Investigating the Robustness of Reading Difficulty Models for Russian Educational Texts

Ulyana Isaeva\(^1\) and Alexey Sorokin\(^{1,2(\text{Cor})}\)

\(^1\) Moscow State University, Moscow, Russia
ya.uvi23@yandex.ru, alexey.sorokin@list.ru

\(^2\) Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Abstract. Recent papers on Russian readability suggest several formulas aimed at evaluating text reading difficulty for learners of different ages. However, little is known about individual formulas for school subjects and their performance compared to that of existing universal readability formulas. Our goal is to study the impact of the subject both in terms of model quality and on the importance of individual features. We trained 4 linear regression models: an individual formula for each of 3 school subjects (Biology, Literature, and Social Studies) and a universal formula for all the 3 subjects. The dataset was created of schoolbook texts, randomly sampled into pseudo-texts of size 500 sentences. It was split into train and test sets in the ratio of 75 to 25. As for the features, previous papers on Russian readability do not provide proper feature selection. So we suggested a set of 32 features that are possibly relevant to text difficulty in Russian. For every model, features were selected from this set based on their importance. The results obtained show that all the one-subject formulas outperform the universal model and previously developed readability formulas. Experiments with other sample sizes (200 and 900 sentences per sample) prove these results. This is because feature importances vary significantly among the subjects. Suggested readability models might be beneficial for school education for evaluating text relevance for learners and adjusting those texts to target difficulty levels.

Keywords: Readability assessment · Text reading difficulty · Russian · Educational texts · Robustness · Text simplicity

1 Introduction

Text complexity [2], reading difficulty [1], readability [2] and comprehensibility [15] are text characteristics which are often undistinguished in the literature related to text complexity. They all refer to measuring to what extent the structure of a text affects the effort that the reader needs to understand the text. This characteristic is one of the crucial criteria for moderating course materials of educational programs since the content of educational material (e.g. textbooks)
influences the results of education. The concept is studied at the intersection of psychology and linguistics [11].

For more than 40 years there have been held numerous studies aiming at defining a proper set of parameters which could become a basis for a universal automated tool for measuring the readability of texts in Russian. In recent years a step forward was made by a series of surveys introduced by V. D. Soloviev, M. I. Solnyshkina and their coauthors [4,12,13]. These papers share the authors’ experience of using a wide variety of features for evaluating the complexity of Russian school textbooks on Social Studies and provide several Russian readability linear formulas which showed good quality in application to textbooks and outperformed the older formulas.

In this work, we provide a pilot study of the extensibility of the above-mentioned models (trained on texts on Social Studies) to textbooks on other school subjects: Biology and Literature. We suggested some new features which may be useful for evaluating the readability of textbooks. After feature selection, we trained individual linear models for each of the three subjects. Another model was trained on a mixed dataset of texts on all the 3 subjects. These models outperformed the existing readability models for Russian. We provide an analysis of the selected features, which demonstrated that adding new morphological and syntactic features results in significant improvement in model performance. Also, we show that there is little coincidence in feature sets for different subjects, which supports the assumption that readability models trained on texts on one subject are not always extensible to other subjects.

2 Related Work

The history of readability formulas begins in the XX century. One of the first formulas was introduced in 1949 by Flesch [3]. It is based on two text parameters: ‘average sentence length’ and ‘average syllables per word’. This formula was adjusted to use for many other languages.

First Russian readability formulas developed by Matskovsky and Mikk [8,9] aimed at evaluating a learner’s ability to understand texts. At that time a narrow set of features was used: it consisted mostly of features related to sentence and word length, the familiarity of the words to a reader, and their abstractness [11]. One of the most important steps in the direction of modern view at measuring readability was made in the late 1970s with the development of the theory of information by Claude Shannon. A big shift was made towards using not only the above-mentioned quantitative features but also a wide range of qualitative features such as parts of speech and syntax phrases [11].

With the development of computational tools began the modern period of study of text readability. The one who contributed greatly to the development of modern Russian readability formulas was I. Oborneva, who provided an adaptation of the Flesch Reading Ease formula to the Russian language [10]. To do this she carried out a contrastive study of the average length of English and Russian words using ‘Dictionary of the Russian Language’ edited by S. Ozhegov and
‘English-Russian Dictionary’ edited by V. Muller. I. Oborneva analyzed a hundred of fictional English texts with Russian translations and found the average word lengths to be 3.29 and 2.97 syllables for Russian and English respectively. The formula is as follows:

\[
FR_E = 206.836 - (1.52 \times ASL) - (65.14 \times ASW)
\]  

In 2018 Oborneva’s formula was developed further by V. Solovyev and his colleagues. In the paper [13] the formula was applied to the Russian Readability Corpus constructed of school textbooks on Social Studies and its coefficients were adjusted. The authors suggested an extended number of features and computed the values of correlation coefficients between each feature and the target variable. The results obtained showed that the most correlated features, i.e. most effective for evaluating text readability of textbooks on Social Studies are ‘frequency of content words’ and ‘number of adjectives per sentence’. Considering this, the authors suggested a new formula. It represents a modification of the Flesch-Kincaid-Grade (FKG) [6] formula with a new feature – ‘number of adjectives per sentence’. The modified formula allows to predict the grade of the text fragment with the RMSE of 0.51\(^1\) on their dataset:

\[
FKG(ASL, ASW, ADJ) = 1.59 + (0.23 \times ASL) - (1.48 \times ASW) + (3.97 \times ADJ)
\]

Another paper by the same group of authors [4] covers an investigation of the impact of 24 lexical, syntactic, and frequency features on readability assessment. This work provides a correlation analysis of features that demonstrated that features traditionally used for evaluating readability (‘average sentence length’ and ‘average words per sentence’) have the highest degree of correlation with the target variable (grade level). Also, among the most correlated parameters, there are such features as ‘average number of coordinating chains’ and ‘average number of participial constructions’. This proves that information about syntactic structure can be made use of when assessing text readability. Using the ridge regression approach, the authors selected a subset of 16 features that have the maximum influence on the target variable. However, they did not evaluate the performance of model on the selected set of features.

Despite this, most questions concerning the complexity of Russian texts remain unsolved. One of these questions is the robustness of developed formulas across domains and genres. We address this problem in the current study.

\[^1\] We replicate the calculation of RMSE for [12] model and our dataset and obtain much larger error. Since the exact test set and model predictions are not available, we have no explanation of such difference.
3 Methodology

3.1 Data Collection and Sampling

For our task, we collected a corpus of Russian schoolbook texts. Some of them were obtained by OCR processing and some were taken from the database provided by the authors of [4,12,13] at https://kpfu.ru/slozhnost-tekstov-304364.html. The dataset size and structure are provided in Table 1. Since the size of our corpus was not large enough to make statistically significant conclusions, we had to utilize a text sampling technique. That is, we evaluated our model, not on contiguous sentences or paragraphs, but random samples from the text. Note that this approach was also used in the previous studies on the KPFU dataset.

Table 1. Dataset size (in sentences) and structure.

| Grade/Subject | Biology | Literature | Social studies | Total |
|---------------|---------|------------|----------------|-------|
| 5             | 4836    | –          | 2326           | 7162  |
| 6             | 4995    | 10525      | 2949           | 18469 |
| 7             | 6798    | 5003       | 3942           | 15743 |
| 8             | 3424    | 15995      | 7122           | 26541 |
| 9             | 6650    | 21075      | 6465           | 34190 |
| 10            | 14834   | 8010       | 20822          | 43666 |
| 11            | –       | 6780       | 22218          | 28998 |
| Total         | 41537   | 67388      | 68544          | 174769|

The most controversial issue in text sampling was the sample size. In [12] sample size 500 was suggested as one which allows obtaining readability assessment for a sample close to that of the text from which the sample was taken. This value was derived empirically, so we carried out our own investigation of the impact of sample size on readability prediction accuracy. We took 500 as the basic value for the sample size and also performed several experiments with sample sizes 200 and 900 to ensure robustness.

3.2 Evaluation of Recent Readability Formulas for Russian

Our initial assumption is that existing readability formulas trained on texts on a particular school subject may appear to underperform when applied to texts on other subjects. So, we applied the formula (2) to texts on Biology, Literature, Social Studies, and to the entire dataset (texts on 3 subjects). The results are provided in Table 2. It is quite expectable that the RMSE values vary across the subjects, also, it is important to notice that the model intended for texts on Social Studies shows better performance on Literature texts.

Considering this, further in this study we will treat differently the subsets (texts on different subjects) of our dataset and train individual independent
Table 2. Model performance across different subjects (500 sentences per sample).

| Subject         | Biology | Literature | Social studies | Total |
|-----------------|---------|------------|----------------|-------|
| RMSE            | 1.36    | 1.03       | 1.15           | 1.22  |
| $R^2$           | 0.43    | 0.52       | 0.46           | 0.49  |

models for them. Also, we will build a general model for the entire dataset and compare its performance to that of the individual models to find the answer to our research question about readability models extensibility (universality).

3.3 Introducing a Feature Set

The number of features used in readability models and observed in the theoretical papers differs widely, starting from the first Flesch-Kincaid Grade with 2 features to large studies like [2], where the total number of features viewed in the paper is 87. Generally, the more features are suggested as possibly relevant to the target variable, the better performance may be achieved by the final model based on selected features provided the features are relevant to the problem under consideration.

Thus, we developed a set of 32 features which might be useful for assessing the readability of schoolbook texts. To the best of our knowledge, a wider set of features for Russian readability models had never been observed, though a comparable number of the features (26) appeared in [4]. The features are divided into 3 groups based on their nature: the first group consists of features traditionally used for text readability estimation, the second is for features based on morphological features of the words, and the third one contains averaged data about syntactic characteristics of the sentences. All morphosyntactic information was obtained using the UDPipe [14] tagger and parser.

- Group 1. Traditional features:
  1. Average number of words per sentence.
  2. Average number of characters per word.
  3. Average number of syllables per word.
  4. Average frequency of words in the text.
  5. Average Juilland coefficient [5].
- Group 2. Features based on POS-tags: computed as a proportion of words with the feature among all words in a sentence:
  6. Finite verbs.
  7. Non-finite verbs.
  8. Verbs in active voice.
  9. Verbs in passive voice.
  10. Active participles.
11. Passive participles.
12. Adverbs.
13. Adverbial participles.
14. Nouns.
15. Full adjectives.
16. Short adjectives.
17. Personal pronouns.
18. Other pronouns.
19. Prepositions.
20. Conjunctions.
21. Negative particles.
22. Particles.
23. Content words.

– Group 3. Syntactic features: computed for every sentence and averaged for a text:

24. Number of clauses.
25. Syntax tree depth.
26. Length of adverbial participle phrases.
27. Length of participle phrases.
28. Number of embedded NPs with genitive case.
29. Ratio of subordinate clauses.
30. Number of relative clauses.
31. Ratio of infinite clauses.
32. Maximum distance between a node and its descendant.

3.4 Feature Selection

Feature selection is one of the most important steps in model building, since non-informative or duplicate predictors may adversely affect the model quality. Moreover, the more features there are in the model, the more complex it becomes, which requires more computing resources [7].

As indicated above, the suggested set of 32 features is a hypothetical one and is subject to a thorough feature selection procedure. This section will cover the feature selection methods which we applied to our features. These are univariate feature selection, recursive feature elimination, selection by model coefficients, and decision trees. Univariate feature selection gave the best results, so we will provide the results only for this method here.

Since all the feature values are numerical and the target variable is numerical too, it is convenient to view the task of readability assessment as a regression task. We do not aim at obtaining the predicted readability level as an integer, because float numbers are not less illustrative in this case. This assumption is supported by the fact that first readability formulas were built as linear models and so were many modern ones [4,10,12].
The selected subsets of features formed a basis for 4 linear regression models: one for every subject and one universal model trained on the texts on all the 3 subjects. We used the Ordinary Least Squares Regression algorithm implementation provided at the ScikitLearn library (LinearRegression). To evaluate the models, we computed 2 parameters: root mean squared error (RMSE) and determination coefficient ($R^2$). RMSE was chosen as the basis for model comparison.

**Univariate Feature Selection.** The univariate feature selection method used in this work aims at evaluating the predictors from the initial feature set based on some criterion and selecting a subset of features that will form the basis of the model. We chose 2 selection criteria: mutual information and F-score function between a predictor and the target variable since they are both commonly used for linear regression tasks. Figure 1 presents a dependency between the number of features selected by the method and the performance of a model based on this number of features. These results were obtained by training a model for every number of features and validating using 10-fold cross-validation.

![Fig. 1. Model RMSE dependency on the number of selected features. Solid lines represent results obtained with mutual information criteria and dashed line represent that of F-score function. The values of RMSE are negative to visualize the effect as ‘the higher – the better’.](image-url)
It can be observed that for every model there is a too small and a too large number of features with which the model’s quality decreases. Approximately from 1 to 7 features the error decreases significantly with each added predictor and reaches a plateau (with a slight decrease) until it grows again when the model starts to overfit due to an excess of features.

Since we seek to find a model with the best possible quality and at the same time face a restriction on the number of features, for every model we chose a point at the start of the plateau with the least RMSE. These points are marked with circles in Fig. 1. There was the only exception for the Biology model because after the plateau there is a notable decrease in error. That is, the best point is 26 features, being much larger than 7 features (which is the first peak point marked with a cross on Fig. 1). We set this model to be based on 7 features for the sake of interpretability.

Summarizing, the following features were kept in the final model:

- **Biology (7 features)**: average number of words (1), average number of syllables per word (3), verbs in active voice (8), personal pronouns (17), syntax tree depth (25), number of embedded noun phrases (28), maximum linear distance between parent and child (32).
- **Literature (10 features)**: average number of words (1), average number of characters per word (2), finite verbs (6), verbs in active voice (8), nouns (14), full adjectives (15), personal pronouns (17), content words (23), syntax tree depth (25), number of embedded noun phrases (28).
- **Social studies (14 features)**: average number of words (1), average number of characters per word (2), average number of syllables per word (3), average word frequency (4), non-finite verbs (7), verbs in active voice (8), passive participles (11), adverbs (12), personal pronouns (17), other pronouns (18), negative particles (21), particles (22), syntax tree depth (25), number of embedded noun phrases (28).
- **General (14 features)**: average number of words (1), average number of characters per word (2), average number of syllables per word (3), average word frequency (4), average Juilland coefficient (5), finite verbs (6), verbs in active voice (8), adverbs (12), nouns (14), personal pronouns (17), content words (23), syntax tree depth (25), number of embedded noun phrases (28), ratio of infinite clauses (31).

## 4 Results

### 4.1 Evaluation of the Models

To evaluate the models’ quality, we created 10 random splits of our dataset into train and test sets in proportion 75/25. The estimates presented in Table 3 are averaged over these 10 evaluations. The sample size is set to default 500 sentences. For comparison, we also evaluate the model 2 on our data.\(^2\)

\(^2\) We do not evaluate the adjusted FRE-formula (1), [10] as it predicts the abstract readability score, not the school grade.
Table 3. Model performance across different subjects (500 sentences per sample) compared to formula 2.

| Subject               | Biology | Literature | Social studies | All subjects |
|-----------------------|---------|------------|----------------|--------------|
| Opt. number of features | 7       | 10         | 14             | 14           |
| RMSE (our model)      | 0.52    | 0.45       | 0.49           | 0.77         |
| RMSE (2)              | 1.36    | 1.03       | 1.15           | 1.22         |

As it was stated above, we validated these results on different sample sizes. The values of RMSE depending on sample size are presented in Table 4.

Table 4. Model performance across different subjects evaluated with different sample sizes (RMSE values).

| Subject             | Biology | Literature | Social studies | All subjects |
|---------------------|---------|------------|----------------|--------------|
| Sample size = 200   | 0.48    | 0.54       | 0.39           | 0.59         |
| Sample size = 500   | 0.54    | 0.44       | 0.43           | 0.63         |
| Sample size = 900   | 0.35    | 0.35       | 0.33           | 0.54         |

It can be observed that within all the experiments the universal models tend to be inferior compared to the single-subject models. However, introducing separate models for different subjects or domains may be inefficient under some circumstances, so it is necessary to define some criteria to assess a model independently from other models and decide on whether a model is acceptable. We set this RMSE threshold to be 0.5 because in terms of defining whether a text is appropriate for students of a particular grade such an error seems to be acceptable. Of course, this assumption needs a more proper investigation in future works.

4.2 Analysis of Selected Features

To analyze which features were selected we plotted a bar chart that presents the coefficients from the 4 linear models according to the selected features. An important thing to notice here is that the so-called traditional features (1–3) turned out to be among the most ‘powerful’ in terms of readability prediction.

Four features (#3, #8, #17, and #25) were selected for all the four models. So we decided to measure the contribution of the other selected features to each of the models over these 4 features. The results are present in Table 5.
Table 5. Model performance across different subjects evaluated with different sample sizes (RMSE values).

| Subject                | Biology | Literature | Social studies | All subjects |
|------------------------|---------|------------|----------------|--------------|
| Number of features     | 4       | 4          | 4              | 4            |
| RMSE                   | 0.65    | 0.59       | 0.61           | 1.13         |
| Number of features     | 7       | 10         | 14             | 14           |
| RMSE                   | 0.52    | 0.45       | 0.49           | 0.77         |

Notice that for all 3 individual subjects the model on 4 features achieves performance comparable to the one of the optimal model. However, for the entire dataset the relative gap between these two classes of models occurs to be more significant. It implies that feature weights vary considerably between the universal models and the subject-specific models.

5 Discussion

The formulas suggested in this paper differ significantly from existing readability formulas, first of all, in terms of the number of predictors. A modern formula applied to texts on different subjects showed noticeable variance in the values of RMSE which is chosen as the main parameter to evaluate and compare readability models. The new models developed separately for each of the 3 subjects (Biology, Literature, and Social Studies) outperformed the previous formulas.

This was explained during the analysis of the features selected by the univariate feature selection method. The observation showed that there is not much coincidence in feature sets for different subjects and thus a significant decrease in model performance is expected when the model trained on one subject is applied to texts on other subjects. Though there is a common “core” of 4 features that are kept for all the domains, even their weights vary significantly across subjects.

Also, an increase in models’ quality is explained by newly added features which were chosen according to the specific characteristics of texts on each of the subjects. Although traditional features such as ‘average sentence length’ and ‘average syllables per word’ turned out to be the most informative for assessing text readability, we managed to obtain a significant improvement in models’ performance using features referring to linguistic characteristics of a text: morphology and syntax. Further improvement may be achieved by taking into account lexical characteristics but we leave it for future work. We also expect that lexical characteristics tend to be more domain-dependent than more abstract features (Fig. 2).
Fig. 2. Linear models’ coefficients for the selected features. The features which have 0 were not selected to any of the four models.
6 Conclusion

The results obtained in this pilot study represent the next step in the development of readability models for Russian educational texts. To the best of our knowledge, there have not been any investigations of the universality of readability models for educational texts. The presented findings encourage future studies on readability formulas for different subjects or groups of subjects.

The main result of the paper is that the most relevant features vary from subject to subject, nevertheless ‘traditional features’ (such as average sentence length or word length in syllables) are present in all feature sets.

Acknowledgements. We thank V. Solovyev, M. Solnyshkina, V. Ivanov et al. for publishing their database of schoolbook texts that we used in our study. This contributed greatly to the findings obtained in our work. We are also very grateful to anonymous AIST reviewers whose thorough comments helped to improve the paper.

References

1. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 193–200 (2004)

2. Dell’Orletta, F., Wieling, M., Venturi, G., Cimino, A., Montemagni, S.: Assessing the readability of sentences: which corpora and features? In: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 163–173. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/W14-1820. http://aclweb.org/anthology/W14-1820

3. Flesch, R., Gould, A.J.: The Art of Readable Writing, vol. 8. Harper, New York (1949)

4. Ivanov, V., Solnyshkina, M., Solovyev, V.: Efficiency of text readability features in Russian academic texts. In: Komp’juternaja Lingvistika I Intellektual’nye Tehnologii, pp. 267–283 (2018)

5. Juillard, A., Chang-Rodríguez, E.: Frequency dictionary of Spanish words. Technical report (1964)

6. Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch (1975)

7. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6849-3

8. Matskovsky, M.: Problems of readability of printed material. In: Semantic Perception of Speech Messages in Conditions of Mass Communication, Nauka, pp. 126–142 (1976)

9. Mikk, Y.: On factors of comprehensibility of educational texts. Ph.D. thesis, Tartu University (1970)

10. Oborneva, I.: Automatic assessment of the complexity of educational texts on the basis of statistical parameters. Ph.D. thesis, Moscow State Pedagogical University (2006)
11. Solnyshkina, M.I., Kiselnikov, A.S.: Text complexity: study phases in Russian linguistics. Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya 6(38), 86–99 (2015). https://doi.org/10.17223/19986645/38/7

12. Solovyev, V., Ivanov, V., Solnyshkina, M.: Assessment of reading difficulty levels in Russian academic texts: approaches and metrics. J. Intell. Fuzzy Syst. 34(5), 3049–3058 (2018)

13. Solovyev, V., Solnyshkina, M., Ivanov, V., Batyrshin, I.: Prediction of reading difficulty in Russian academic texts. J. Intell. Fuzzy Syst. 36(5), 4553–4563 (2019)

14. Straka, M., Straková, J.: Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 88–99 (2017)

15. Zakaluk, B., Samuels, S.: Issues related to text comprehensibility: the future of readability. Revue québécoise de linguistique 25(1), 41–59 (1996)