Studying Text Complexity in Russian Academic Corpus with Multi-Level Annotation

Marina Solnyshkina[1], Valery Solovyev[2], Vladimir Ivanov[3], and Andrey Danilov[4]

[1] Kazan Federal University, Kazan, Russia
mesoln@yandex.ru
[2] Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru
[3] Innopolis University, Kazan, Russia
nomemm@gmail.com
[4] Kazan Federal University, Kazan, Russia
tukai@yandex.ru

Abstract. The problem of compiling a large multi-level annotated corpus of Russian academic texts was sparked by the demand to measure complexity (difficulty) of texts assigned to certain grade levels in terms of meeting their cognitive and linguistic needs. For this purpose we produced a corpus of 20 textbooks on Social Studies and History written for Russian secondary and high school students. Measuring text complexity called for linguistic annotations at various language levels including POS-tags, dependencies, word frequencies. Three complexity formulas are compared as an example of using a corpus to study the complexity of texts.

Keywords: multi-level, annotated corpus, Russian academic texts, text complexity, POS-tags, dependencies, word frequencies.

1 Introduction

Automatic multi-level analysis of language implies utilizing a large corpus or a number of corpora which are viewed to be of great value for several research tasks [24]. In this paper we present the ongoing project carried out at Kazan Federal University (Russia) aimed at compiling and annotating a corpus of Russian academic texts.

To the best of our knowledge, no prior corpus-based research has been specifically conducted with the aim of estimating text complexity of Russian educational materials on Social studies. The specific, though sporadic, studies of Russian text readability did not go beyond using mere collections of limited texts of a specific type or genre: fiction (mostly for academic purposes) [17], legal [8], academic texts (chemistry, mathematics, economics) [26, 14, 20, 27]. Most of the research carried out in the area was based on English and other Germanic languages for native and/or non-native readers [3, 6, 10, 16, 22, 23]. The shortage of previous corpus-based research on text complexity of modern Russian academic texts provides a strong justification for pursuing the current study. Our objective is to introduce a multi-level annotated corpus of

Russian academic texts with the ultimate goal of disseminating its potential in Russian discourse research.

It is the authors hope that this proliferation will contribute to detailed examination, identification and measurement of Russian text features. The paper is organized in the following way: In section Background we first give an introduction to the problem of text complexity, we also present the empirical approach to the problem applied in modern multidisciplinary studies. In section Corpus Description we provide information on the corpus collection regarding the type of the texts collected, the size of the corpora and the ultimate goal behind the corpus collection. In same Section we also provide information on preprocessing of the corpus and the multi-level process of the annotation. In Section 4 we briefly describe our experiments conducted with the compiled corpus and in the conclusion section we offer the authors' insights into the areas of possible utilization of the corpus and the perspectives of the work.

2 Background

The earliest studies on readability dating back to late 19th century were mostly aimed at developing readability formulas and utilized a limited number of quantitative features: average sentence length, average word length and word frequency [13, 4, 5]. Given the simplicity of the models and availability of the variables, the readability formulas have been the focus of harsh criticism since they appeared for the first time. Modern advances in natural language processing (NLP) allowed obtaining lexical and syntactic features of a text, as well as automatically train readability models using machine-learning techniques [23]. Text readability studies based of ngram models were successfully conducted by American researchers [9] and later on, based on syntax simplicity/complexity, discourse characteristics (narrativity, abstractness, referential and deep cohesion, etc., extended to assessing a particular text profile and its target audience see [16].

Modern researchers of English develop NLP tools of new generation providing accurate and valid analyses on various dimensions of texts and measure complex discourse constructs using surface-level linguistic features such as text structure, vocabulary or the number of unique words in a text, givenness or the number of determiners and demonstratives in a text, anaphor or the number of all pronouns lexical diversity, connectives and conjuncts which together with anaphor are indicators of text coherence, future as an indicator for situational cohesion, syntactic complexity measured through the number of words per sentence, and the number of negations [7]. Based on systemic language parameters text features are to be specified for one language only. Thus, every modern NLP tool as well as a readability formula are applicable to one language in particular. E.g. parameters measured for English cannot be applied to estimating Russian texts complexity as Germanic languages have limited morphology in comparison with Russian [23] and all text features need to be validated in a corpus of a considerable size.

Owing to the existing lack of available corpora Russian discourse studies at the moment are viewed as underdeveloped [25]. Russian academic texts began being used in readability studies only in 1970-s [21], but with a short break during 1990-s the studies in the area were quite extensive. Nowadays researchers view the following text readability features as cognitively significant: number of syllables, number of words, sentence count, average sentence length, abstract words count, homonyms counts, polysemous words counts, technical terms counts, etc. [20]. Ivanov V.V. tested correlations of 49 factors, among which the strongest correlations are identified for the percentage of short adjectives, the percentage of finite verb form, the Flesch-Kincaid Grade Level Score, the Flesch Reading Ease Score [13], the Coleman and Liau index, average number of words per sentence, percentage of complex sentences, percentage of compound sentences, percentage of abstract words [11]. Karpov N. et al. [26] conducted a series of experiments utilizing a number of machine-learning models to automatically rank Russian texts based on their complexity. For the purpose the authors compiled two subcorpora: (1) a corpus of texts generated by teachers for learners of Russian as a foreign language (at http://texts.cie.ru); (2) 50 original news articles for native readers. They assessed 25 text parameters of each text in the corpora, such as sentence length, word length, vocabulary, parts of speech classification. For the last fifteen years, readability of Russian academic texts has been actively discussed at conferences in Russia and abroad as well as in numerous publications [21] but readability studies are still far from being systematic and irregularities in reporting make it difficult to draw firm conclusions [23] mostly due to corpora limitations.

The problem of defining Russian text complexity features can be studied on a massive corpus containing academic texts used in modern schools. Unfortunately neither Russian National Corpus nor Corpora of Russian (http://web-corpora.net/?l=en) though being large and widely used in studies of lexical, syntactic and discourse features cannot be used for the purposes of our research based on the fact that they do not provide access to modern Russian academic texts.

3 Corpus Description

For the purposes of the study we compiled a corpus of two sets of textbooks on Social Studies and History written for Russian secondary and high school students. The total size of the corpus of 20 textbooks is more than 1 million tokens.

The first collection of 14 texts from textbooks on Social Studies by Bogolubov L. N. marked “BOG” by Nikitin A.F. marked “NIK” aimed for 5 – 11 Grade Levels. In our study, Grade Levels means the class number for which the textbook is intended. It was selected to teach the predictive model and define independent variables of the text variation. The second collection of 6 texts from textbooks on History by different authors aimed for 10 – 11 Grade Levels. Both sets of textbooks are from the “Federal List of Textbooks Recommended by the Ministry of Education and Science of the Russian Federation to Use in Secondary and High Schools”.

To ensure reproducibility of results, we uploaded the corpus on a website thus providing its availability online. Note, however, that the published texts contain shuffled order of sentences. The sizes of BOG and NIK subcollections of texts are presented in Table 1.

Table 1. Properties of the preprocessed corpus on Social Studies.

	Tokens		Sentences		Words per sentence	
Grade	**BOG**	**NIK**	**BOG**	**NIK**	**BOG**	**NIK**
5-th	--	17,221	--	1,499	--	11.49
6-th	16,467	16,475	1,273	1,197	12.94	13.76
7-th	23,069	22,924	1,671	1,675	13.81	13.69
8-th	49,796	40,053	3,181	2,889	15.65	13.86
9-th	42,305	43,404	2,584	2,792	16.37	15.55
10-th	75,182	39,183	4,468	2,468	16.83	15.88
10-th*	98,034	--	5,798	--	16.91	--
11-th	--	38,869	--	2,270	--	17.12
11-th*	100,800	--	6,004	--	16.79	--

In the Table 1 star sign (*) denotes advanced versions of books for the corresponding grade; sign ‘-‘ denotes absence of a textbook for the corresponding grade.

Data on the collection of books on history is presented in Table 2. The first column lists textbook authors and the class number.

Table 2. Properties of the preprocessed corpus on History.

Author / Grade	Tokens	Sentences	Words per sentence
Soboleva / 10-th	81544	7116	11.46
Volobuyev 10-th	40949	3676	11.14
Guryanov / 11-th	100331	9393	10.68
Petrov / 11-th	85409	8536	10.01
Plenko / 11-th	63804	5292	12.06
Ponomarev / 11-th	44833	4003	11.2

3.1 Corpus Preprocessing

For the convenience, we have preprocessed all texts from the corpus in the same way. Common preprocessing included tokenization and splitting text into sentences. During the preprocessing step we excluded all extremely long sentences (longer than 120 words) as well as too short sentences (shorter than 5 words) which we consider outliers. Clearly, such sentences can be not outliers at all in another domain, but for the case of school textbooks on Social Studies sentences shorter than 5 words are outliers. Sentence and word-level properties of the preprocessed dataset are presented in Tables 1 and 2.

Extremely short sentences mostly appear as names of chapters and sections of the books or as a result of incorrect sentence splitting. We omit those sentences, because the average sentence length is a very important feature in text complexity assessment and hence should not be biased due to splitting errors. At the same time sentences with five to seven words in Russian can still be viewed as short sentences, because the average sentence length (in our corpus) is higher than ten.

Table 1 demonstrates that values of Word per sentence (ASL) as it is generally expected, increase with the grades.

3.2 Multi-level Annotations in Corpus

All annotations in the corpus are performed on three levels: text-level, sentence- level and word-level. At the text-level meta-annotations refer to a number of sentences and a set of tokens, an author and a grade-level of a given text. At the word-level we have part-of-speech tag for each word. POS-tagging has been performed with the use of the TreeTagger for Russian (http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/). The tagset is available from the website of the project. As example we provide distribution of major PoS-tags among texts on Social Studies, Table 3. We also annotate each lemma in the corpus with its relative frequency measured in the large corpus of Russian texts, Russian National Corpus.

At the sentence-level the corpus contains annotations of sentence boundaries, the tokens are assigned to sentences as well as a dependency tree of each sentence. For dependency parsing we use pretrained neural models (https://github.com/MANASLU8/ CoreNLPRusModels) for Stanford Dependency Parser for Russian (https://nlp.stanford.edu/software/stanford-dependencies.shtml). Finally, at the moment, we are adding semantic annotations to the corpus. The semantic annotations are based on the very large Russian Thesaurus (RuThes) [28]. Concepts of the RuThes are mapped to the Wordnet thesaurus that allows to process textual content at semantic level.

Table 3. Unique words in each of four PoS-tags that appear in textbooks; normalized by 1000 words.

	NOUN		**VERBS**		**ADJECTIVES**		**ADVERBS**	
Grade	**BOG**	**NIK**	**BOG**	**NIK**	**BOG**	**NIK**	**BOG**	**NIK**
5-th	--	69.7	--	48.6	--	77.6	--	10.7
6-th	69.1	69.4	48.8	42.2	81.2	96.6	11	11.3
7-th	71.4	63.6	39.5	37.8	100.3	90.8	9.3	9.9
8-th	43	53.5	22.2	27.9	111.3	114.6	6.1	7
9-th	38.3	46.5	21.3	24.2	119.4	114.8	5.5	6.6
10-th	33.5	50.1	17.3	22.8	124.5	130.6	4.4	6.6
10-th*	28.6	--	14.7	--	122.3	--	4	--
11-th	--	43.4	--	23	--	124.2	--	6.2
11-th*	30.7	--	14	--	143.7	--	3.9	--

4 Studies of Text Readability and Complexity

First of all, the corpus can be used to adjust readability formulas in Russian. Second, even very simple statistics provided in the Table 3 can be useful in text complexity studies. For example, one can see that average number of unique adjectives grow when grade level increases. At the same time average number of adverbs (as well as verbs) decreases. Both observations correspond with idea that texts become more descriptive. However, with assistance of the data it is possible to measure the correlation.

In this study, 3 formulas (our formulas [29], Matskovskiy Readability Formula [30] and Oborneva's Readability Formula [17]) were applied to 5 Social Studies and 7 History textbooks for grades 10 – 11. In the formulas below, GL denote the grade level.

In paper [29] we provided readability formula GL = 0.36ASL + 5.76ASW – 11.97, where ASL and ASW means average of words per sentence and means average of syllables per word respectively. Below, this formula is labeled RRF. In [30] Matskovskiy M.S. computed the first readability formula for the Russian language: GL = 0.62ASL + 0.123X + 0.051, where X is the percentage of three syllable words in the text. In [17] Oboroneva I. introduced readability formula readability formula GL = 0.5ASL + 8.4 ASW – 15.59.

In an attempt to verify the features defined as contributing to text readability but not measured by the existing readability formulas, we compared the 11 texts under study in order to see what metrics better correlate with the grade level. The data are presented in table 4.

The Fig. 1 below shows, that Oboroneva's formula positioned them as textbook comprehensible only by people with at least 16 – 17 years of formal schooling, i.e. with Bachelor or Master's Degree. It is clear from the table that grade level predictions based upon the equation of regression of Oborneva I. do not coincide with the actual grade levels, the difference is marked in 6 years in the case of textbooks on History. As for Matskovskiy's Readability formula which was initially developed to compute readability of media texts only, it proves to be quite reliable in assessing readability of academic texts also (compare columns 'Grade' and 'Matskovskiy' in Table 4).

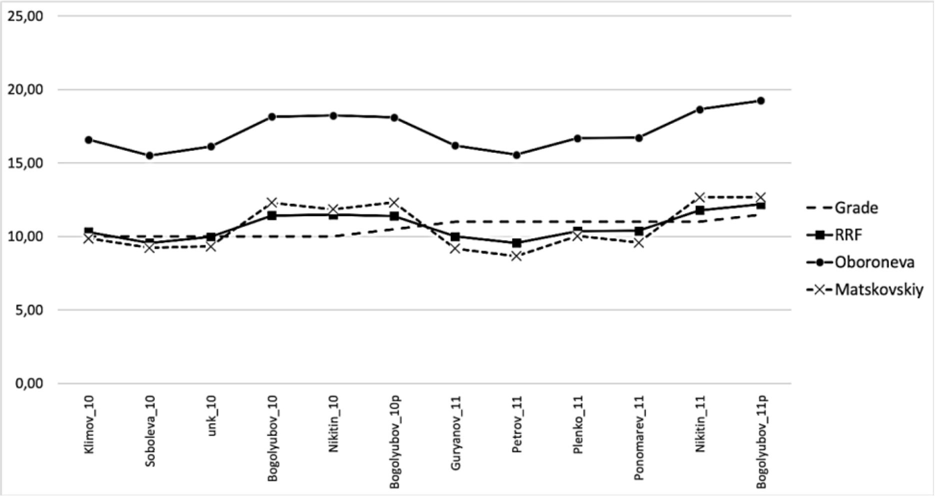

Fig. 1. Predictions of grade levels. Ground truth is represented with a dashed line.

Table 4. Comparison of three readability formulas using Social Science and History textbooks.

Book	ASL	ASW	Fraction of 3-sylables words	TRUE_GRADE	RRF	Oboroneva	Matskovskiy
Guryanov_11	11.14	3.12	0.18	11.00	10.01	16.19	9.19
Klimov_10	12.45	3.09	0.17	10.00	10.31	16.60	9.88
Petrov_11	10.43	3.09	0.18	11.00	9.57	15.56	8.67
Plenko_11	12.52	3.10	0.18	11.00	10.38	16.69	10.03
Ponomarev_11	11.64	3.15	0.19	11.00	10.39	16.73	9.59
Soboleva_10	11.75	3.00	0.15	10.00	9.57	15.53	9.23
BOG_10	15.88	3.07	0.20	10.00	11.44	18.15	12.31
BOG_10*	16.06	3.06	0.19	10.50	11.41	18.11	12.33
BOG_11*	16.03	3.19	0.22	11.50	12.19	19.25	12.68
NIK_10	15.06	3.13	0.20	10.00	11.49	18.24	11.85
NIK_11	16.19	3.11	0.21	11.00	11.79	18.66	12.68

5 Discussion

Thus, there are two reasons which make future research into Russian texts readability relevant. First, the recent reports from educators call for improving reading comprehension in secondary and high schools throughout the country [2, 1]. Researchers also testify to Russian students lack of interest in reading caused by inappropriate selection of educational materials [20]. The Corpus is a valuable instrument for discourse studies as its data and flexible search system provide a solid foundation for comparative research of modern Russian texts and enables deep insights into patterns and dependencies of different text features. The Corpus is also viewed by the authors as a powerful tool for discovering new aspects and regularities of Russian discourse.

Acknowledgements

This research was financially supported by the Russian Science Foundation, grant #18-18-00436, the Russian Government Program of Competitive Growth of Kazan Federal University, and the subsidy for the state assignment in the sphere of scientific activity, grant agreement 34.5517.2017/6.7. The Russian Academic Corpus (section 3, 3.1 in the paper) was created without support from the Russian Science Foundation.

References

1. Kompetentnostnyy podkhod v vysshem professionalnom obrazovanii (pod redaktciyey A.A. Orlova, V.V. Gracheva), Tula (2012).
2. Berezhkovskaja E. Problema psihologicheskoj negotovnosti k polucheniju vysshego obrazovanija u studentov mladshih kursov. M.: Prospec. (2017).
3. Britton, B.K., & Gulgoz, S. Using Kintsch’s computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, pp. 329-404 (1991).
4. Chall J., Dale E. Readability revisited: The new Dale-Chall readability formula. Brookline Books (1995).
5. Coleman M., Liau T. L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60:283-284 (1975).
6. Cornoldi, C., & Oakhill, J. (Eds.). Reading comprehension difficulties: Processes and intervention. Hillsdale, NJ: Erlbaum (1996).
7. Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. Analyzing discourse processing using a simple natural language processing tool (SiNLP). Discourse Processes, 51, pp. 511-534 (2014).
8. Dzmitryieva A. Iskusstvo yuridicheskogo pis’ma: kolichestvennyy analiz resheniy Konstitutsionnogo Suda Rossiyskoy Federatsii [The art of legal writing: a quantitative analysis of the Russian Constitutional Court rulings]. Sravnitel’noe konstitutsionnoe obozrenie, no.3, pp. 125-133. (In Russian) (2017).
9. Heilman M., Thompson K. C., Callan J., and Eskenazi M. Combining lexical and grammatical features to improve readability measures for first and second language texts. In

Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-07), pp. 460-467, Rochester, New York (2007).

10. Jackson, G. T., Guess, R. H., & McNamara, D. S. Assessing cognitively complex strategy use in an untrained domain. In N. A. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society. pp. 2164-2169. Amsterdam, The Netherlands: Cognitive Science Society (2009).
11. Ivanov V. K voprocu o vozmonosti ispolzovanija lingvisticeskix xarakteristik slonosti teksta pri issledovanii okulomotornoj aktivnosti pri ctenii u podrostkov [Toward using linguistic profiles of text complexity for research of oculomotor activity during reading by teenagers]. Novye issledovanija [New studies], 34(1):4250 (2013).
12. Karpov N., Baranova J., and Vitugin F. Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (2014).
13. Kincaid J. P., Fishburne R. P. Jr., Rogers R. L., and Chissom B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN (1975).
14. Krioni N., Nikin A., and Fillipova A. Avtomatizirovannaja sistema analiza slozhnosti uchebnyh tekstov [The automated system of the analysis of educational texts complexity]. Vestnik UGATU (Ufa), 11(1):28 (2008).
15. McNamara, D.S. Reading both high and low coherence texts: Effects of text sequence and prior knowledge. Canadian Journal of Experimental Psychology, 55, pp. 51-62 (2001).
16. McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition & Instruction, 14, pp. 1-43 (1996).
17. Obobroneva I. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov. M.: RAS Institut soderzhaniya i metodov obucheniya (2006).
18. Okladnikova S. Modelkompleksnoj ocenki citabelnosti testovyx materialov na etape razrabotki [A model of multidimensional evaluation of the readability of test materials at the development stage]. Prikaspijskij journal: upravlenie i vysokie texnologii, 3:6371 (2010).
19. Popova Ja.I., Shishkevich E.V. Standartizacija uchebnoj literatury srednej shkoly po kriteriju udobochitaemosti In Sevastopol’skij nacional’nyj universitet jadernoj jenergii i promyshlennosti. Nauchnye vedomosti BelGU. Ser. Gumanitarnye nauki. 12. No. 6. pp. 142-147 (2010).
20. Shpakovskiy Y. et al. Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta. PhD thesis (2007).
21. Solnyshkina M., Harkova E, and Kiselnikov A. Comparative coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. English Language Teaching, 7(12):65 (2014).
22. Ozuru, Y., Rowe, M., OReilly, T., & McNamara, D. S. Wheres the difficulty in standardized reading tests: The passage or the question? Behavior Research Methods, 40, pp. 1001-1015 (2008).
23. Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, CA: 16 June 2016. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp.289-300 (2016).
24. Sinclair, J. Corpus Evidence in Language Description, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) Teaching and Language Corpora. London/New York: Longman, pp. 27-39 (1997).

25. Ivanov V.V., Solnyshkina M.I., & Solovyev V.D. Efficiency of text readability features in Russian academic texts. In Computational Linguistics and Intellectual Technologies, V.17, pp. 277–287 (2018).
26. Karpov N., Baranova J., and Vitugin F.. Single-sentence readability prediction in Russian. In International Conference on Analysis of Images, Social Networks and Texts, pp. 91-100. Springer (2014).
27. Ustinova, L. V., Fazylova L. S. Avtomatizacija ocenki slozhnosti uchebnyh tekstov na osnove statisticheskih parametrov. Vestnik Karagand. un-ta. Ser. Matematika. 1. pp. 96-103 (2014).
28. Loukachevitch, N. V., Lashevich, G., Gerasimova, A. A., Ivanov, V. V., & Dobrov, B. V. Creating Russian Wordnet by conversion. Kompjuternaja Lingvistika i Intellektualnye Tehnologii, 15, pp. 405-415 (2016).
29. Solovyev V., Ivanov V., & Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics, Journal of Intelligent & Fuzzy Systems, vol.34, no.5, pp. 3049-3058 (2018).
30. Matskovskiy, M.S. Problemy chitabelnosti pechatnogo materiala [Problemsof printed material readability]. In: Dridze, T.M. & Leontev, A.A. (eds) Smyslovoe vospriyatie rechevogo soobshcheniya v usloviyakh massovoy kommunikatsii [Semantic perception of verbal communication in the context of mass communication]. Moscow: Nauka (1976).