Lexical Features of Text Complexity: the case of Russian academic texts

Anna Churunina[1*], *Marina* Solnyshkina[1], *Elzara* Gafiyatova[1], and *Artem* Zaikin[1]

[1]Kazan Federal University, Kremlyovskaya St, 18, Kazan, Republic of Tatarstan, 420008, Russian Federation

Abstract. The work presented in this paper is a part of an ongoing project that investigates academic text features indicative of its complexity at different grade levels. In this study we examine comparative complexity of Social science texts used in Russian secondary and high schools. Based on the metrics of ten descriptive and four lexical features assessed for seven classroom textbooks we claim lexical diversity, frequency, abstractness and the number of terminological units to be statistically significant predictors of text complexity. The total size of the Corpus of over 160.000 tokens comprising two sets of textbooks ranging from the 5th to the 11th grades provides a satisfactory level of its representativeness and as such a solid foundation for statistical validity of the results. We employ RusAC, an online text analyzer, to compute lexical features of texts and the effect of the four lexical features on text complexity is confirmed with a mixed analysis of variance. The study fills a gap both in corpus linguistics as regards a systematic approach to Russian academic texts and in text complexity studies as regards the description of secondary and high school textbooks.

1 Introduction

As a focus of numerous studies for over fifty years, the problem of assessment of Russian texts linguistic complexity is still viewed theoretically valuable [1-3]. The research in the area is aimed at designing an algorithm identifying a "target reading audience" and validating a list of text features which effect its complexity. The latter is especially significant nowadays due to the increased information flow and cognitive density of modern academic texts [4]. The three lexical features with the highest impact on academic text complexity validated in the recent studies are lexical diversity, frequency and abstractness [5]. The total count of terms is viewed as an additional predictor of text complexity in studies on reading comprehension [6, 7].

2 Literature review

After the 1890s when the research on text complexity assessment for native and foreign language speakers began, discussions on shortening the list of parameters defining text complexity have been ongoing quite intensively [3].

* Corresponding author: silmarill1397@gmail.com

2.1 Readability

Readability determines the level of reading ease of a text and is measured based on solely quantitative parameters: 1) number of sentences in the text; 2) average number of syllables in the words of the text. The most popular readability formula, i.e. Flesh-Kincaid Grade level, ranks text appropriateness for a certain school grade [8]:

$$FKG = (0{,}39 \times ASL) + (11{,}8 \times ASW) - 15{,}59, \quad (1)$$

where 'ASL' stands for average sentence length, 'ASW' stands for average syllables per word. The index obtained lies corresponds a grade level.

The equation adapted for the Russian language and validated in numerous studies, proved its reliability when applied to academic texts [9]:

$$FKG\ (SIS) = 208{,}7 - 2{,}6 \times ASL - 39 \times ASW, \quad (2)$$

However, though quantitative readability measures, i.e. average sentence and word length, enable researchers to compare text descriptive metrics and quite ubiquitously applied, do not allow to compare texts in an objective way. They ignore numerous text characteristics which can influence its readability. A word, though long, may be so frequently used in the discourse and thus present no difficulty for readers. E.g. words international and morphological have the same ASW, i.e. five syllables, but their frequency registered in COCA is strikingly different: 158.89 vs 1.54 per mln. [10]. Thus, due to its frequency the word international presents much less difficulty for an average potential reader than the word morphological.

By now studies aimed at extending the list of 'qualitative' features affecting text complexity have been going on for over a century. Nowadays researchers integrate text features estimating not only descriptive metrics, but morphological (parts of speech, distribution, etc.), lexical, syntactical and discourse parameters [3].

2.2 Morphological Distribution

Morphological categories distribution has been a focus of linguists' interest for decades [11-13].

Defining 'progression towards a more 'academic' style', which is in fact progression towards a higher degree of complexity, D. Biber [14] indicates higher scores of nouns and groups, fewer verbs and verb groups, more nominalisations of verbs and adjectives, and a greater number of abstract nouns and long words. D. Biber also validates a high level of 'informational density' of Social Science texts realized in the above mentioned morphological and lexical categories [14].

2.3 Lexical features

Vocabulary range and its awareness appeared on the list of text complexity parameters as early as 1900s, since many researchers now and then view vocabulary features as fundamental to reading comprehension and correlation between verbature or a person's vocabulary size and reading comprehension is an acknowledged fact [15].

Type-token ratio (TTR) has been widely used in assessments of texts lexical diversity since 1957, when M. Templin introduced it [16]:

$$TTR = \frac{\text{word types}}{\text{word tokens}}, \quad (3)$$

where 'word types' are unique, i.e. not repeated, words in a text and 'word tokens' are total amount of words in a text.

In early 2000s research showed that unique words distribution is not linear for corpora of various sizes due to the fact that words tend to repeat themselves: the bigger the corpus, the more repeated words it comprises. Hence, only relatively small amount of words is going to increase along with the corpus enlargement, which causes extra difficulties while comparing corpora of different sizes. Thus, it was suggested to assess TTR per 1000 words [14].

One of the first text complexity formulas, the Dale–Chall formula [17], employs vocabulary lists to rate books for grade levels. The ratio of listed words in a text provides the data to measure complexity of a text and correlate it with a grade level. In 1981, Anderson & Freebody also claimed the ratio of difficult words in a reading text to be the best predictor of text complexity [18].

Another feature directly influencing text complexity is lexical frequency: the more high frequency words are used in the text, the easier it is for the reader. (cf. international and morphological above). The research validating frequency as a function of complexity has integrated into corpus linguistics and is finalized in online servers as Lexile [19].

Russian frequency indices estimated with the help of Frequency dictionary [20] and are also successfully used to assess Russian texts complexity [21-22].

In cases when frequency lists are unavailable or corpora lack frequency annotation, researchers resort to simpler lexical metrics, e.g. the number of terminological units or nomenclature in a text to assess its complexity. R.V. Mayer (2016) introduces a new text complexity notion, didactic complexity, which rests on the metric of the number of terms in a text alongside with mathematical symbols count and information density of a text [23].

A significant number of models and ideas have been developed to estimate text abstractness after abstractness or degree of abstractness was validated as a metric of text complexity. Among the most popular are numerical indices or ratings of abstractness and scales of abstractness/ concreteness of different range: from 1 to 5 [24], 0 to 9 [25], 1 to 7 [26].

In summary, text complexity is a developing notion, not a well-defined concept. Though the features presented above estimate text complexity based on conventional metrics only they provide a fine-grained assessment of text age, cognitive and linguistic appropriateness.

3 Methods and Material

In this paper, we aim at identifying the effect of lexical parameters cluster on academic text complexity. The set of lexical parameters comprises (1) lexical diversity, (2) number of terminological units, (3) words frequency and (4) abstractness of text vocabulary. The research focus is in quantifying differences at various grade levels (5-11) thus providing the data to automate text complexity assessment.

The research data are 70 academic texts extracted from school textbooks "Social science, Grades 5-11" [27] with the total size of over 160.000 tokens. We used the online service RusAC [22] to compute texts lexical features and performed a preliminary mixed analysis of variance (Spearman) to define the effect of the four lexical features.

The research data are Social Science students textbooks written by Bogolyubov (2012-2014) for grades 5th -11th recommended for all Russian schools by the Ministry of Education.

4 Analysis

On Stage 1 we compiled Russian Social Science Academic Corpus (RSSAC) as a subcorpus of Russian Academic Corpus [24] and estimated the following: (1) the size of RSSAC; (2) the size of each textbook in RSSAC and (3) the size of 10 samplings from each textbook. All the samplings from the textbooks were coded with the grade number, the subject title and the number of the sampling as '5SS1', where '5' stands for 'the 5th grade', 'SS' - for 'Social Science' and '1' is Sampling #1. (see Table 1)

Table 1. Size and structure of Russian Social Science Academic Corpus.

Grades	Textbook	Sampling
5	10083	1008
6	10135	1013
7	11226	1122
8	24027	2402

Grades	Textbook	Sampling
9	21184	2118
10	38440	3844
11	52803	5280
RSSAC	**167898**	**2398**

The problem of corpus representativeness is associated not only with its size but its quality and genre range [14]. Russian Social Science Academic Corpus used for the current study is defined as representative based on the fact that it represents a certain language variety, i.e. classroom books texts used to teach Social science in Russia. (https://4ege.ru/obrazovanie/60190-utverzhden-federalnyj-perechen-uchebnikov-na-5-let.html). We also defined the size of the sampling in each textbook based on the formula designed by D. Biber [14]:

$$T(s) = T(t) : 20, \quad (4)$$

where T(s) is the number of tokens in a sampling and T(t) is the number of tokens in a textbook. The size of samplings varies from 1008 in the 5th grade to 5280 in the 11th grade.

On Stage 2 we processed each of the 10 samplings in each textbook with the help of Russian texts' processing service RusAc [28], that provided statistics on the following metrics: 1) total amount of words; 2) total amount of syllables; 3) total amount of sentences; 4) average amount of words per sentence; 5) average amount of syllables per word; 6) adjectives count; 7) adverbs count; 8) pronouns count; 9) nouns count; 10) verbs count; 11) words frequency (based on Sharoff's dictionary); 12) Flesch-Kincaid Grade (SIS); 13) abstractness index; 14) type-token ratio (TTR); 15) terms count.

Table 2. Text complexity metrics.

Metric/grade	5	6	7	8	9	10	11
Tokens	1008,30	1013,50	1122,60	2402,70	2118,40	3844,00	5280,30
Syllables	2510,10	2505,10	3006,70	6561,60	5956,30	11101,70	15705,50
Sentences	85,10	92,00	107,90	209,30	188,80	297,50	384,10
Words/sent	11,90	11,08	10,52	11,66	11,32	12,98	13,77
syllable/word	2,49	2,47	2,68	2,73	2,81	2,89	2,98
Adj.	129,30	122,80	152,40	360,80	329,60	648,60	954,40
Adverbs	47,90	46,40	43,50	107,00	73,30	138,40	185,20
Pronoun	98,20	99,20	116,20	243,40	219,80	394,00	560,50
Nouns	345,20	330,70	422,20	906,60	846,60	1517,20	2162,50
Verbs	168,90	178,20	184,70	329,90	278,80	468,50	602,40
Frequency	134,31	143,79	128,27	117,97	116,18	113,10	104,65
FKG	6,66	6,26	7,24	7,97	8,30	9,34	10,12
Abstr	-1,69	-1,89	-1,89	-2,00	-1,50	-2,17	-2,05
TTR	0,63	0,63	0,63	0,54	0,52	0,51	0,47
Terms	18,90	18,00	38,10	60,00	124,20	91,50	167,00

5 Research results

The results of the research conducted are presented in Fig,1 – 4.
Fig.1 demonstrates a steady growth of FKG from 6,26 (6th grade) to 10,12 (11th grade). However, as we can see their readability rates are in many cases below the corresponding proficiency level of the target audience.

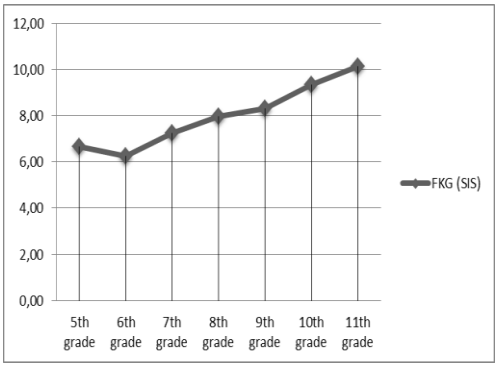

Fig. 1. Flesh-Kincaid Grade (SIS).

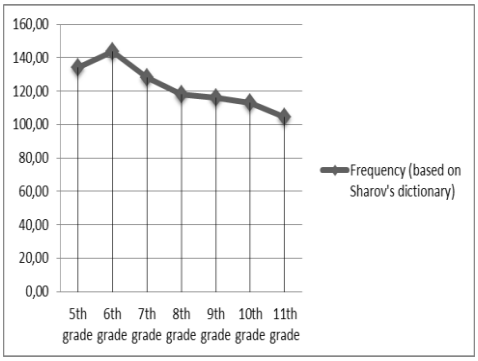

Fig. 2. Frequency rates.

Frequency level of the texts under study decreases as their complexity grows (see Fig. 2). The frequency variable lies within the range from 143,79, the highest in the 6th grade, to 104,65, the lowest in the 11th grade.

Type-token ratio for normalized texts' extracts (1000 words) is within the range from 0,61 (9th grade textbook) to 0,64 (7th and 10th grades textbooks) (see Figure 3). That means that from 61% to 64% of all words used in the texts are unique, which is viewed as an average for this text types [9].

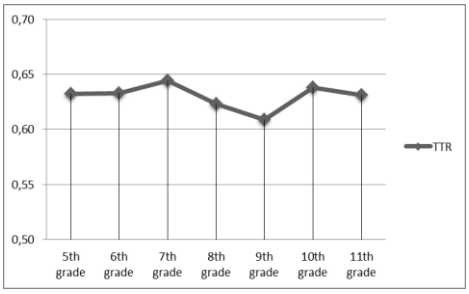

Fig. 3. TTR rates for normalized texts.

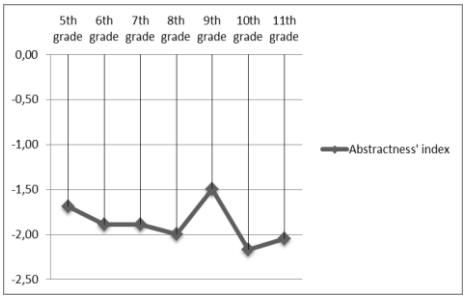

Fig. 4. Abstractness' index rates.

The abstractness indices are also indicative of the overall increasing text complexity (see Figure 4).

The fluctuations of the graph can be explained by the fact that abstractness of the narration is not always achieved exclusively by abstract lexical units. According to the analysis data, the distribution of parts of speech is also changing with the rising of the grades' number. The adjectives' rate in the texts rises from 0,128 for 5th grade to 0,180 for 11th grade. At the same time, average rate of adverbs, on the contrary, decreases from 0,047 for 5th grade textbook to 0,035 for 11th grade textbook. Similarly, the average rate for nouns and pronouns is increasing, when verbs rates are falling, which indicates the increase of abstractness of the texts. Nouns rates are changing from 0,342 (5th grade textbooks) to 0,409 (11th grade textbooks) and pronouns' rates are changing from 0,097

(5th grade textbooks) to 0,106 (11th grade textbooks). Verbs rates are changing from 0,167 to 0,114 respectively. Thus, abstractness increases from grade to grade not only because of the number of abstract words in the text, but also due to the changes in morphological distributions.

The average number of terms per text also tends to increase across grades, though a fluctuation in the 10th grade may testify to the revising character of the reading material in the book (see. Fig 5).The total number of terms ranges between 243 (6th grade) and 2844 (10th grade). The average number of terms gradually increases from 18 (6th grade) to 167 (11th grade). Though the average amount of terms for the 10th grade textbook is relatively lower than in both the 9th and 11th grade textbooks, this also can be explained with specific topics selected by the authors for each textbook. While texts for the 9th grade are focused on politics, the 11th grade texts are centered on economy and social stratification; texts for 10th grade are focused mostly on social mechanisms and human activities. The latter is presented mostly with everyday words, not terminological nomenclature.

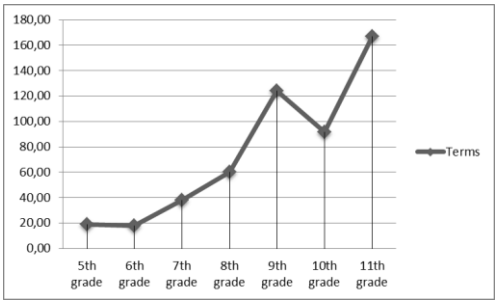

Fig. 5. Terms' count.

We also conducted a mixed analysis of variance (Spearman) to confirm statistical significance of the features estimated (See Table 3).

Table 3. Spearman's values.

Metrics	**P value**
words	0,96
syllables	0,93
sentences	0,96
Words/sent.	0,57
Syllab/word	0,96
Adj.	0,96
Adv.	-0,86
Pron.	0,82
Nouns	0,93
Verbs	-0,96
frequency	-0,96
FKG	0,96
Abstr	-0,61
TTR	-0,86
TTR 1000	-0,25

Spearman Rank Order Correlations are confirmed for the following text features: total amount of words, total amount of syllables, total amount of sentences, average amount of syllables per word, adjectives rate, adverbs rate, pronouns rate, nouns rate, verbs rate, word frequency, FKG, and TTR. The statistically significant metrics have p value <0,05.

6 Conclusion

In this paper we have presented a multi-factor analysis of seven Russian textbooks on Social science. The analysis of 14 text features performed with the help of RusAC, an

online tool designed to assess conventional, morphological and lexical metrics indicated a statistically significant correlation of text complexity with its diversity, frequency, abstractness, number of terminological units . The findings lead us to believe that RusAC is a useful tool for researchers, teachers, and test developers. The results of the research are applicable to match academic texts and test materials with potential target readers. We view identifying syntactic text parameters effecting its complexity as the research perspective.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research, grant No. 19-07-00807.

This work was also supported by the Russian Science Foundation, grant No. 18-18-00436.

This paper is published with financial support of Russian Foundation for Basic Research, project № 20-012-22046.

References

1. A.N. Kolmogorov, *Three approaches on defining the 'amount of information' term* Problems of Information Transmission, V. 1, № 1. 3 – 11 (1965)
2. I.V. Oborneva, Automated complexity evaluation of academic texts based on statistical metrics : diss. … PhD thesis : 13.00.02. 165 p. Moscow (2006)
3. M.I. Solnyshkina, E.V. Harkova, M.B. Kazachkova, *The Structure of Cross-Linguistic Differences: Meaning and Context of 'Readability' and its Russian Equivalent 'Chitabelnost'*, in Journal of Language & Education, 6 (1), 103-119 (2020)
4. H. Nesi, *Information density in a corpus of university student writing*, in Proceedings of the International Conference CORPUS LINGUISTICS 2017. St Petersburg: St. Petersburg State University. pp. 66-71 (2017)
5. D.S. McNamara, A.C. Graesser, P.M. McCarthy & Z. Cai, Automated evaluation of text and discourse with Coh-Metrix. Cambridge, MA: Cambridge University Press. 289 p (2014)
6. M.M. Nevdakh, *Analysis of academic text's informational features with the application of multivariate statistical analysis*, in Applied Computer Science, Publ. by SEI "MFIU Synergy", No. **4**, p. 117 -130 (2008)
7. Y.F. Shpakovsky, Assessment of perception difficulty and an academic text's complexity optimization : (based on Chemistry academic texts): Diss. Abstract … PhD thesis in Philology, Minsk, 21 p (2007)
8. J.P. Kincaid, Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Naval technical training command. Memphis, TN: Naval Air Station. 40 p (1975).
9. V.V. Ivanov, M.I. Solnyshkina, V.D. Solovyev, Efficiency of text readability features in Russian academic texts. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. – Vol.2018-May, Is.17. - pp.267-283 (2018)
10. Corpus of Contemporary American English. URL: https://www.english-corpora.org/coca/ (Access: 21.10.2020)
11. A. F. Zhuravlev, An experience of quantitative and typological investigation of spoken registers, in Varieties of urban spoken language: a collection of research articles – Raznovidnosti gorodskoy ustnoy rechi, Moscow, Nauka, pp. 84–150 (1988)

12. O. B. Sirotinina, Spoken language within the system of functional styles of the Russian literary language: grammar, 3rd edition (Librekom, Moscow, 2009)
13. J. H. Greenberg, "A quantitative approach to the morphological typology of language," International Journal of American Linguistics, vol. **26**(3), pp. 178–194 (1960)
14. D. Biber, University Language: A corpus-based study of spoken and written registers. John Benjamins Publishing Co. Amsterdam. 271 (2006)
15. S.A. Stahl, *Differential word knowledge and reading comprehension*, in Journal of Reading Behavior, No. **15**, pp.33-50 (1983)
16. M. Templin, Certain language skills in children. Minneapolis: University of Minnesota Press (1957)
17. James R. Layton, "A Chart for Computing the Dale-Chall Readability Formula above Fourth Grade Level." Journal of Reading, vol. 24, No. **3**, pp. 239–244. JSTOR (1980) www.jstor.org/stable/40031660. (accessed 21 Oct. 2020).
18. R.C. Anderson, P. Freebody, *Vocabulary knowledge*, in J. Guthrie (Ed.), Comprehension and teaching: Research reviews (pp. 77-117). Newark, DE: International Reading Association (1981)
19. Stenner 2001 Stenner, A. J. The Lexile Framework: A common metric for matching readers and texts. California School Library Journal, **25**(1), 41-42 (2001)
20. S.A. Sharov, O.N. Lyashevskaya, An introduction to frequency dictionary for Russian frequency language. M.: Azbukovik, p. 21 (2009)
21. H. Zinsmeister, S. Birzer, D. Batinić, LeStCor: Levelled Study Corpus of Russian URL: http://lestcor.org/about_the_project/ (Access: 21.10.2020).
22. M. Solnyshkina, V. Solovyev, V. Ivanov & A. Danilov, *Studying Text Complexity in Russian Academic Corpus with Multi-Level Annotation*, in CEUR Workshop Proceedings V. 2303, International Workshop on Computational Models in Language and Speech, CMLS 2019. Kazan, Russia. pp. 1-11 (2019)
23. R.V. Mayer, *Assessing the complexity of academic text in natural sciences*, in Modern Education, No. **4**, pp. 56 – 64 (2016)
24. V. Solovyev, M. Andreeva, M. Solnyshkina, R. Zamaletdinov, A. Danilov and D. Gaynutdinova, «Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach,« 2019 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, pp. 403-408 (2019)
25. Xu, X., Li, J., 2020. Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. PLOS ONE.. doi:10.1371/journal.pone.0232133 June 22, 202013/ 16
26. A.M. Borghi, E. Zarcone, Grounding Abstractness: Abstract Concepts and the Activation of the Mouth. Front. Psychol. 7:1498. (2016)
27. L.N. Bogolyubov, L.F. Ivanova, N.F. Vinogradova, N.I. Gorodetskaya etc., Civics. 5-11 grades: Student's Book, M.: Prosveshcheniye. (2012-2014)
28. RusAC URL: http://tykau.pythonanywhere.com/ (Access: 21.10.2020).