Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

AUTOMATED TEXT READABILITY ASSESSMENT FOR RUSSIAN SECOND LANGUAGE LEARNERS[1]

Laposhina A. N. (antonina.laposhina@gmail.com),
Veselovskaya T. S. (tatianus2006@yahoo.com),
Lebedeva M. U. (m.u.lebedeva@gmail.com),
Kupreshchenko O. F. (ofkupr@gmail.com)

Pushkin State Russian Language Institute (Moscow, Russia)

This paper presents an outline of the readability assessment system construction for the purposes of the Russian language learning. The system is designed to help educators easily obtain the information about the difficulty level of reading materials. The estimation task is posed here as a regression problem on data set of 600 texts and a range of lexico-semantic and morphological features. The scale choice and annotated text collection issues are also discussed. Finally, we present the results of the experiment with learners of Russian as a foreign language to evaluate the quality of a predictive model.

Keywords: readability, text complexity, reading difficulty, graded readers

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СЛОЖНОСТИ РУССКОГО ТЕКСТА КАК ИНОСТРАННОГО

Лапошина А. Н. (antonina.laposhina@gmail.com),
Веселовская Т. С. (tatianus2006@yahoo.com),
Лебедева М. Ю. (m.u.lebedeva@gmail.com),
Купрещенко О. Ф. (ofkupr@gmail.com)

Государственный институт русского языка
им. А. С. Пушкина (Москва, Россия)

[1] This research has been supported by the RFBR grant No.17-29-09156.

1. Introduction and related works

Today's information and text-rich world opens great opportunities for personalized learning, but at the same time, it sets the task of estimation and selection the suitable information. Suitable is understood as relevant to the educational purposes on the one hand and interesting and meaningful for this particular student on the other.

As R. Reynolds notes, tools for automatic identification of complexity of a given text would help to avoid one of the most time-consuming steps of text selection, allowing teachers to focus on pedagogical aspects of the process. Furthermore, these tools would also make it possible for learners to find appropriate texts by themselves [Reynolds, 2016].

In general, automated text difficulty assessment is the task of labeling a text with a certain difficulty level, such as grades, the age of the student, CEFR[2] levels, or some other abstract scale. The need of estimating texts by difficulty is not new: it starts from the beginning of the 20th century in a context of school education with quite simple formulas based on words and sentences length.

Nowadays both methods and possible application areas of such systems have widely expanded. Originated in the field of school education, researches on estimation of text complexity and search of appropriate ways of its simplification can play a significant role in specific applications where the accessibility of information is extremely important: for instance, readability assessing of government documentation for the general public[3], applications helping readers with dyslexia [Rello et al., 2012] or with intellectual disabilities [Feng et al., 2009], other groups of poor readers. Finally, the issue of finding educational texts with appropriate difficulty level for the second language learners is our particular interest. In modern NLP researches readability assessment posed as a data-driven machine learning task, is using a variety of text features from habitual word length to complex syntactic [Schwarm and Ostendorf, 2005] and discourse features [Pitler and Nenkova, 2008], features from statistical language models [Collins-Thompson and Callan, 2004], etc.

The task of text complexity estimation for the second language learners has some peculiarities. Thus, Heiman indicates the greater role of grammatical features in the second language readability research compared to native language one [Heilman et al., 2008]. The differences in the vocabulary level are also worth noticing. In our previous research [Laposhina, 2017] we have found out that the lexical group of features demonstrates one of the best correlation scores with text complexity in Russian. Perhaps, this is due to the difference in vocabulary acquisition of native and foreign languages. Walker et al. notes the disparity of reading in the native and the second or foreign language: when we first learnt to read in our first language, we already knew at least 5,000 words orally [Cunningham 2005], whereas we are usually plunged into reading a second language at an early stage, when we know very little of the language. L2 readers are constantly confronted with vocabulary they do not

[2] Common European Framework of Reference for Languages.

[3] https://plainlanguage.gov

know [Walker, 2013]. Moreover, the differences in readability assessing for a second language include sufficiently clear and rigorous scale levels, knowledge and skill requirements for each level, word lists, and vocabulary.

There are a few readability researches for the Russian as a Foreign Language. R. Reynolds builds a six-level Random Forest classifier with a range of lexical, morphological, syntactic, and discourse features and obtains F-score of 0.671. Better results were shown in binary classification task with two adjacent reading levels (e.g. A1-A2): F-score here is about 0.8–0.9. The author also provides information about feature`s information gain. [Karpov et al. 2014] use Classification Tree, SVM, and Logistic Regression models for binary classification of 4 CEFR levels (A1-C2, A2-C2, and B1-C2). The design of the given classification task seems not to fit the author's objective 'to retrieve appropriate material for their (students) language level' [Karpov et al., 2014], as the classification of adjacent reading levels is absent. A predictive model was trained on the base of 219 texts and 25 features including sentence and word length, the percentage of words from vocabulary lists and the number of several POS. The most predictive one were word lists. The authors also examine the sentence-level readability classification on 'B1 level and lower' and 'higher than B1' using transformed Dale-Chall model. [Sharoff et al. 2008] use Principal Component Analysis (PCA) in the aim to find the range of features that make a text difficult to read across a variety of languages without requiring complex resources, such as parsers. In order to realize that, they use word and sentence length, Flesch Readability Formula, average number of some specific word forms and coverage by frequency lists. The two main components from PCA can be interpreted as grammatical and lexical dimensions of difficulty. Authors also present the results of the experiment on using this system in actual language teaching.

2. Readability Assessment

As noticed by [Kevyn Collins-Thompson 2014], a machine-learning approach to readability prediction consists of three basic steps:

- First, a gold-standard training corpus of individual texts is constructed.
- Second, a set of features is defined that are to be computed from a text.
- Third, a machine learning model learns how to predict the gold standard label for a text from the text's extracted feature values.

Our work has been done in the established tradition. In section 1, the scale choice and training data set construction is discussed. Section 2 is devoted to feature extraction and selection; section 3 represents machine-learning algorithms training; and finally section 4 presents an evaluation experiment with a real educational life.

2.1. Scale choice and corpus constructing

For text complexity research a scale selection is being required: this will determine the way of corpus annotation and the type of machine learning task. Discussing traditional readability formulas, the text is considered to be suitable based on the

reader's age or grade, but this differentiation does not reflect information about real reader's competence. This situation is clearly illustrated by the authors of the project on the personalization of the readability metrics Lexile[4]: in their video presentation they show a family who has come to a store to buy kid's sneakers; searching for a suitable pair, parents do not use child's individual shoe size, but focus on his age. The authors of this project offer an abstract numerical index that consists of text metrics and the vocabulary of a particular student as a scale.

An abstract scale is also widely used among readability studies: from 0 to 100 [Orphee De Clercq, 2017], 1 to 5 [Pitler and Nenkova, 2008], binary—easy/difficult or suitable/not suitable for this level, triple—simple/average/difficult [Selegey et al., 2015]. Quite easy and effective way to get annotated training data may be using parallel collections of texts: e.g. Simplified VS Normal Wikipedia, [Sharoff et al, 2008], Children VS Adult version of Encyclopedia Britannica [Schwarm and Ostendorf, 2005]. Regarding the multi-level scale, it can be graded reader collections such as Weekly Reader, an educational newspaper with texts targeted at different grade levels [Weekly Reader, 2004].

As for a second language readability studies, the most common decision here is to use a standard grading scale for foreign language proficiency which is already developed for assessment and certification of foreign students ([Reynolds, 2016]; [Karpov, 2014]; [Schwarm and Ostendorf, 2005]). For European languages, this is the CEFR[5] level system that is measured with a six-level scale from A1 (Beginner) up to C2 (Proficiency). This system of levels has several advantages:

1. The availability of the specific regulatory documents that clarify the requirements for knowledge of vocabulary, grammar and syntax for each level.
2. Independence from such subjective categories as grade / age / number of years of study. These levels have a specific amount of language material that a person who claims to have a certificate of appropriate level should know.
3. The textbooks contain information on what levels they are intended.
4. There is a correlation with the real-life situations (for example, it is necessary to have a certain level to enter Russian universities, get a job in Russia, get Russian citizenship, get permission to teach Russian, etc.). Thus, the level of text complexity becomes a less abstract category.

level	A1	A2	B1	B2	C1	C2	C2+
number of text	108	120	106	97	39	75	48

Table 1. *Corpus content distribution*

There are 6 basic CEFR levels, translated into numerical form (A1 = 0, A2 = 1 etc) at the core of our scale. So the complexity of the text is presented as an increasing value, that reflects the concept of process of language acquisition more naturally,

[4] https://lexile.com

[5] https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

than 6 closed classes. Our corpus contains about 600 texts from the CIE resource[6] and several textbooks. Authors of these books provided the information about the target level. The content distribution is shown in **Table 1**.

For the C2 (the level of an educated native speaker) texts from news portals and articles from popular magazines on various subjects were used. However, it is obvious that the reading difficulty of texts marked as native speaker level can also differ greatly. Therefore, we added to our scale the C2+ level, which will include texts supposedly perceived as difficult by Russian native speakers: texts of laws, articles from the popular science magazine N+1[7], noted by the editors as complicated (they defined complexity as an amount of the scientific background in this field which is needed to understand the article).

We have faced several issues while collecting corpus: a) information about level could be absent in the textbook; b) this information may not contain a clear indication of the CEFR levels (B1-B2, «advanced», «for the second semester»); c) the difficulty level reflects the author's subjective evaluation. Therefore, in the future we are planning to perform expert or crowdsourcing annotation of our text collection to fix these limitations and to get more objective information about complexity of given texts using average score from several annotators.

2.2. Feature extraction and selection

We select the features to extract taking into account the following principles: 1) the features should reflect the information provided in the regulatory documents. 2) following [Sharoff et al. 2008], we believe that the features should be quite simple and reproducible if we are talking about the real usage of this system in language learning.

First, we extract some basic text metrics such as average and median word length, sentence length, average number of syllables per word, percentage of 'long' words (more than 4 syllables), average number of punctuation marks per sentence. This group of features is easy to get but it is still capable to show high correlation with a difficulty level in an obvious way: the longer the text and the words in it, the more likely it is difficult to read.

[François and Miltsakaki 2012] in their readability study have found that the best prediction performance was obtained using both classic (readability formulas) and non-classic features. Considering this, we have applied as a feature 5 commonly used readability formulas, which are using following parametres:

1. Flesch–Kincaid: (words / sentences) + (syllables / sentences);
2. Coleman Liau index: (characters / words) + (sentences / words);
3. Automated Readability Index: (characters / words) + (words / sentences);
4. Dale-Chall formula: ('difficult' words that are out of Dale's 3000 simple words list / all words) + (words / sentences);
5. Simple Measure of Gobbledygook: (words more than 4 syllables / sentences).

[6] http://texts.cie.ru

[7] https://nplus1.ru

More information about readability formulas adaptation for Russian is available in [Begtin, 2015].

Following previous researches (e.g. [Pitler and Nenkova, 2008]; [Zeng et al., 2008]; [Laposhina, 2017]), we paid attention to the group of lexical features: there are subsets of features based on coverage by vocabulary lists for each level ("lexical minimums"), frequency lists by Lyashevskaya and Sharoff[8] and Brown [N. Brown, 1996], and number of words from some specific word lists: abstract words, emotional words, verbs of motion, modal constructions, Dale's list of 3000 "simple words"[9], lists of 1000 and 2000 basic words from the Basic English Project[10]. As for the last ones, we realize the roughness of the English word lists` translation, but even approximate information on their correlation to the text complexity in Russian can motivate our further study in this field.

The next feature subset provides data about grammatical information. The percentage of POS or grammatical forms is counted here for a sentence and for a whole text, e.g. 'percent of nouns in a sentence', 'percent of nominative case in a text'.

To estimate the impact of these features in Russian second language readability assessment, the Pearson and Spearman correlation coefficients and p-value were calculated[11]. Top-30 features contains all groups of features but in different proportions.

Feature	Pearson coefficient	p-value	Spearman coefficient	p-value
A2 word list coverage of a text	-0.85	1.3e-171	-0.87	5.6-e186
Formula SMOG	0.75	2.6e-110	0.74	6e-108
Mean sentence length	0.72	3.6e-100	0.71	1.1e-96
10000 frequency word list coverage of a text	-0.69	2.2e-86	0.70	1.3e-90
Dale 3000 word list coverage of a text	-0.68	1.6e-84	0.70	1.3e-92
Abstract words list coverage of a text	0.58	3.9e-57	0.60	2.3e-63
Percentage of neuter words per text	0.55	1.3e-49	0.60	5e-68
Median number of punctuation per sentence	0.55	1.4e-49	0.55	7.3-50
Percentage of words in genitive case per text	0.50	2.8e-39	0.60	6.1e-60

Table 2. *Examples of correlation coefficient for different groups of features*

The highest correlation was shown by the lexical minimums coverage—this fact not only confirms the connection between the lexical minimums and text difficulty, but also characterizes the corpus content, which consists mostly of the textbooks, designed, in turn, according to the lexical minimums,—this has led to a vicious circle. All five readability formulas, sentence and word length information have also shown top results. The presence of features from specific word lists as Dale Word List and

[8] http://dict.ruslang.ru/freq.php

[9] https://en.wikipedia.org/wiki/Dale–Chall_readability_formula

[10] http://ogden.basic-english.org/

[11] https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.pearsonr.html

Basic English translated versions encourage us to continue research in this area and to develop similar lists for Russian.

The top morphological features are presented by the percentage of neuter nouns, words in nominative and genitive cases, and participles. Most of the grammatical features have positive correlation (e.g. the high proportion participles can indicate passive forms and specific Russian participle constructions which cause difficulties in understanding among foreigners; a number of neuter nouns may be connected with a number of special terms and abstract concepts). In contrast, there is a negative correlation between the percentage of words in nominative case and the difficulty level, as the less often the nominative case occurs in the sentence, the larger a proportion of the oblique cases is in it, which is also difficult. Mean sentence length, number of prepositions and conjunctions may indicate a syntactical aspect of difficulty.

A number of linear correlations between these text features was detected, e.g. connection between different readability formulas or lexical minimums, frequency lists. We will keep it in mind while model fitting.

2.3. Regression model

The aim of this part of work was to predict the correct assessment of a text from the continuous-valued scale from 0(A1) to 6(C2+). In order to do this, we have experimented with two linear regression algorithms: ordinary least squares Linear Regression and Ridge Regression (linear least squares with l2 regularization) by scikit-learn[12]. The mathematical objective of this techniques is to minimize mean squared error.

	Linear Regression	Ridge regression
all 149 features		
explained variance	0.73	0.83
mean squared error	0.67	0.49
44 best correlation features		
explained variance	0.82	0.84
mean squared error	0.49	0.46

Table 3. *Model Evaluation*

The models were built with:

a) all 149 features

b) 44 features with correlation by Pearson > 0.3.

To evaluate the results we have used a standard metrics as explained variance score and mean squared error. The best result was achieved by Ridge regression based on 44 best correlation features. We assumed that Ridge Regression better results may be explained by its resistance to multicollinearity of features. Twenty-fold cross-validation test showed accuracy 0.82 (±0.05) for Ridge Regression and 0.80 (±0.07) for Linear Regression.

[12] http://scikit-learn.org

To visualize the output of an algorithm a confusion matrix was constructed. Rows represent here an actual level by the corpus data while columns represent predicted levels.

Standart \ Prediction	A1	A2	B1	B2	C1	C2	C2+
A1	21	3	0	0	0	0	0
A2	9	15	4	1	0	0	0
B1	1	7	12	0	0	0	0
B2	0	5	9	17	5	0	0
C1	0	0	0	3	4	0	0
C2	0	0	0	7	14	3	0
C2+	0	0	0	0	2	5	3

Table 4. *Confusion matrix*

Table 4 shows, that mispredictions more than 1 level are only 10% of a test set, that is quite encouraging. It's also interesting to note 'the direction' of errors: algorithm more often underestimates the difficulty (47 VS 12), especially at high levels. One of the reasons of such phenomenon may be connected with the peculiarity of the corpus content: texts in B2 and C1 textbooks are aimed at confident users of Russian and provide information on complex grammatical constructions and various functional styles of the Russian language, so they can be more difficult, than the usual news articles that we collected for the C2 level.

Text	*Predicted level*
Tale story 'Masha and bears'	2.2 (B1)
Article from travel blog (1 000 words)	2.9 (B1)
A. Chekhov. Basic Education (a novel)	3.1 (B2)
A. Pushkin. The Captain's Daughter (3 000 words)	4.2 (C1)
The contract for renting an apartment	4.5 (C1)
V. Nabokov. Lolita (3 000 words)	5.9 (C2)

Table 5. *Example of predictions*

The examples of system working with authentic Russian texts are shown below. These results correspond both to the intuition of the expert teachers and to the requirements of the state standards for Russian learners, where reading authentic texts with minimal adaptation are appropriate for readers beginning with B1 level and above. More details on the evaluation experiment will be described in the next section.

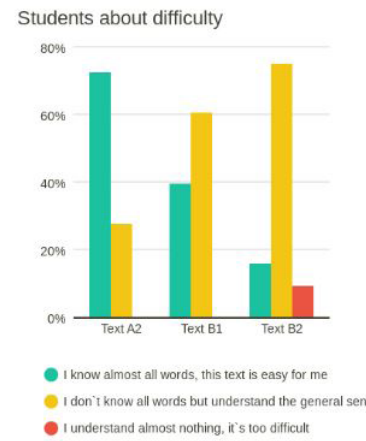
2.4. Evaluation

To test the accuracy of our approach to automatic text complexity measurement and to estimate its applicability in real educational life, we have proceeded an experiment with 78 international students at B1 level of Russian language proficiency.

It took place at the Pushkin State Russian Language Institute in February 2018. Three authentic texts on the similar topic with minimal adaptation were prepared; our system evaluated them as A2, B1 and B2 respectively. The students were asked to read each text without dictionary, to mark unknown vocabulary, to do post-reading quiz and to note how difficult to understand these texts were.

	Text A2	Text B1	Text B2
Level assessement by the algorithm	1.63	2.69	3.02
Level assessement by teachers	1.17	1.7	2.35
Median reading time (minutes)	4	5	6
Number of words out of B1 word list	11 (6%)	21 (10%)	31 (15%)
Mean number of words marked by students as unknown	0.9	3.11	5.9
Students anwered correct at least 2 of 4 post text quiz	96%	88%	76%
Students answered correct all 4 post text quiz	42%	20%	15%

Table 6. *Survey results*



The core insights from this study are shown below. The scale of text difficulty is readily seen here: the more difficult by our algorithm text is, the more words and syntactic constructions are marked by students as unknown and the less percentage of correct quiz answers are given. During personal interview students also easily ordered given texts by difficulty level highlighting that text 3 is the most difficult. Besides, they avoided to pick the option 'I understand almost nothing, this text was too difficult' in the questionnaire: this can be caused both by psychological factors, when intermediate-level students are not comfortable to admit such an overgeneralized option and by weakness of our program due to it's tendency to overestimate the real level of the text difficulty. We will take it into account while our further research.

3. Conclusion and further work

In this article we presented a supervised approach for text complexity assessment for Russian as a Second Language using linear regression. The best result was performed by Ridge Regression algorithm, trained on the 44 best correlation features set. As our further work we can point out such directions as:

1. Corpus expansion, adding the segment with authentic texts, mainly annotated by different experts;
2. Searching for new lexico-semantic features (polysemantic words, idioms and collocations, archaisms and historicisms, conversational vocabulary, genre-specific words seem particularly promising).

References

1. *Begtin, I. V.* (2014), What is "Clear Russian" in terms of technology. Let's take a look at the metrics for the readability of texts: the blog of the company "Information Culture" [Chto takoe "Ponjatnyj russkij jazyk" s tochki zrenija tehnologij. Zagljanem v metriki udobochitaemosti tekstov: blog kompanii "Informacionnaja kultura"], available at: http://habrahabr.ru/company/infoculture/blog/238875/
2. *Brown, N.* (1996), Russian Learners' Dictionary: 10,000 Russian Words in Frequency Order, Routledge, 1996.
3. *Collins-Thompson, K.* (2014), Computational assessment of text readability: a survey of current and future research. In: François, Thomas and Delphine Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics 165:2, pp. 97–135.
4. *Collins-Thompson, K., & Callan, J.* (2004), A language modeling approach to predicting reading difficulty. In Proceedings of HLT-NAACL 2004, pp. 193–200.
5. *Feng, L., Elhadad, N., & Huenerfauth, M.* (2009), Cognitively motivated features for readability assessment. In The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), pp. 229–237.
6. *François, T., Miltsakaki, E.* (2012). Do NLP and machine learning improve traditional readability formulas? Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Association for Computational Linguistics, 2012, 49–57.
7. *Heilman, M. J., Collins, K., Callan, J., & Thompson, M. E.* (2007), Combining lexical and grammatical features to improve readability measures for first and second language texts. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, Rochester, New York, USA, pp. 460–467.
8. *Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M.* (2007), Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Proceedings of HLT-NAACL'07, pp. 460–467.
9. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.
10. *Laposhina, A.,* (2017), Relevant features selection for the automatic text complexity measurement for Russian as a foreign language. [Analiz relevantnyh priznakov dlya avtomaticheskogo opredeleniya slozhnosti russkogo teksta kak inostrannogo], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017), Issue 17, p.1–7.
11. *Pitler, E. & Nenkova, A.* (2008), Revisiting readability: a unified framework for predicting text quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–195.
12. *Rello, L., Saggion, H., Baeza-Yates, R., Graells, E.* (2012), Graphical schemes may improve readability but not understandability for people with dyslexia. Proceedings of NAACL-HLT 2012.

13. *Reynolds R.* (2016), 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.
14. *Schwarm, S. E. & Ostendorf, M.* (2005), Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 523–530.
15. *Sharoff S., Kurella S., Hartley A.* (2008), Seeking needles in the web's haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference, (TaLC-8), Lisbon, Portugal.
16. *Walker A., White G.* (2013), Technology Enhanced Language Learning: connecting theory and practice, Oxford University Press.
17. *William H. DuBay* (2006), The Classic Readability Studies. Impact Information, Costa Mesa, California.