Readability Formula for Russian Texts: A Modified Version

Marina Solnyshkina[1(✉)], Vladimir Ivanov[2], and Valery Solovyev[1]

[1] Kazan Federal University, 18, Kremlyovskaya Street, Kazan, Russia
maki.solovyev@mail.ru, mesoln@yandex.ru
[2] Innopolis University, 1, Universitetskaya Street, Innopolis, Russia
v.ivanov@innopolis.ru

Abstract. The authors of the article offer new readability formulas for academic texts which provide a comparatively higher degree of accuracy than other Russian readability formulas. The results achieved are due to using original syntactic, lexical and frequency metrics ignored in previous research on Russian readability. The methods applied by the authors include Ridge and linear regression. The new readability formulas were computed on the Corpus of secondary school textbooks on Social Studies and then validated on the Corpus with the total size of 1 mln. tokens. The perspectives of the research lie in further modification of the formula for texts of various genres.

Keywords: Text readability formula · Academic texts
Russian language

1 Introduction

Modern readability studies have lately become interdisciplinary and continue engaging more researchers all over the world. The main reason for this is obvious: the increased number of failures to reach readers with a printed (electronic) text. As target audiences of companies, authorities and organizations become more receptive to audio and visual signals, quality requirements to printed texts grow exponentially and the task to enhance reading outcomes is becoming crucial. As the world itself and human activities are becoming more sophisticated the task to create comprehensible texts has become even more difficult: writers have to use more elaborate words and longer constructions to describe the world. One of the areas where improving the quality of reading materials is especially important and even indispensable is education. It is also true about Russia: after all social and political changes in 1990-s and 2000-s, the country is experiencing a real educational crisis [15,16]. Analysis also proves that the complicated language of school textbooks remains one of the burning issues in Russia today[1]. The quality of educational (printed) material depends largely on the skill of the

[1] https://www.hse.ru/news/122263399.html.

I. Batyrshin et al. (Eds.): MICAI 2018, LNAI 11289, pp. 132–145, 2018.
https://doi.org/10.1007/978-3-030-04497-8_11

author, expertise and experience of the editor. At the moment the Ministry of Education is expected to develop new standards of school textbooks expertise[2]. In this regard, the creation of reliable and universally accepted methods of automated verification of text complexity and readability is an urgent task. Another aspect in readability studies is an increasing need for leveled texts, i.e. texts profiled for different readers, in various areas. When performed manually it is time consuming, resource intensive and extremely costly. Therefore an automated tool performing the same functions is very desirable. In this paper we present the research aimed at measuring text complexity of Russian academic texts and offer results of our studies on various metrics of academic texts which successfully allow to profile a text for potential readers' linguistic abilities correlated with a particular grade level.

The current study was conducted to answer two research questions:

(1) How do 'classical' readability metrics work in the corpus of Russian academic texts?
(2) How do the new metrics, offered by the authors of the article, correlate with readability of Russian academic texts?

2 Related Work

The history of Text readability studies are the research aimed at extending the list of text metrics correlating with text complexity. It is also a history of criticism of readability formulas and doubts that the ideal formula profiling the text and the reader may never be derived. In the middle of the last century [5] and [2] proved that the correlation between factors that affect text comprehension is so great that only a few are enough to measure text complexity, but in their search for discovering correlations between text metrics and comprehension levels, researchers only increased the number of metrics: all over the world there have been conducted thousands of experiments with over 200 different text features. At present, there are over two hundred formulas of readability: Gunning fog index, Coleman Liau index, Flesch Kincaid Grade Level etc. for texts in many languages: English, French, German, Dutch, Swedish, Russian and other languages. Below we offer a brief history of views on each of the variable used by the authors in the current research.

2.1 Average Length of Sentences and Words, ASL, ASW

The very first two metrics introduced by Rudolf Flesch and Kincaid in 1948, i.e. ASL (average sentence length (in words)) and ASW (average word length in syllables) [2], are nowadays core components in the majority of readability formulas (see [1]). E.g. English Flesch reading ease, FRE = 206,835 – 1,015 ASL – 84, 6 ASW. It distributes ASL and ASW within the readability range as follows:

[2] https://www.kommersant.ru/doc/3614360.

- 100: The text is very easy to read. The average sentence length is 12 or fewer words. There are no words longer than 2 syllables.
- 65: The text is written in plain English. The average sentence length is from 15 to 20 words. An average word consists of 2 syllables.
- 30: The text is rather difficult to read. Sentences contain up to 25 words. Words are disyllabic.
- 0: The text is very difficult to read. An average sentence is 37 words long. An average word has more than 2 syllables.

2.2 Percentage of Long Words in Text, PLW

PLW has been calculated differently in various studies depending on the unit of measurement: either it is a number of characters in a word (letters) or syllables. E.g. one of the variables in Carl-Hugo Björnsson (1968) readability formula for Swedish known as Lix, is the percentage of long words, i.e. words of more than six letters: Lix = WL + SL, where, WL = percentage of words of more than six letters; SL = average number of words per sentence. But as the parameter is different in languages of different morphological types: in analytical languages words are shorter as they have fewer affixes (e.g., in English or in Spanish), while in synthetic languages with their highly developed system of morphemes (e.g. German, Ukranian) words are typically longer. Based on the discriminant analysis of 49 text variables in Russian academic texts [4] arrived at the conclusion that the best correlation (among others) between text metrics and readability are (1) the percentage of words of 11 letters and more and (2) the percentage of words of 13 letters and more.

In many readability formulas for European languages the word length is measured in syllables and researchers use a variable of the so-called 'complex words' which are typically defined as polysyllabic or multisyllabic words. In English, Spanish and French a polysyllabic word is a word made up of three or more syllables. For instance, the SMOG Readability Formula for English computes readability in the following way: SMOG grade = 3 + Square Root of Polysyllable Count (McLaughlin, 1969). Matskovskii [7] proposes to calculate readability of Russian texts based on the percentage of words of 4 or more syllables: Russian text readability = (0.62 × ASL) + (0.123 × percentage of words in the text of 3 or more syllables) + 0.051. Another indirect reason to consider 4, not 3 syllable words as 'complex' in Russian is proposed by I.V. Oborneva [8] who proved that on average an English word (2.97 syllables) is one syllable shorter than a Russian word (3.29 syllables). Thus, in our studies we also observe correlation of 4-syllable words with text readability.

2.3 Type Token Ratio, TTR

For years researchers have been offering different views and developing tools to measure 'lexical diversity' or 'lexical density' of a text. One of the metrics used in many studies is the so-called 'lexical richness of the text', i.e. type-token ratio, the ratio of types of words (unique words) to the total number of

words (tokens) in the text [11]. However, later it was proved to be very sensitive to the size of the text. Thus, a number of reformulation of TTR have been offered since that. For example, for Swedish texts a common lexical density measure is OVIX, a word variation index (see [12]), calculated with the help of the natural logarithm. Cvrček and Chlumská [13] introduced two more metrics: (1) standardized (normalized) TTR, the sTTR, which is calculated for every thousand words and (2) zTTR, calculated as the ration of the observed TTR and the reference TTR. Unfortunately none of TTR metrics performs accurately and stable enough in a discourse (see [9,14]).

3 Corpus Description

For the research purposes we compiled a Corpus of two collections of texts (see [17]). The first collection of 7 texts derived as a result of OCR and postprocessing of textbooks on Social Studies by L.N. Bogolubov is marked in the Corpus as "BOG". The textbooks used cover the range of 6 – 11 Grade Levels of secondary and high schools in the Russian Federation. The second collection of 7 texts of textbooks on Social Studies by A.F. Nikitin marked "NIK" in the Corpus comprises the Grade levels of 5 – 11. 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"[3]. In the study we refer to the joined collection of textbooks as Russian Readability Corpus (RRC). To ensure reproducibility of results, we uploaded the corpus on the website[4], but for copyright purposes we had to shuffle the order of sentences in the uploaded texts of the Corpus. This shuffling, indeed, does not affect the values of features under study as they do not depend on sentence order. Table 1 below provides a numerical description of the RRC.

Table 1. Corpus parameters

Document	Tokens	Sentences	Syllables	Document	Tokens	Sentences	Syllables
1	19,412	1,482	39,964	8	23,019	2,275	42,512
2	26,72	1,907	60,977	9	19,619	1,399	40,739
3	58,391	3,441	138,509	10	28,349	2,009	59,239
4	50,828	2,977	121,407	11	48,844	3,614	108,523
5	90,12	5,051	218,984	12	53,273	3,389	123,358
6	117,251	6,25	287,068	13	47,267	2,711	112,487
7	116,12	6,326	299,019	14	45,943	2,549	111,995

[3] http://www.fpu.edu.ru/fpu/.
[4] http://kpfu.ru/slozhnost-tekstov-304364.html.

4 Processing of Texts in the Corpus

For the sake of convenience, we have processed all the texts of the Corpus in the same way. The preprocessing included tokenization, splitting text into sentences (or rather 'sequences of text separated by periods') and part-of-speech tagging (using the TreeTagger for Russian[5]). During the preprocessing stage we categorized two types of outliers and excluded (1) excessively long sequences of words (longer than 120 words) and (2) sequences shorter than five words. The long sequences proved to be either quotations from legal acts, e.g. Constitution of the Russian Federation, or lists generated by the textbook authors to save space in the textbooks. Short sequences of words (separated from nearby sentences with periods) appear to be either names of chapters and sections of books or results of incorrect sentence splitting. The quotations from legal acts were excluded on the presumption that they do not present academic discourse patterns but are typical of legal discourse. The lists and titles with grammatical structures of Nominal Phrases are viewed as outliers and were also excluded from the Corpus as they lack complete grammatical structure of a sentence. The exclusion of those 'sentences' from the Corpus is viewed by the authors as an important preprocessing stage as their metrics may, to a greater extend, influence the average sentence length used in all existing readability formulas. The research shows, that in the Russian discourse, the average sentence length depending on the genre and type of a text varies from ten to 30.

To ensure reproducibility of results, we uploaded the corpus on a website thus providing its availability online[6]. As textbooks we use are protected by copyright rules we shuffled the order of sentences in the Corpus thus limiting the possibility to use the texts in the Corpus for research only.

4.1 Sampling from the Corpus

The RRC contains 14 documents and thus by no means presents a representative sample of the population of all the school textbooks under study. Building a larger corpus is difficult, as it would violate some of the key principles: we either use new texts from different domains, or texts will come from different authors with different writing styles. Both cases of this kind may add noise to the dataset.

In order to overcome the issue of collection size, the following procedure of sampling from the corpus is suggested. The first issue in sampling from the corpus, as Biber (1990) puts it, "concerns the sampling of texts: how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions?". Having compared the internal variations of the two texts in the corpus, Biber (1990) concludes that text samples of 1000 words are representative for the text categories under study.

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
[6] http://kpfu.ru/portal/docs/F1554781210/shuffled.zip.

Both dependent variables (such as readability value) and independent variables (such as ASL) measured for a sample of a text should be close to the complexity value of the whole text. This assumption means, that starting from a certain subset (or sample) of sentences, text complexity of the sample will be almost the same as text complexity of the whole document from corpus. The sample corpus size was set to 5000 tokens. However, preserving order of tokens and sentences is important, otherwise the sampled texts will be less natural, even though they could carry the main features of the documents from the corpus. Thus, we sample 5000 token sequences from each document. We calculate features for readability analysis using the described sampling technique.

5 Text Features for Readability Analysis

In this study we have explored an extended feature set for text readability modeling:

- average number of words per sentence (ASL);
- average number of syllables per word (ASW);
- percentage of long words in text (PLW);
- type-token ratio (TTR), in four variants:
  - type-token ratio for all tokens (TTR);
  - type-token ratio for Nouns only (TTR_N);
  - type-token ratio for Verbs only (TTR_V);
  - type-token ratio for Adjectives only (TTR_A);
- TTR-based ratio of (TTR_A + TTR_N)/TTR_V; at a later stage we denote this feature as(NAV).;
- a relation between number of unique words in text: (number of unique Adjectives + number of unique Nouns)/(number of unique Verbs); at a later stage we denote this feature (UNAV).

The target feature for prediction is the grade level of the text. The feature is represented as real number. The relevance of the set of the above listed features for English text complexity modeling was studied by McNamara et al. in [21]. The features were calculated in the following way.

$$ASL = \frac{\text{total words}}{\text{total sentences}}$$

The ASL metric, i.e. the average sentence length (in words), is a core component in the majority of all readability formulas (see [1]).

$$ASW = \frac{\text{total syllables}}{\text{total words}}$$

The number of syllables was calculated as the number of vowels in a word. In Russian this heuristic gives appropriately good results. If a word does not contain vowels (e.g. some prepositions) it is attached to the adjacent word with vowels.

$$PLW = \frac{\text{total words with 4 and more syllables}}{\text{total words}}$$

$$TTR = \frac{\text{total unique tokens}}{\text{total tokens}}$$

Table 2. Features calculated in Russian readability corpus

Document	ASL	ASW	PLW	TTR	TTR_N	TTR_V	TTR_A	NAV	UNAV	GRADE
1	13.1	2.06	0.17	0.36	0.42	0.52	0.57	1.91	2.19	6
2	14.01	2.28	0.22	0.37	0.38	0.54	0.52	1.67	2.93	7
3	16.97	2.37	0.25	0.35	0.36	0.54	0.45	1.48	2.81	8
4	17.07	2.39	0.26	0.32	0.32	0.52	0.4	1.39	2.81	9
5	17.84	2.43	0.26	0.34	0.35	0.55	0.42	1.39	3.66	10
6	18.76	2.45	0.26	0.34	0.35	0.55	0.45	1.46	3.06	10.5
7	18.36	2.58	0.29	0.35	0.35	0.57	0.4	1.33	3.78	11.5
8	10.12	1.85	0.1	0.37	0.42	0.53	0.54	1.80	2.77	5
9	14.02	2.08	0.18	0.37	0.4	0.56	0.52	1.66	2.55	6
10	14.11	2.09	0.18	0.38	0.4	0.57	0.56	1.68	2.82	7
11	13.52	2.22	0.23	0.38	0.39	0.58	0.47	1.49	2.84	8
12	15.72	2.32	0.25	0.36	0.36	0.56	0.46	1.47	3.30	9
13	17.44	2.38	0.26	0.37	0.38	0.58	0.46	1.45	4.47	10
14	18.02	2.44	0.27	0.32	0.32	0.53	0.39	1.35	3.55	11

6 Model Selection

The problem of readability prediction can be formulated as a regression model. Indeed, most popular readability formulas are simple linear models that use one or several text features. In this section we analyze several regression models for readability prediction. As candidate regression models we consider a simple linear regression, a polynomial regression (i.e. a case when regression is built with the use of polynomial features). Additionally, we measure relative importance of the selected features. To this end, we use a feature selection technique that is based on the F-test. Finally, we apply regularization in order to find a subset of features most useful in prediction. In the end of the section we provide new formulas for readability prediction along with their performance evaluation based on the mean squared error (MSE) measure.

6.1 Linear Models and Feature Selection

Univariate Linear Regression Tests. First, we select features based on the Pearson correlation coefficient between each parameter and the grade level. We use the whole dataset from Table 2 as input data and select top-K (K = 1..8) features. The results of the experiment are presented in form of the ordered list of feature tuples.

- ASW, p-value = $8.76 \cdot 10^{-7}$
- ASL, p-value = $2.71 \cdot 10^{-6}$
- PLW, p-value = $3.79 \cdot 10^{-6}$
- NAV, p-value = $1.03 \cdot 10^{-5}$
- TTR_A, p-value = $4.76 \cdot 10^{-5}$
- TTR_N, p-value = $3.07 \cdot 10^{-4}$
- UNAV, p-value = $1.95 \cdot 10^{-3}$
- TTR, p-value = 0.0215
- TTR_V, p-value = 0.379

The second approach to feature selection is the recursive feature elimination. Method starts with full set of features and tries to eliminate features one by one. The result of this method is elimination of two features: TTR and TTR_V. In comparison to the previous technique it confirms that those two features can be eliminated from further investigation.

6.2 Building a Linear Model with Regularization

After elimination of "TTR" and "TTR_V" we can build a simpler and robust linear model for prediction. The common approach to build such formula is to regularize the coefficients in linear regression. This can be done in several ways, including Lasso (L1) regularization, Ridge regression and Elastic-Net regularization [22]. In Table 3 we provide results of building linear regression with the three approaches to regularization. The higher the absolute value of a coefficient, the more important corresponding feature is.

Table 3. Building a linear readability models with regularization

Regularization type	PLW	UNAV	TTR_N	TTR_A	ASL	ASW	NAV
Lasso	0.00	0.78	0.00	0.00	0.32	2.45	−2.10
Ridge	0.35	0.82	−0.35	−0.95	0.36	1.84	−1.69
Elastic-Net	0.00	0.83	0.00	−0.15	0.44	1.23	−1.46

Table 3 shows that ASL and ASW are useful features as well as NAV and UNAV. Corresponding values in a column (weights of a feature) for each feature are close to each other in different regularization techniques. These features could be a basis for a more robust readability formula.

The resulting linear formulas with 2, 3 and 4 features are presented in Table 4. The table contains coefficients (weighs) of corresponding features. In order to measure performance of models, we use well-known measures: mean squared error (MSE) and mean absolute error (MAE).

$$MSE = \frac{1}{N} \sum (Y_{predicted} - Y_{observed})^2$$

Table 4. Coefficients of linear models.

Model name	ASL	ASW	UNAV	NAV	Intercept
M_0	0.28	6.2	-	-	−10.12
M_1	0.24	3.48	0.75	−2.38	−1.87
M_2	0.26	3.55	-	−3.74	2.07
M_3	0.25	4.98	0.89	-	−9.53
M_4	-	-	0.89	−8.42	18.6

$$MAE = \frac{1}{N} \sum |Y_{predicted} - Y_{observed}|$$

Finally, we have run a brute-force search for a formula that can give lowest MSE in training set. The following formulas with 4 and 5 regressors are provided below:

$$F_4 = 0.83UNAV - 6.73TTR\_A + 0.24ASL + 3.36ASW - 2.41$$

$$F_5 = 0.81UNAV - 5.47TTR\_A + 0.24ASL + 3.28ASW - 0.6NAV - 1.79$$

Additionally, we experimented with a polynomial regression (of degree 2 and 3). The selected features were squared before fitting a linear model and the Ridge regression was then applied to select better features.

6.3 Building a Quadratic Model with Regularization

An alternative way to build a readability formula is making multiplication of features and hence producing more complex quadratic model. For instance, given a list of 3 features: ASL, ASW and NAV, one could generate the following list of 6 feature products: ASL ASL, ASW ASL, ASW ASW, ASL NAV, ASW NAV and NAV NAV. These new features are used to fit a linear regression model, making it possible to explore combinations of existing features as terms in the readability formula. Note, that initial three features are added to the final set of features. Thus, the resulting formula will have 10 parameters overall (9 for features and 1 for the intercept). In the experiment with a quadratic model, the initial set of features is limited to the following: ASW, ASL, UNAV and NAV. After generation of feature products, the set of features contains 15 features (1 for the intercept, 4 initial features, and 10 features generated as pair products).

We also tried three different regularization techniques, but Lasso and Elastic-Net performed outrageously bad. In contrast, Ridge regression performed better than the existing linear models. In fact, during validation the absolute error exceeded 1.0 only three times. The formula for the quadratic model (Q) is the following (the intercept was fitted to zero):

Q = – 0.124ASL + 0.018 ASW – 0.007 UNAV + 0.007 NAV –0.003 ASL^2 +
+ 0.184 ASL ASW + 0.097 ASL UNAV – 0.158 ASL NAV + 0.09 ASW^2 +

+ 0.091 ASW UNAV + 0.023 ASW NAV – 0.157 UNAV^2 –
– 0.079 UNAV NAV + 0.058 NAV^2 .

In the rest of this section we provide evaluation of the derived formulas and compare them with the existing formulas for readability of Russian texts.

6.4 Evaluation of Models Performance

Given the small size of the corpus, to evaluate formulas we use Leave-One-Out Cross-validation (LOOCV). In this setting test set contains a single text and the training set contains all remaining documents. Thus, in corpus of 14 documents it is possible to generate 14 splits. For each such split we build a model, evaluate MSE and then calculate the average. The result of LOOCV is provided in Table 5.

Table 5. Performance of models measured with LOOCV.

	Linear							Quadratic
	M_0	M_1	M_2	M_3	M_4	F_4	F_5	Q
LOOCV MSE	0.76	0.74	0.68	0.67	1.13	**0.62**	0.83	0.54
MSE on training set	0.42	0.25	0.34	0.28	0.58	0.24	0.24	0.18
LOOCV MAE	0.73	0.76	**0.72**	0.68	0.85	**0.72**	0.84	0.68
MAE on training set	0.55	0.44	0.49	0.45	0.62	0.44	0.44	0.37

6.5 Comparison to Existing Readability Formulas

The Flesh Reading Ease formula was adopted for the Russian language only in the late 1970-s: first by M.S. Matskovskiy in 1976. Later, I.V. Oborneva has proposed a readability formula for Russian. In 1976, M.S. Matskovskiy computed the first readability formula for the Russian language:

$$Z_1 = 0.62ASL + 0.123X_3 + 0.051,$$

where Z_1 is text readability (or difficulty), ASL is the average sentence length (in words); X_3 is the percentage of words of more than 3 syllable in the text. Another formula which became quite popular in Russian readability studies is the one developed by I.V. Oborneva (2005):

$$Z_2 = 0.5ASL + 8.4ASW - 15.59,$$

To compute the coefficients in the formula, the researcher compared:

– the average length of syllables in English and Russian words in 100 parallel English-Russian literary texts and;
– the percentage of multi-syllable words in dictionaries for Russian and English.

I.V. Oborneva concluded that an average English word is formed of 2.97 syllables, while an average Russian word consists of 3.29 syllables. We evaluate the formulas Z_1, Z_2 on the corpus compiled for the study and compare the results of readability prediction for each text separately. Results of the comparison are provided in Figs. 1 and 2.

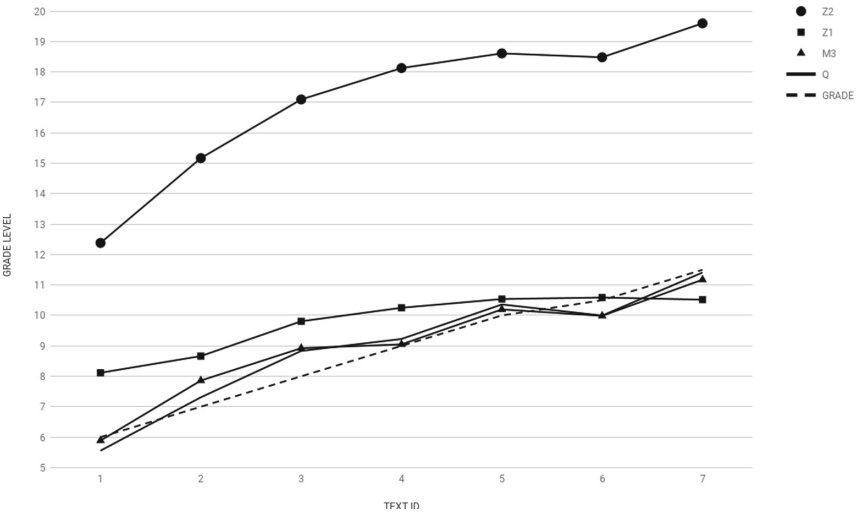

Fig. 1. Comparison of the formulas Z_1, Z_2, M_3 and Q on the "BOG" subcorpus of the Russian academic corpus. Dashed line represents the ground truth.

7 Analysis of Results

By now there have been formed two approaches to automatic assessment of text complexity. The classical approach, which we pursue in this paper, implies selecting a limited number of most relevant parameters for estimating text complexity and developing a text complexity formula based on the linear regression method. Another approach presupposing selecting the largest possible number of parameters - 100 or more - and applying a classifier such as Random Forest. The second approach is applied, in particular, in works of Reynolds [18], Laposhina [19] and Sadov [20]. The drawback of this approach is lack of transparency for the end-user. As for the first approach it proved to be useful for testing and applying TTR metrics measured on the Corpus of Russian academic texts. For the first time in Russian readability studies we applied a two-step method including assessment of correlation of coefficients, using Ridge regression and other methods of selecting the most informative parameters and finally we applied modified TTR parameters. As a result, in the new linear formula designed to measure Russian texts readability we use three parameters: ASW, ASL, UNAV ((number of unique Adjectives + number of unique Nouns)/(number of unique Verbs)).

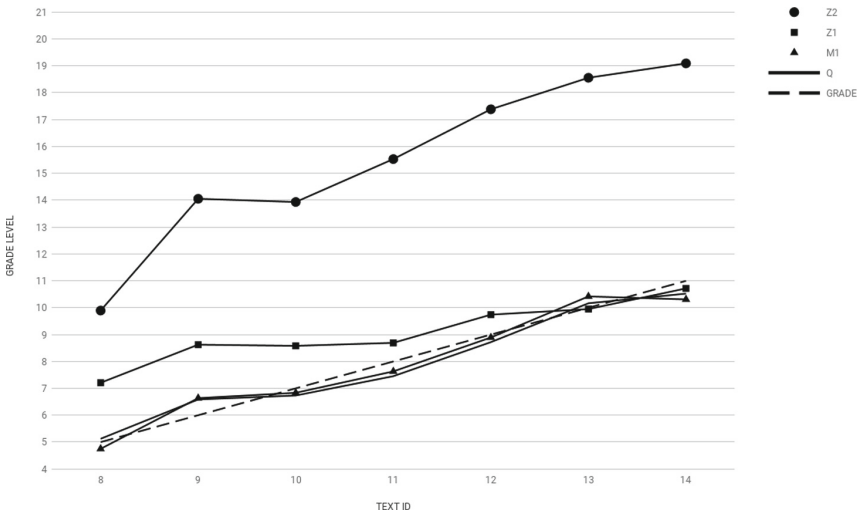

Fig. 2. Comparison of the formulas Z_1, Z_2, M_3 and Q on the "NIK" subcorpus of the Russian academic corpus. Dashed line represents the ground truth.

The curves in Figs. 1 and 2 show that the new formula provides a much higher accuracy measuring text complexity than the formula of I.V. Oborneva and a better accuracy than the formula of M.S. Matskovskiy. Another formula offered by the authors is also innovative in terms of its being not linear but quadratic. It also provides a higher accuracy in comparison with linear ones. However, it is not intuitively perceivable, and the achieved improvement in accuracy is not comparatively much higher. Taking into account that in this and previous works of the authors a rather large number of basic parameters (at the sentence level) of a text were explored and obtained improvements in accuracy is relatively not high, we can make a preliminary conclusion that as a result of the studies conducted we managed to develop models close to optimal.

8 Conclusion

The article offers new formulas to measure the level of complexity of Russian texts. This study is carried out on a text corpus of secondary and high school textbooks in Social Studies that we compiled earlier. We show that the previously proposed formulas do not correctly determine complexity level of academic texts in Russian. Solving the problem we studied and applied a number of parameters which were never used in Russian text complexity assessment, though successfully applied for assessing English and other languages text complexity. We offer original metrics in two innovative readability formulas, i.e. a quadratic (introduced for the first time in Russian readability studies) and linear. The accuracy of both exceeds the accuracy of all previously computed readability formulas for Russian texts. The latter does not imply the research conducted is to be viewed

as final, allowing no further studies or disputes on the matter. As the number as well as sets of linguistic metrics are almost infinite, some other combinations of text metrics may provide better results. The predominant number of studies on Russian text complexity so far have been performed on morphological, lexical and syntactic levels: neither paragraph nor text level features have ever become text complexity metrics. That is where we see perspectives of Russian readability studies.

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 (Sect. 3 in the paper) was created without supporting by the Russian Science Foundation.

References

1. Solnyshkina, M.I., Harkova, E.V., Kiselnikov, A.S.: Comparative Coh-Metrix analysis of reading comprehension texts: Unified (Russian) State Exam in English vs Cambridge First Certificate In English. English Lang. Teach. **7**(12), 65–76 (2014)
2. Flesch, R.: A new readability yardstick. J. Appl. Psychol. **32**, 221–233 (1968)
3. McLaughlin, G.: SMOG grading: a new readability formula. J. Reading **12**(8), 639–646 (1969)
4. Nevdakh, M.M.: Research of information characteristics of educational text using methods of multidimensional statistical analysis. Appl. Inform. **4**(16), 117–130 (2008)
5. Lorge, I.: Predicting readability. Teacher's Coll. Rec. **45**, 404–419 (1944)
6. Flesch, R.: Estimating the comprehension difficulty of magazine articles. J. Gen. Psychol. **28**, 63–80 (1943)
7. Matskovskii, M.S.: Problems of Readability of Printed Material. Semantic Perception of a Speech Message in Mass Communication, pp. 126–142. Nauka, Moscow (1976)
8. Oboroneva, I.V.: The automated estimation of complexity of educational texts on statistical parameters. Diss. Ped. n. M., 2006. 165 p
9. Falkenjack, J., Jonsson, A.: Classifying easy-to-read texts without parsing. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) (2014)
10. Falkenjack, J., Heimann, M., Jönsson, A.: Features indicating readability in Swedish text. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 27–40 (2013)
11. Piotrovsky, R.G. and others: Mathematical linguisticstick. Textbook. manual for ped. in-tov. M.: Higher School, 383 p. (1977)
12. Hultman, T.G., Westman, M.: Gymnasistsvenska. Liber, Lund (1977)
13. Cvrček, V., Chlumská, L.: Simplification in Translated Czech: A New Approach to Type-Token Ratio-Russian Linguistics, pp. 309–325. Springer, Dordrecht (2015). https://doi.org/10.1007/s11185-015-9151-8
14. Romanishin, G.V.: The study of the lexical wealth of scientific texts in New information technologies in automated systems: materials of the nineteenth scientific and practical seminar. M.: IPM them. M.V. Keldysh. - 352 p. (2016)

15. Karmanova, D.: Crisis of Russian higher education: towards the issue of aspectization labyrinth. J. Soc. Hum. Res. **1**, 78–84 (2012)
16. Stepanov, V.I., Stepanova, O.T.: The crisis of education in Russia: the ways and causes of the exit. In: Non-State-Walled Education in Russia, Novosibirsk (1996)
17. Ivanov, V.V., Solnyshkina, M.I., Solovyev, V.D.: Efficiency of text readability features in Russian academic texts. Comput. Linguist. Intellect. Technol. **17**, 277–287 (2018)
18. Reynolds, R.: Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 289–300 (2016)
19. Laposhina, A.N.: Analysis of relevant characteristics for automatic assessment of complexity of Russian texts used in courses for Russian as a foreign language [Electronic resource]: URL: http://www.dialog-21.ru/media/3993/laposhina.pdf. Accessed 10 July 2018
20. Sadov, M.A.: Development of an approach for measuring Russian text readability. Master course thesis. NRU HSE. 2018
21. Crossley, S., Allen, D., McNamara, D.: Text readability and intuitive simplification: a comparison of readability formulas. Read. Foreign Lang. **23**(1), 84–101 (2011)
22. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge (2007)