Russian Language Studies
Русистика
ISSN 2618-8163 (Print); ISSN 2618-8171 (Online)
2023 Vol. 21 No. 2 212–227
http://journals.rudn.ru/russian-language-studies

DOI: 10.22363/2618-8163-2023-21-2-212-227
EDN: ZYHDWM

Research article

Lexical diversity as a predictor of complexity in textbooks on the Russian language

Anna A. Churunina, Marina I. Solnyshkina, Iskander E. Yarmakeev

Kazan (Volga Region) Federal University, Kazan, Russian Federation
mesoln@yandex.ru

Abstract. The parametric model of the text as a research problem is of paramount importance in modern linguistics and education, since it opens up new approaches to understanding the processes of comprehending texts of various types. In the current study, 17 Russian language textbooks for elementary school were employed to identify correlations between lexical diversity indices and other complexity predictors. The total volume of the corpus compiled for the study is 439,938 words. The two-stage research algorithm included the evaluation of the reference values of text features at the basic level (word length, sentence length, the number of unique, non-repeating words and the number of word forms), evaluation and subsequent contrasting of complexity predictors, i.e. lexical diversity and readability indices. All calculations were performed with the automatic text analyzer RuLingva. The study revealed a positive dynamic of readability and no evidence of lexical diversity increase across grades. An average level of vocabulary diversity and overlaps of every 4th word in the text are fixed. No indication of correlation between text readability and lexical diversity is found. The obtained results can be useful to researchers, textbook authors, and teachers selecting textbooks. The prospects are seen in implementing functional and epidigmatic stratification of the vocabulary of the Russian textbooks under study.

Keywords: elementary school textbooks, text complexity, complexity predictors, readability

Article history: received 13.12.2022; accepted 14.02.2023.

Acknowledgments: This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

For citation: Churunina, A.A., Solnyshkina, M.I., & Yarmakeev, I.E. (2023). Lexical diversity as a predictor of complexity in textbooks on the Russian language. *Russian Language Studies, 21*(2), 212–227. http://doi.org/10.22363/2618-8163-2023-21-2-212-227

Introduction

Text complexity is one of the factors that affect reader perception and understanding of the text. In the modern scientific paradigm, the assessment of complexity is based on the calculation of textual parameters and ends up with predicting

the target reader audience. At the same time, the target audience itself is identified either through the formal learning period (Kupriyanov et al., 2022) or the volume of readers' vocabulary, as, for example, on the platform Lexile.[1] In the first case, we traditionally calculate the text relevance index or the so-called "readability", and in the second case, we estimate the correspondence between the lexicons of the reader and the book. With a certain degree of convention, *readability* is also referred to as syntactic difficulty (Schnick, Knickelbine, 2003), as it depends on sentence length and lexical length as semantic difficulty. Both methods are sufficiently reliable for assessing text complexity and are often used when selecting texts for different reader audiences (Lennon, Burdick, 2004).

Researchers are particularly interested in the difficulty of educational texts because the perception of an instructional text largely determines the success of learning. The problem that has been studied for more than a century is still relevant now. The first works published in 19th century in Russia (Rubakin, 1895), France (Javal, 1878) and England (Sherman, 1893) approach the problem from different sides, but are similar in one aspect: it is important to solve this problem not only for linguistics and educational system, but for the prosperity of the country. At the end of the nineteenth century N.A. Rubakin wrote: "...nothing characterizes the degree of social development, the degree of social culture so much as the level of the reading public at a given historical moment" (Rubakin, 1895: 1). In the Russian biblio-psychological tradition a complex approach is being formed, comsidering both reader's characteristics and text parameters: "...it would be useful to have a look at the reading public itself, to study this public in quantitative and qualitative relations" (Rubakin, 1895: 5). Rubakin especially insists on studying the reader: "How much has been done so far to study the reading public? The Russian reader, both 'grey', 'semi-cultural', and the most intelligent, remains unknown" (Rubakin, 1895: 6).

For more than a century of research on text complexity, dozens of books, hundreds of articles have been published, and the topic has been discussed at numerous conferences (What Do Leaders Need to Know about Text Complexity and Close Reading 2016, What Do Principals Need to Know about Text Complexity and Close Reading 2017, Text Complexity DE Challenge 2022, Educational Challenges 2022: Functional Literacy – Investing in the Future!, Managing the Development of Functional Literacy of Students, GermEval 2022 Workshop on Text Complexity Assessment of German Text, and others). Researchers studying these scientific problems unite in associations (Reading Rockets, The International Literacy Association, International Reading association, Russian Reading Association, etc.). Successful research laboratories and centers such as the Harvard Reads Lab[2] at Harvard University, the SoLET Lab at Arizona State University,[3] the Tex-

[1] The Lexile Framework for Reading – Lexile. Retrieved from https://lexile.com/

[2] Projects at Harvard. Retrieved from https://projects.iq.harvard.edu/reads_summer_learning/home

[3] Science of Learning and Educational Technology. Retrieved from https://soletlab.asu.edu/

tometr[4] project at Pushkin State Russian Language Institute, the Research Laboratory "Text Analytics"[5] at Kazan (Volga Region) Federal University, and others.

In the modern linguistic paradigm, the complexity of nonfiction texts is usually treated as a construction and calculated through estimating the number of elements and the variety of connections between them (morphological, lexical, syntactic, and discursive (Solnyshkina et al., 2022). Researchers name up to 200 text parameters as complexity predictors. Among the most verified for many languages are lexical diversity and readability (Graesser et al., 2004). Lexical diversity is interpreted as "the range and variability of vocabulary that a speaker (and the writer. – *A.Ch., M.S., I.Ya.*) realizes in a text" (McCarthy, Jarvis, 2007: 459). Readability as a property of a text perceived by the reader is calculated on the average word length and sentence length in the text (Kincaid et al., 1975).

Of all various complexity predictors validated by contemporary authors (Solnyshkina et al., 2022), the lexical diversity or richness of the lexicon of educational texts is the least studied question (Kharchenko, 2017). At the same time, it is important to emphasize that numerous works are devoted to the richness of the vocabulary of fiction authors (see: Vasilyev, Zhatkin, 2020): a wide palette of methods for studying the language of a fictional text – from tropes to syntax preferences, from creating concordances and dictionaries to analyzing intertextuality – has been developed within the modern scientific paradigm (see: Fateeva, 2013). The choice of fictional texts and authors to research the richness of a writer's language is never random: works with the richest language, the subtlest shades of meaning, and lexical findings are chosen, each of them is strictly documented and illustrated by carefully selected quotations. And it is understandable: the influence of the writer's word on the reader cannot be overestimated.

As for educational texts, philology "has not yet tended to treat <them> as carefully as artistic fabric" (Kharchenko, 2017: 23). There are practically no studies on the richness of the vocabulary of Russian language manuals and textbooks. To confirm this, let us point to three publications (Veselovskaya, 2020; Laposhina et al., 2018; Kupriyanov et al., 2022). At the same time, experts have special requirements to the language of the textbook: it should "talk" to the student in lively language, use figurative, memorable comparisons that evoke vivid associations in the mind (see: Donskoy, 1985: 162). The textbook on the Russian language is in the focus as a textbook on "subject of subjects" (Buslaev, 2019), which plays a meta-disciplinary role and largely determines not only the academic success of the student, but also the ability to realize themselves in life. The language of Russian language textbooks is designed to have a "pronounced semantic orientation of grammar and orthographic material"; contribute to "the formation of aesthetic taste of students by means of the language itself" and be characterized

[4] Textometr – text complexity analysis online. Retrieved from https://textometr.ru/

[5] The Research Laboratory "Text Analytics". Retrieved from https://kpfu.ru/philology-culture/struktura-instituta/otdelenie-russkoj-i-zarubezhnoj-filologii-imeni/kafedra-inostrannih-yazikov/nil-39intellektualnye-tehnologii-upravleniya

by “a wide use of <...> material that has a value-and-sense orientation” (Lvova, 2013: 65).

An interesting and particularly significant issue when selecting educational materials for a particular target audience is the question of the optimal range of lexical diversity, which are always assessed in linguistic expertise of academic publications in English (see: McCarthy, Jarvis, 2010). For texts in Russian, it is currently really relevant to identify “diagnostic” criteria for describing norms, i.e. the range of lexical diversity in academic texts of a particular subject area. It is important to describe texts with an extremely rich language and without repetitions, which ensure the coherence of the text. This makes the text extremely difficult to comprehend. Opposed to the texts of this type are texts with numerous repetitions and such a monotonous vocabulary that the reader loses interest and refuses to read them. Establishing the vocabulary range of the most popular textbooks can form the basis for a typology of lexical diversity in texts of different genres and varying degrees of complexity. A research niche in Russian philology and linguodidactics remains the issue of this parameter dynamics as the complexity of a textbook text increases.

It is significant that the term “lexical diversity”, according to Ngram Viewer[6] data, was first recorded and has been functioning in Russian discourse since the 1920s (Figure 1).

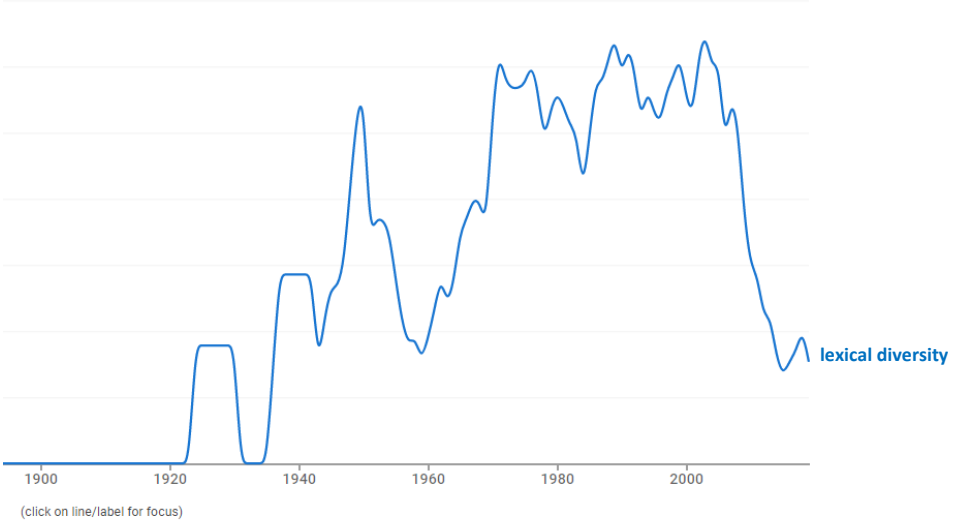

Figure 1. Frequency fluctuations of term “lexical diversity” in Russian discourse

The context of the term semanticizes its intensional as “lexical richness” or the author's lexicon. For example, “The expressive character of the speech is supported by the remarks accompanying the speech; their number in any melodrama is extensive, and the *lexical variety* shows the melodramatist's search for

[6] Google books Ngram Viewer. Retrieved January 15, 2023, from http://books.google.com/ngrams

vivid and unmistakable tones of speech" (Ngram Viewer. (1927). *Poėtika*, (3); "The *lexical diversity* of Pushkin's letters is extremely rich" (Ngram Viewer. (1937). *Izvestia of the USSR Academy of Sciences*). Modern contexts confirm the semantic stability of the term: "It has been shown that the *lexical diversity* and variety of word combinations, compound and complex constructions in the speech of a parent when his child is 1 year old conditions the same characteristics of speech diversity at the age of 4 years" (Chernov, D.N. (2013). Sociocultural conditionality of language competence of a child. *Ngram Viewer*). "Let us first consider the *lexical diversity* of the text. Let us note that in this story Chekhov did not give his characters his usual grotesque surnames and names" (Ulin, V. (2013). Literary Institute. *Ngram Viewer*). "The *lexical diversity* of nouns naming rituals and celebrations testifies not to idle life, but to the bright, characteristic elements of the peasant way of life based on ancient traditions" (Ngram Viewer. (2007). *Lexical Atlas of Russian Folk Vocabulary*).

Since scientific style texts have a high index of lexical diversity (McCarthy, Jarvis, 2010; Richards, 1987), it is obvious that texts for high school students with a higher degree of "scientificity" compared to texts for younger students, should have a higher index of lexical diversity. Consequently, the lexical diversity of educational texts of one subject block, and this is the *hypothesis* of the study, grows from grade to grade. Thus, **the aim of the research** is (1) to identify the dynamics of lexical diversity in Russian language textbooks and (2) to establish the relationship between readability and lexical diversity indices.

Methods and materials

The study was carried out on the material of Russian educational texts for the younger grades from the Educational Corpus of the Russian Language (ECRL[7]), which currently exceeds 8 million words. To preserve copyrights, the Corpus is used as a closed one exclusively for scientific projects; only its demonstrative sample – randomly shuffled texts of social studies textbooks (CORAT[8]) – is in open access. The core of the CORAT consists of elementary, middle, and high school educational and examination texts, including texts for Unified State Examination and the Main State Exam in all subject areas. The corpus also includes texts for studying Russian as a foreign language. The representativeness and balance of the ECRL has been proven in a number of studies (Kupriyanov et al., 2022; Solovyev et al., 2018), which makes it very valuable for studying the current state of scientific and academic style.

The corpus of the study amounted to 439, 938 word forms, it included the texts of 17 Russian language textbooks for grades 2–4, included in the Federal

[7] Database State Registration Certificate No. 2020622254.

[8] The Research Laboratory "Text Analytics". Retrieved from https://kpfu.ru/philology-culture/struktura-instituta/otdelenie-russkoj-i-zarubezhnoj-filologii-imeni/kafedra-inostrannih-yazikov/nil-39intellektualnye-tehnologii-upravleniya

list of textbooks approved for use in state-accredited educational programs of primary general, basic general, secondary general education in organizations involved in educational activities.[9] All textbooks were published between 2009 and 2020.

Complexity parameters were calculated with the automated text analyzer RuLingva[10] (see: Solovyev et al., 2018), created by a team of Russian scientists to automate routine arithmetic and research operations with Russian texts. The descriptive text parameters include the number of words, sentences, syllables, repeated and non-repeated words, one-, two-, three- and four-syllable words, etc. RuLingva can make lists of terms, notional parts of speech, as well as certain morphological categories and discourse markers extracted from the analyzed text. RuLingva was developed in the framework of the Russian Science Foundation project "Complexity of texts in Russian"[11] with two main goals: to identify and describe typological parameters of academic texts and to develop methods of their ranking by levels of complexity. The RuLingva text ranking by level of difficulty is based on the identified correlations of text parameters and typical reader characteristics (age, education, vocabulary volume).

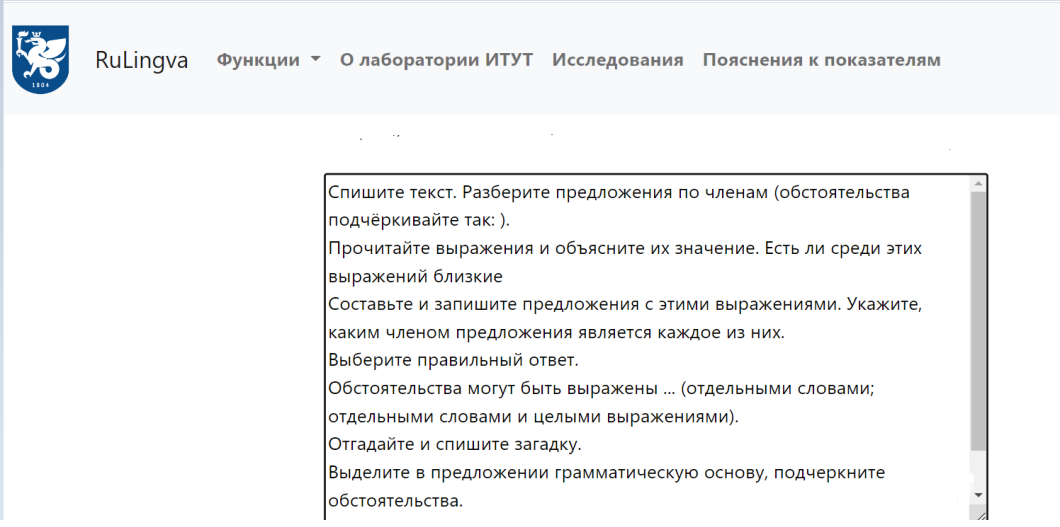

Figure 2. RuLingva Interface

Currently, RuLingva performs automatic linguistic analysis of texts up to 50,000 words and evaluates 47 parameters of Russian texts (Figure 2), including the number of word forms and words, average word length of the loaded text in syllables, average sentence length in words, lexical diversity and readability indices, connectivity, abstractness index, number of terms, number of morphological

[9] The federal list of textbooks. Retrieved from https://fpu.edu.ru/

[10] RuLingva. Retrieved from https://rulingva.kpfu.ru/

[11] The card of the project supported by the Russian Science Foundation. Retrieved from https://rscf.ru/prjcard_int?18-18-00436

parameters, etc. RuLingva allows uploading and saving data in excel spreadsheet format (Figure 3).

32	Genitive Case (Noun)	61
33	Dative Case (Noun)	28
34	Accusative Case (Noun)	73
35	Instrumental Case (Noun)	26
36	Prepositional Case (Noun)	20
37	Present Tense (Verb)	49
38	Future Tense (Verb)	1
39	Past Tense (Verb)	29
40	Interrelation of verbs and nouns	0.38
41	Interrelation of adjectives and nouns	0.2
42	The percentage of nouns in genitive case	0.21
43	The number of terms on social studies	14
44	The number of one-syllable words	135
45	The number of two-syllable words	148
46	The number of three-syllable words	135
47	The number of four-syllable words	108

Загрузить результаты

Figure 3. List of parameters assessed by RuLingva

According to the modern approach in Russian and foreign linguistics (see: Biber, 2006; Solnyshkina et al., 2022), two groups of words are evaluated in the lexical diversity coefficient: repetitive and non-repetitive. That is why automated calculation of lexical diversity seems rather non-trivial: a significant shortage of such evaluation is its "sensitivity" to the length of the text: the longer the text is, the more functional words it contains, and the lower the lexical diversity is (see line 28, Figure 4). This parameter is accurate enough only if the length of the passage does not exceed 1000 word forms (Biber, 2006; Vakhrusheva et al., 2021).

RuLingva allows to calculate the average lexical diversity of the whole text regardless of its length (TTRavg – type token ratio average) by dividing the text into passages of 1000 word forms, measuring lexical diversity separately in each passage and suggesting the arithmetic average (see line 29, Figure 4).

28	TTR	0,15
29	TTRavg	0,32
30		

Figure 4. TTR metrics

As part of this study, the following predictors of complexity were calculated for each of the textbooks: (1) the number of word forms, (2) the number of unrepeated words, (3) the index of lexical diversity, (4) average word length (in syllables), (5) average sentence length (in words), and (6) the Flesch – Kincaid readability index (see Tables 1–3). These quantitative parameters were chosen because they identify the basic set of already studied and described indicators, allowing to interpret the numerical data obtained in the analysis of texts (Kupriyanov et al., 2022). *The number of word forms* in the text and *the number of non-repeating words* are believed to have a direct impact on *lexical diversity index* (type-token ratio, TTR, lit. *word-to-word forms ratio* (Graesser et al., 2004: 1)), which is calculated as the ratio of non-repeating words (word types) to the total volume of text in word forms (word tokens) (Templin, 1957). When TTR = 1.0, none of the words in the text are repeated. Obviously, this kind of text can only be created artificially because the lack of lexical repetition makes it difficult to perceive the text. Low values of TTR (< 0.5) signal a high repetition of words, which positively affects the speed of text processing by the reader. The target audience for this type of texts is users with limited vocabulary (language learners or elementary school students) (Malvern et al., 2004). Vocabulary diversity is interpreted in this case as the vocabulary used by the author of the text, reflecting his/her ability to use certain lexical units (Fergadiotis, Wright, 2011). It is a measure of the speech act success, including speech-language pathology situations and cross-cultural communication (Fergadiotis et al., 2013; Owen, Leonard, 2022).

Average word length and *average sentence length* as predictors of text complexity are used to calculate the readability index. The formula for calculating the readability of Russian texts was based on Flesch – Kincaid Grade Level formula (Kincaid et al., 1975), but it considered systemic differences between Russian and English languages (Solnyshkina et al., 2018):

$$\text{Readability} = 208.7 - 2.6 \times \text{ASL} - 39 \times \text{AWL},$$

where ASL – average sentence length in words; AWL – average word length in syllables.

The readability formula ranks texts by grade, i.e. according to the learning period needed for the reader to comprehend the text. For example, if the calculated readability is 2.5, then the text is addressed to 2nd or 3rd graders, and if the value is between 3.0 and 4.0, then it is addressed to 3rd and 4th graders, etc.

Results

The study of lexical diversity dynamics in Russian language textbooks for elementary school and its possible correlation with readability revealed the specifics of the language used in Russian school textbooks. In terms of readability texts in the studied textbooks are highly likely to cause difficulty in understanding for the target audience, because the calculated indices are on average one or two

levels higher than expected. The index of vocabulary richness in the textbooks ranges from 0.33 to 0.55, which is average for textbooks. The revealed dynamics of lexical diversity showed an uneven change in the Russian language instructional texts complexity both within one line of textbooks and within the entire corpus of texts studied as a whole. No correlation was found between text readability and lexical diversity, the growth of lexical diversity index from grade 2 to grade 4 was not detected.

Discussion

Tables 1–3 show the data obtained during the analysis of the corpus of texts according to the six difficulty parameters.

Table 1

Complexity predictors of Russian textbooks for the 2nd grade

No.	Author, year	Grade	Complexity predictors					
			Tokens	Types	TTR	Average word length, syllables	Average sentence length, words	FKGL
1	Ramzaeva T., 2011[12]	2	13 689	2961	0.48	2.18	5.68	2.63
2	Zheltovskaya L., Kalinina O.; 2012[13]	2	26 877	4632	0.47	2.34	6.79	3.93
3	Klimanova L., Babushkina T.; 2012[14]	2	8001	2622	0.55	2.17	7.54	3.27
4	Nechaeva N., 2013[15]	2	19 168	4138	0.49	2.25	8.29	3.98
5	Soloveychik M., Kuzmenko N.; 2013[16]	2	20 422	2777	0.41	2.22	7.33	3.44
6	Kanakina V., Goretskiy V.; 2017[17]	2	25 020	4626	0.45	2.38	6.63	4.11
Average			18 863	3626	≈ 0.48	2.26	7.04	3.56

[12] Ramsaeva, T.G. (2011). *Russian language. 2 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ., Drofa Publ. (In Russ.)

[13] Zheltovskaya, L.Ya., & Kalinina, O.B. (2012). *Russian language. 2 grade : Textbook in 2 parts*. Moscow: Drofa Publ. (In Russ.)

[14] Klimanova, L.F., & Babushkina, T.V. (2012). *Russian language. 2 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ. (In Russ.)

[15] Nechaeva, N.V. (2013). *Russian language. 2 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ. (In Russ.)

[16] Soloveychik, M.S., & Kuzmenko, N.S. (2013). *Russian language. 2 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ., Binom Publ. (In Russ.)

[17] Kanakina, V.P., & Goretskiy, V.G. (2017). *Russian language. 2 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ. (In Russ.)

Table 2

Complexity predictors of Russian textbooks for the 3rd grade

No.	Author, year	Grade	Complexity predictors					
			Tokens	Types	TTR	Average word length, syllables	Average sentence length, words	FKGL
1	Ramzaeva T.; 2009[18]	3	20 763	3886	0.50	2.34	6.49	3.82
2	Ivanov S., Evdokimova A., Kuznetsova M. et al; 2013[19]	3	39 318	5498	0.47	2.31	8.05	4.21
3	Kanakina V., Goretskiy V., 2013[20]	3	30 700	4410	0.43	2.56	6.26	5.02
4	Klimanova L., Babushkina T.; 2014[21]	3	31 424	5530	0.49	2.39	7.07	4.34
5	Soloveychik M., Kuzmenko N.; 2014[22]	3	27 343	3468	0.41	2.26	7.61	3.81
6	Zelenina L., Khohlova T.; 2015[23]	3	28 713	2998	0.33	2.62	6.80	5.58
Average			29 710	4298	≈ 0,44	2.41	7.05	4.46

[18] Ramzaeva, T.G. (2009). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ., Drofa Publ. (In Russ.)

[19] Ivanov, S.V., Evdokimova, A.O., Kuznetsova, M.I., Petlenko, L.V., & Romanova, V.Yu. (2013). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Ventana-Graph Publ., Rossiiskii Uchebnik Publ. (In Russ.)

[20] Kanakina, V.P., & Goretskiy, V.G. (2013). *The Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ. (In Russ.)

[21] Klimanova, L.F., & Babushkina, T.V. (2014). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ. (In Russ.)

[22] Soloveychik, M.S., & Kuzmenko, N.S. (2014). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ., Binom Publ. (In Russ.)

[23] Zelenina, L.M., & Khohlova, T.E. (2015). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ. (In Russ.)

Table 3

Complexity predictors of Russian textbooks for the 4th grade

No.	Author, year	Grade	Complexity predictors					
			Tokens	Types	TTR	Average word length, syllables	Average sentence length, words	FKGL
1	Zelenina L., Khohlova T.; 2012[24]	4	29 906	4138	0.41	2.6	7.45	5.71
2	Kanakina V., Goretskiy V.; 2013[25]	4	33 716	4739	0.44	2.6	6.62	5.39
3	Ramzaeva T., 2013[26]	4	30 020	4861	0.49	2.36	6.09	3.82
4	Klimanova L., Babushkina T.; 2014[27]	4	30 014	4966	0.47	2.43	7.42	4.69
5	Zheltovskaya L., Kalinina O.; 2020[28]	4	24 844	4936	0.50	2.41	7.76	4.7
Average			29 700	4728	0.46	2.48	7.07	4.86

The average readability ranges from 2.63 to 5.7, with an average of 3.56 for the second-grade texts, 4.46 for the third-grade texts, and 4.86 for the fourth-grade texts. Apart from second-grade textbooks, the readability index corresponds to the grade. For second-grade textbooks, the readability index fluctuates between 2.63 and 4.11, which means that for the most part they are significantly (1.5 to 2.5 points) above the norm (see: Solnyshkina et al., 2020).

The texts show a gradual increase in the average number of unrepeated words from grade 2 to grade 4. This index gradually increases from an average of 3,626 words for grade 2 textbooks to 4,728 words for grade 4 textbooks.

The average lexical diversity index ranges from 0.3 to 0.55 with an average of 0.46 for the entire corpus of texts, which indicates a high number of repeated

[24] Zelenina, L.M., & Khohlova, T.E. (2012). *Russian language. 4 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ. (In Russ.)

[25] Kanakina, V.P., & Goretskiy, V.G. (2013). *The Russian language. 4 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ. (In Russ.)

[26] Ramzaeva, T.G. (2013). *Russian language. 4 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ., Drofa Publ. (In Russ.)

[27] Klimanova, L.F., & Babushkina, T.V. (2014). *Russian language. 4 grade: Textbook in 2 parts*. Moscow: Prosveshcheniye Publ. (In Russ.)

[28] Zheltovskaya, L.Ya., & Kalinina, O.B. (2020). *Russian language. 4 grade: Textbook in 2 parts*. Moscow: Drofa Publ. (In Russ.)

lexical units in the texts of the studied textbooks. The obvious reason is the specificity of the texts included in the textbooks on the Russian language and the chosen period of study, which is characterized by methodical repetition of learning activities in order to form a skill. The textbooks contain instructions to exercises of a certain pattern facilitating perception and understanding of the instructions by students.

As we noted above, a text with high lexical diversity is considered to be more complex (Richards, 1987). Two texts with the same number of word forms and non-repeating words are similar in lexical diversity and richness, while two texts with the same number of word forms and different numbers of non-repeating words have different lexical diversity. Notably, the textbook with the lowest lexical diversity of 0.33 in the corpus under consideration is not the second-, but the third-grade textbook.[29] One would expect fourth-grade textbooks to have a higher level of lexical diversity, since students of this age should have a higher level of language proficiency, but even in the fourth grade the level of lexical diversity does not rise above 0.55. Thus, the hypothesis of the study is not confirmed, because there is no growth in lexical diversity even in the textbooks of the same line. For example, the dynamics of lexical diversity in the textbooks edited by T.G. Ramzaeva is quite contradictory: 0.48 (2^{30}) – 0.5 (3) – 0.49 (4). Lexical diversity indexes do not grow in the line of textbooks edited by M.S. Soloveychik and N.S. Kuzmenko: the index is 0.41 for all levels. The negative dynamics in the lexical diversity was revealed in the textbooks edited by L.F. Klimanova, T.V. Babushkina (0.55 (2) – 0.49 (3) – 0.47 (4)), and positive dynamics was observed only in the 3rd–4th grade textbooks edited by L.M. Zelenina, Khohlova T.E.: 0.33 (3) – 0.41 (4). However, in the latter case the index of lexical diversity is below average, which indicates, on the one hand, numerous repetitions in the text, i.e. the absence of a real wealth of vocabulary, and, on the other hand, provides coherence and easy understanding.

A deeper discussion should touch the identified lack of correlation between readability and lexical diversity: regardless of the readability, the texts in the textbooks have an average lexical diversity. For example, the lexical diversity in the textbook edited by T.G. Ramzaeva with a readability index of 3.82 and in the textbook edited by L.F. Klimanov and T.V. Babushkina with a readability index of 4.34 is the same and amounts to 0.49.

In some cases lexical and syntactic complexity are balanced. For example, in the textbook edited by L.F. Klimanova and T.V. Babushkina for the 2nd grade the relatively high lexical diversity (0.55) is balanced by a lower readability – 3.27, and in the 4th-grade textbook edited by L.M. Zelenina and T.E. Khohlova the relatively low lexical diversity corresponds to a higher readability – 5.58.

[29] Zelenina, L.M., & Khohlova, T.E. (2015). *Russian language. 3 grade: Textbook in 2 parts.* Moscow: Prosveshcheniye Publ. (In Russ.)

[30] The number in parenthesis shows the grade.

Conclusion

An adequate level of linguistic complexity of learning materials is believed to be crucial for students' development. Among a wide range of complexity predictors, lexical diversity and readability are of paramount importance because of their high "demonstrative" potential, their ability to reflect both syntactic and lexical parameters of the text. Our results provide researchers, textbook developers, and practitioners with data on qualitative differences in the textbooks studied and can be used by scholars and practitioners in developing instructional materials and in linguistic expertise. Data on the lexical diversity of instructional texts can become the basis for automatic determination of text type and can be used, for example, in text profilers and search browsers. It can also be useful for the examination of educational materials when writing textbooks and developing test materials and tests of different levels. In the light of the data obtained, the expansion of the corpus of research and identification of the lexical diversity of middle and high school Russian language textbooks seems very promising. The frequency of the vocabulary used in Russian language textbooks and its connection to the nuclear vocabulary of the Russian language is of special interest.

References

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

Buslaev, F.I. (2019). *Issues of teaching national language*. Moscow: URAIT Publ. (In Russ.)

Donskoy, G.M. (1985). Typological properties of modern textbook. *Problems of Modern Textbook: Typology of School Textbooks: Collection of Articles* (issue 15, pp. 70–86). Moscow: Prosveshchenie Publ. (In Russ.)

Fateeva, N.A. (2013). Intertext as a form of discursive interactions and as an environment of cultural concepts (based on works of Y.S. Stepanov). *Linguistic Parameters of the First Civilization: Proceedings of the First Scientific Conference (in Memory of Y.S. Stepanov)* (pp. 348–358). Moscow: Distance Education Center "Eidos". (In Russ.)

Fergadiotis, G., & Wright, H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation task. *Aphasiology*, *25*(11), 1414–1430.

Fergadiotis, G., Wright, H., & West, T. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), 397–409.

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

Javal, E. (1878). Essai sur la physiologie de la lecture. *Annales d'Oculustique, 79*, 97–117.

Kharchenko, V.K. (2017). On the richness of word-stock and calculation of the coefficient of lexical variety in the "History of the Russian church" by metropolitan Makary (Bulgakov). *Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication*, (3), 21–25. (In Russ.)

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel*. Research Branch Report 8–75. Millington, Tennessee: Institute for Simulation and Training.

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.

Laposhina, A.N., Veselovskaya, T.S., Lebedeva, M.Y., & Kupreshchenko, O.F. (2018). Automated text readability assessment for Russian second language learners. *Dialogue 2018: Proceedings of the International Conference, 17*(24), 396–406.

Lennon, C., & Burdick, H. (2004). *The LEXILE framework as an approach for reading measurement and success.* MetaMetrics, Inc.

Lvova, S.I. (2013). Russian language textbook as a basis for education, development and upbringing of contemporary pupil. *Municipal Education: Innovations and Experiment*, (1), 63–70. (In Russ.)

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and languagedevelopment: Quantification and assessment.* Basingstoke: Palgrave MacMillan.

McCarthy, P.M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*, 459–488.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.

Owen, A.J., & Leonard, L.B. (2002). Lexical diversity in spontaneous speech of children with specific language impairment. *Journal of Speech Language and Hearing Research, 45*, 927–937.

Richards, B. (1987). Type/Token Ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201–209.

Rubakin, N.A. (1895). *Studies on Russian reading public: Facts, numbers, observations*. St. Petersburg: Sklad Izdaniya N.P. Karbasnikova Publ. (In Russ.)

Schnick, Th., & Knickelbine, M. (2003). *The Lexile framework: an introduction for educators*. MetaMetrics, Inc.

Sherman, L.A. (1983). *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn and Co.

Solnyshkina, M., Guryanov, I., Gafiyatova, E., & Varlamova, E. (2018). Readability metrics: The case of Russian educational texts. *Abstracts & Proceedings of ADVED 2018 – 4th International Conference on Advances in Education and Social Sciences* (pp. 676–681). Istanbul: OCERINT.

Solnyshkina, M.I., Harkova, E.V., & 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.

Solnyshkina, M.I., Solovyev, V.D., Gafiyatova, E.V., & Martynova, E.V. (2022). Text complexity as interdisciplinary problem. *Issues of Cognitive Linguistics*, (1), 18–39. (In Russ.)

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.

Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.

Vakhrusheva, A.Y., Solnyshkina, M.I., Kuprijanov, R.V., Gafiyatova, E.V., & Klimagina, I.O. (2021). Linguistic complexity of academic texts. *Issues in Journalism, Education, Linguistics, 40*(1), 88–99). (In Russ.)

Vasilyev, N.L., & Zhatkin, D.N. (2020). "Pushkin Dictionary" by G.A. Shengeli: The unpublished article by the author of the concordance of A.S. Pushkin's poems. *Literary Fact*, (1), 458–476. (In Russ.)

Veselovskaya, T.S. (2020). The linguistic world-image in the Russian language primary school textbooks: A corpus study. *Ethnopsycholinguistics*, (3), 224–237. (In Russ.)

Bio notes:
Anna A. Churunina, Assistant Lecturer, Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests:* text analytics, corpus linguistics, computational linguistics, comparative linguistics. ORCID: 0000-0002-7385-9911. E-mail: churunina.anna@gmail.com

Marina I. Solnyshkina, Doctor Habil. of Philology, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Head and Chief Researcher of "Text Analytics" Research Lab, Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. The author of two monographs and over 65 publications on discourse complexology and text complexity. *Research interests:* text complexity assessment, text comprehension, natural language processing, sociolinguistics, comparative linguistics. ORCID: 0000-0003-1885-3039. E-mail: mesoln@yandex.ru

Iskander E. Yarmakeev, Doctor of Pedagogy, Professor of the Department of Linguistic and Intercultural Communication, Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. Honored worker of higher school of the Republic of Tatarstan, honorary worker of higher professional education of the Russian Federation, member of the Scientific Council on Problems of History of Education of the Russian Academy of Education, and member of the International Pedagogical Academy. *Research interests:* theory and methods of teaching, analytics of academic texts. ORCID: 0000-0002-1103-6469. E-mail: ermakeev@mail.ru

DOI: 10.22363/2618-8163-2023-21-2-212-227
EDN: ZYHDWM

Научная статья

Лексическое разноообразие как предиктор сложности учебников по русскому языку

А.А. Чурунина, М.И. Солнышкина, И.Э. Ярмакеев

Казанский (Приволжский) федеральный университет, Казань, Российская Федерация
mesoln@yandex.ru

Аннотация. Параметрическая модель текста как научная проблема имеет первостепенное значение в современной филологии и образовании, поскольку открывает новые подходы к пониманию процессов восприятия текстов различных типов. В исследовании для идентификации корреляций индексов лексического разнообразия с другими предикторами сложности использовались 17 учебников русского языка для начальной школы. Общий объем корпуса исследования составил 439 938 слов. Двухэтапный алгоритм исследования включал оценку референтных значений текстовых параметров базового уровня (длина слова, длина предложения, количество неповторяющихся слов и количество словоформ), оценку и последующее контрастирование предикторов сложности – индексов лексического разнообразия и читабельности. Все расчеты производились при помощи автоматического анализатора текстов RuLingva. Выявлено, что индекс читабельности изучаемых учебников русского языка демонстрирует положительную динамику. Рост лексического разнообразия от класса к классу не обнаружен. Зафиксирован средний уровень разнообразия лексикона, при котором каждое четвертое слово в тексте повторяется. Корреляции между читабельностью текста и лексическим разнообразием не выявлены.

Полученные результаты могут быть полезны исследователям, разработчикам учебников и учителям в процессе выбора учебника. Текущая перспектива видится в осуществлении функциональной и эпидигматической стратификации лексики изучаемых учебников русского языка.

Ключевые слова: учебники начальной школы, сложность текста, сложности, читабельность

История статьи: поступила в редакцию 13.12.2022; принята к печати 14.02.2023.

Благодарности: Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Для цитирования: *Чурунина А.А., Солнышкина М.И., Ярмакеев И.Э.* Лексическое разноообразие как предиктор сложности учебников по русскому языку // Русистика. 2023. Т. 21. № 2. С. 212–227. http://doi.org/10.22363/2618-8163-2023-21-2-212-227