



<https://doi.org/10.22363/2687-0088-30132>

Research article


Text complexity and linguistic features: Their correlation in English and Russian

Dmitry A. MOROZOV¹  , Anna V. GLAZKOVA² 
and Boris L. IOMDIN³ 

¹*Novosibirsk State University, Novosibirsk, Russia*

²*University of Tyumen, Tyumen, Russia*

³*Vinogradov Russian Language Institute, RAS, Moscow, Russia*

 morozowdm@gmail.com

Abstract

Text complexity assessment is a challenging task requiring various linguistic aspects to be taken into consideration. The complexity level of the text should correspond to the reader's competence. A too complicated text could be incomprehensible, whereas a too simple one could be boring. For many years, simple features were used to assess readability, e.g. average length of words and sentences or vocabulary variety. Thanks to the development of natural language processing methods, the set of text parameters used for evaluating readability has expanded significantly. In recent years, many articles have been published the authors of which investigated the contribution of various lexical, morphological, and syntactic features to the readability level. Nevertheless, as the methods and corpora are quite diverse, it may be hard to draw general conclusions as to the effectiveness of linguistic information for evaluating text complexity due to the diversity of methods and corpora. Moreover, a cross-lingual impact of different features on various datasets has not been investigated. The purpose of this study is to conduct a large-scale comparison of features of different nature. We experimentally assessed seven commonly used feature types (readability, traditional features, morphological features, punctuation, syntax frequency, and topic modeling) on six corpora for text complexity assessment in English and Russian employing four common machine learning models: logistic regression, random forest, convolutional neural network and feedforward neural network. One of the corpora, the corpus of fiction literature read by Russian school students, was constructed for the experiment using a large-scale survey to ensure the objectivity of the labeling. We showed which feature types can significantly improve the performance and analyzed their impact according to the dataset characteristics, language, and data source.

Keywords: *text complexity, machine learning, neural network, corpus linguistics*



For citation:

Morozov, Dmitry A., Anna V. Glazkova & Boris L. Iomdin. 2022. Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics* 26 (2). 426–448. <https://doi.org/10.22363/2687-0088-30132>

Научная статья


Сложность текста и лингвистические признаки: как они соотносятся в русском и английском языках

Д.А. МОРОЗОВ¹  , А.В. ГЛАЗКОВА² , Б.Л. ИОМДИН³ 

¹Новосибирский государственный университет, Новосибирск, Россия

²Тюменский государственный университет, Тюмень, Россия

³Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

 morozowdm@gmail.com

Аннотация

Автоматическая оценка читабельности текста — актуальная и непростая задача, которая требует учёта разнообразных лингвистических факторов. Сложность текста должна соответствовать уровню читателя: слишком сложный текст останется непонятым, слишком простой будет скучным. Исторически для оценки читабельности использовались простые характеристики: средняя длина слов и предложений, разнообразие лексики. Благодаря развитию методов обработки естественного языка набор используемых для оценки параметров текста существенно расширился. За последние годы было опубликовано множество работ, в которых исследовался вклад в сложность текста различных лексических, морфологических, синтаксических признаков. Тем не менее, поскольку использованные методы и корпуса довольно разнообразны, затруднительно делать общие выводы об эффективности различных лингвистических характеристик текста. Более того, не было проведено сравнение влияния признаков для различных языков. Целью настоящего исследования является проведение масштабного сравнения признаков различного характера. Мы экспериментально сравнили семь часто используемых типов признаков (индексы читабельности, традиционные, морфологические, синтаксические, пунктуационные, частотные признаки и тематическое моделирование) на материале трёх русскоязычных и трёх англоязычных корпусов, с использованием четырех распространённых алгоритмов машинного обучения: логистической регрессии, случайного леса, свёрточной нейронной сети и нейронной сети с прямой связью. Один из корпусов — корпус художественной литературы, читаемой российскими школьниками, — был создан для этого эксперимента с помощью масштабного опроса для обеспечения объективности разметки. Мы показали, какие типы признаков могут значительно повысить качество прогнозирования, и проанализировали их влияние в зависимости от характеристик корпуса, его языка и источника текстов.

Ключевые слова: сложность текста, машинное обучение, нейронные сети, корпусная лингвистика

Для цитирования:

Morozov D.A., Glazkova A.V., Iomdin B.L. Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics*. 2022. Vol. 26. № 2. P. 426–448. <https://doi.org/10.22363/2687-0088-30132>

1. Introduction

Text complexity is crucial for the comprehension process. Texts that are too difficult may be hard to understand. In contrast, unnecessarily simple texts may conflict with the reader's level of communicative skills. Hence, text complexity assessment is an essential task that represents a major challenge for developing natural language processing (NLP) tools. Text complexity can be expressed in different ways, ranging from quantitative characteristics to semantic complexity represented by text topics. Numerous studies have been published on evaluating various features for text complexity assessment. The reported results were obtained from text corpora of widely differing sizes and domains. Moreover, the authors used different machine learning (ML) models and text representation techniques (Feng et al. 2010, Ivanov et al. 2018, Cantos & Almela 2019, Isaeva & Sorokin 2020, Deutsch et al. 2020, Glazkova et al. 2021, Martinc et al. 2021). This makes it complicated to achieve an objective evaluation of the impact of different types of features.

The goal of this paper is to perform an extensive evaluation of seven feature types for text complexity assessment that were frequently used in research on the subject. The results allow us to make the text complexity analysis process more defined and transparent. These findings have a broad spectrum of potential applications in education and recommendation systems. We used the following feature types: readability, traditional features, morphological features, punctuation, syntax, frequency, and topic modeling. The features were evaluated on three Russian and three English text complexity corpora and four ML models in order to answer the following research questions (RQ).

- **RQ1:** How do different types of features affect the performance of baselines?
- **RQ2:** Is the impact of these feature types similar in English and Russian?
- **RQ3:** Do feature-enriched models outperform fine-tuned state-of-the-art language models?

This paper is organized as follows. Section 2 contains a brief review of related works. In Section 3, we introduce datasets and models utilized and provide some short background on the feature types we use. Section 4 presents and discusses the results. Section 5 concludes the paper.

2. Related Work

2.1. Readability: methods and approaches

The earliest approaches to automatic readability assessment, developed in the second half of the 20th century, were intuitive, severely limited by the small number

of existing natural language processing tools and the lack of computing power. Most of these readability indices represented linear combinations of simple features, such as average word or sentence length, the proportion of words in the text with a large number of syllables, and the proportion of words included in special lists of “simple” and “complex” words. A more detailed overview of these algorithms is given, for example, by Cantos and Almela (2019).

These algorithms have become widespread in practical tasks, despite their simplicity and seeming naivety. They are still in use in some spheres, including government requirements for insurance, for example, in some US states such as Connecticut (Chapter 699a. Readable language in insurance policies). At the same time, it is quite clear that such simple mechanisms cannot give a reliable result, especially in relation to fiction and poetry.

2.2. New possibilities: more features, more sophisticated models

The rapid development of NLP tools, including neural networks, has made it possible to significantly expand the set of features and to improve the quality of the text complexity assessment. Since the algorithm that solves this problem can be widely used in application-oriented studies, many authors have analyzed the impact of features of different nature (e.g. Feng et al. 2010, Ivanov et al. 2018, Cantos & Almela 2019, Isaeva & Sorokin 2020, Deutsch et al. 2020, Glazkova et al. 2021).

The most intuitive way to noticeably improve the quality of the prediction, which does not require serious modifications, is to use a combination of classical algorithms. Cantos and Almela (2019) analyzed this approach on a corpus containing excerpts from English-as-a-Foreign-Language textbooks. The presented classifier is based on features from Flesch–Kincaid readability test (Kincaid et al. 1975), Coleman–Liau index (Coleman & Liau 1975), Automated readability (ARI) index (Senter & Smith 1967), SMOG grade (McLaughlin 1969) and some other. The precision of the constructed algorithm significantly outperforms separate approaches.

At the same time, significant gains can be achieved using more abstract and complex characteristics. Feng et al. (2010) analyzed the impact of various categories of features on the complexity of the text, such as the number and density of named entities, semantic chains, referential relations, language modeling, syntactic dependencies, and morphology. Ivanov et al. (2018) considered 24 various features. such as average sentence and word lengths, word frequencies, morphological, and syntactic features on the Russian corpus.

The robustness of different features across various corpora with texts of different languages, styles, and genres is also a challenging question. This issue was partly solved by Isaeva and Sorokin (2020), who studied three groups of features, namely, average lengths plus frequencies, morphological, and syntactic ones. The experiments on three corpora of educational texts showed that there is a core of features that are crucial for all texts: the average number of syllables per word, the proportion of verbs in active voice among all words, the proportion of personal

pronouns among all words, and the average syntax tree depth. Deutsch et al. (2020) considered a few combinations of deep learning models with linguistically motivated features in order to determine how much such a combination will improve the quality of predictions.

As in many other areas of natural language processing, state-of-the-art results can be achieved by fine-tuning Bidirectional Encoder Representations from Transformers (BERT) presented by Devlin et al. (2019) and similar models. Martinc et al. (2021) studied unsupervised and supervised approaches, comparing BERT, Hierarchical attention networks, and Bidirectional Long short-term memory networks. The experiments were conducted on a few English and Slovenian corpora. The results suggested that BERT can be used as a high-level baseline for our research.

2.3. Russian datasets

The study of readability for the Russian language is developing more slowly than, for example, for English. This is largely due to the lack of text corpora with the labeled age of readers or readability level. The creation of a corpus of texts readable for students requires a large number of preliminary surveys of respondents of school age. This can be avoided by using the texts of school textbooks. For example, Ivanov et al. (2018) presented Russian Readability Corpus based on the texts of school textbooks on Social Studies for grades 5–11, later expanded by Isaeva and Sorokin (2020). An alternative way would be to use lists of Recommended Reading as done by Iomdin and Morozov (2021). Presumably, many school students read such texts. However, by their nature, such corpora are far from ideal: they contain texts that, in the opinion of adult compilers, should be read at the appropriate age, but they in no way guarantee the fact of reading, much less understanding of these texts. Such lists often include various historical texts that are incomprehensible at the level of an average student without additional explanations. In the above-mentioned list of recommended literature, there are such complex texts from the point of view of the reader as “The Lay of Igor's Campaign” and “The Tale of Bygone Years”¹. Thus, despite the high labor intensity, a large-scale survey of schoolchildren seems to be the optimal way to select texts for the corpus, which makes it possible to achieve the best representativeness.

3. Datasets, features and models

3.1. Books Read by Students corpus

In real practice, text complexity is often identified with the age of the reader, usually a student. Thus, readability indices often predict a grade of school or college in which the text would be understandable (e.g., Kincaid et al. 1975). This makes it

¹ It is interesting that due to the inability to take into account the obsolescence of the vocabulary, traditional indices assign these texts a very low level of complexity (Iomdin & Morozov 2021).

possible to exclude from consideration adult readers, whose reading experience can be incredibly variable and, as a result, can be estimated only with a very large margin of error. In our study, we decided to use this approximation and assess the age of a reader instead of the abstract complexity level.

To create a corpus with a labeled reading age, two stages of experiments have been conducted. At the first stage, respondents of preschool and school-age (or their parents, acting on their behalf) were asked to name a few of the books they had most recently read, together with their authors. The survey involved 1176 respondents under the age of 16, of which more than 1000 were of school age. In order to complete the list of presumably popular books, a similar additional survey was conducted at the linguistic session at the Sirius Educational Center (Russia), the target audience of the survey was mainly students of grades 9–11.

At the second stage, another group of respondents was asked to select the books they read from the list of those most frequently mentioned during the first stage, 93 books in total. The books were distributed over several subsurveys in such a way that each respondent was asked about no more than two books from the series. The experiment involved 1120 respondents, each of whom answered questions about 15 or 16 books.

Due to the insufficient number of respondents of primary school age and in order to exclude the peculiarities of the individual experience of readers, we decided to combine the ages of students into 5 categories: 1–2, 3–4, 5–7, 8–9, and 10–11 grades. The proportion of participants who read the text was calculated for each book and each age category. After that, the youngest category containing at least 25% of respondents who had read the book was assigned a book label. For example, “In Search of the Castaways” (“Les Enfants du capitaine Grant”) was marked as read by 0% of respondents from the category 1–2, 13% of respondents from the category 3–4, and 70% of respondents in the category 5–7, thus, the category 5–7 was assigned.

As a result, 75 books were included into the *Books Read by Students* corpus (**BR**). A complete list is given in Appendix A. 18 texts from the second stage of the experiment did not receive any age label and were not included in the corpus. In such a situation, the level of complexity cannot be established even with a sharp increase in the proportion of readers when moving between age categories. A brief overview of the collected corpus is presented in Table 1.

Table 1. Brief statistics of the Books Read by Students corpus

Category	All	1–2	3–4	5–7	8–9	10–11
Total texts	75	31	14	8	14	8
Total words	4077579	888984	881578	693448	1060079	553490
Total unique lemmas	50930	24513	24616	24185	30670	24645
Avg sentence length	9.41	8.37	9.58	11.19	9.55	10.44
Avg word length	4.94	4.88	5.02	5.06	4.91	4.83
Avg unique lemmas per text	6864.05	3431.6	6033.6	7375.6	6757.0	7375.6

3.2. Alternative datasets

As mentioned above, for the Russian language, there are few corpora with complexity labels. We decided to compare BR with two of them: *Fiction Previews (Fic)* presented by Glazkova et al. (2021) and *Recommended Literature (RL)*, which we constructed in the previous study from the list of recommended literature for schoolchildren created by the Russian Ministry of Education (Iomdin & Morozov 2021). All collected texts were divided into fragments of 70 sentences each. This allowed us to considerably increase the size of datasets without significant loss of labeling quality (Isaeva & Sorokin 2020).

For English, there are a few corpora with complexity labels; we used three of them. *Common Core State Standards (CC)*² is a corpus designed to represent text complexity levels for each grade in the USA. *OneStopEnglish (OSE)* corpus was specially created for training readability models (Vajjala & Lucic 2018). It consists of 189 text samples, each in three complexity versions. *CommonLit (CL)* corpus was presented at a Kaggle competition³. The main difference of this corpus from the rest ones is continuous labeling set instead of classes. An overview of the datasets is shown in Table 2.

Table 2. Some statistics of the datasets.

BR — Books Read by Students, RL — Recommended Literature, Fic — Fiction Previews, CC — Common Core State Standards, CL — CommonLit, OSE — OneStopEnglish

Characteristics	BR	RL	Fic	CC	CL	OSE
Total texts	5795	9230	58184	219	2834	567
Total categories	5	3	2	6	1	3
Total words	2897003	4888290	26252666	84014	491944	381137
Total unique words	55577	103875	304731	10007	24449	13611
Avg words/text	984.75	1053.28	918.64	450.46	199.65	757.82
Avg words/sentence	13.92	14.95	13.12	22.24	24.94	22.04
Avg sentences/text	70	70	70	23.26	9.46	34.98

3.3. Linguistic Features

According to the related works, we identified seven types of features, which can be used to assess the text complexity: 1) readability indices, e.g., the Flesch–Kincaid readability test and the SMOG grade; 2) traditional features, e.g., the average word length and type/token ratio; 3) morphological feature, e.g., the proportion of nouns and verbs; 4) punctuation, e.g., the number of semicolons; 5) syntactic features, e.g., the average syntactic tree depth and number of clauses; 6) frequencies, e.g., the percentage of tokens included in the list of the most frequent words; and 7) topic modeling. In total, we collected 128 features for English and 126 for Russian of types 1–6. Additionally, we evaluated 100 topics using Latent Dirichlet allocation (Blei et al. 2003). To the best of our knowledge, such a wide

² <http://www.corestandards.org/>

³ <https://www.kaggle.com/c/commonlitreadabilityprize>

range of features was considered for the first time in relation to Russian text complexity models. A full list of features can be found in Appendix B.

For evaluation we used the following libraries: readability (Readability 0.3.1 2019), pymorphy2 (Korobov 2015), nltk (Loper & Bird 2002), gensim (Rehurek & Sojka 2010), spacy (Honnibal & Montani 2017), deeppavlov (Burtsev et al. 2018), and API of readability.io. The source code for our methods is available at GitHub (Readability 2021).

3.4. Models

We used the following machine learning algorithms as baselines:

1 Linear Support Vector Classifier (LSVC). LSVC was built with the l2 penalty and the squared hinge loss. We fitted LSVC on bag-of-words (BoW) text representations with a maximum length of 10000. Scikit-learn (Pedregosa et al. 2011) was used for implementation.

2 Random Forest (RF). We used 100 estimators and the Gini impurity to measure the quality of a split. The implementation details are the same as those for LSVC.

3 Feedforward Neural Network (FNN). The hyperparameters used are identified in Table 3. We employed the Adam optimizer (Kingma & Ba 2015). The model was implemented using Keras (Chollet et al. 2015). Each model was trained with early stopping for a maximum of 100 epochs and patience of 20. We utilized Sentence Transformers text representations obtained using the all-mpnet-base-v2 model (Reimers & Gurevych 2019) for the English corpora and the distiluse-base-multilingual-cased model (Reimers & Gurevych 2020) for the Russian ones.

4 Convolutional Neural Network (CNN). The training details are the same as for FNN. The model was implemented using FastText embeddings for English (Mikolov et al. 2018) and Russian (Kutuzov & Kuzmenko 2016).

Table 3. Hyperparameters for neural baselines

Hyperparameters	FNN	CNN
Number of convolutional layers	-	2
Number of pooling layers	-	2
Number of convolutional filters	-	256
Filter size	-	256
Number of fully connected layers	3	1
Size of fully connected layers	1024	32
Activation (hidden layers)	Tanh	relu
Activation (output layer)	softmax (classification) linear (regression)	
Dropout	0.5	

We randomly shuffled all the Russian corpora and the CL dataset and split them into train and test sets in the ratio of 3:1. The splitting was conducted in such a way that all fragments of one book belonged either to the train set or to the test

one. Due to the small number of documents in OSE and CC corpora, we used five-fold cross-validation on these datasets to obtain more reliable results. For all of the models above, we systematically evaluated each type of linguistic features applying the Min-Max technique for normalization.

To compare the scores obtained with the results of a few state-of-the-art models, we used BERT-base and RuBERT (Kuratov & Arkhipov 2019) for English and Russian corpora respectively. Each model was fine-tuned for 3 epochs using Transformers (Wolf et al. 2020) with the learning rate of $2e-5$ using the AdamW optimizer (Loshchilov & Hutter 2018). We set batch size to 4 and maximum sequence size to 512. To validate our models during the development phase, we divided labelled data using the train and validation split in the ratio 90:10.

We used mean absolute error (MAE) and weighted F1-score to compare the results. MAE is calculated as an arithmetic average of the absolute errors $e_i = y_i - x_i$, where y_i is the prediction, x_i is the true value, n is the number of values:

$$MAE = \frac{\sum_{i=1}^n e_i}{n}. \quad (1)$$

The weighted F1-score calculates the standard F1-score for each label, and finds their average, weighted by the number of true instances for each label. The formula of the standard F1-score is:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \quad (2)$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative predictions respectively.

4. Results and Discussion

We report the results in terms of the MAE (for the CL corpus) and weighted F1-score (for the other corpora) in Table 4. The gray highlight presents the values that outperform the baseline, the values that outperform the baseline by at least 1% are underlined. The best results are shown in bold. Appendix C contains the overall results expressed by several common metrics.

Based on the results, we can identify four performance categories, see Table 5, that describe the impact of various linguistic features (**RQ1**). In most cases, the considered features improved the model performance. Meanwhile, it was only morphological features that gave a positive impact in most classifiers for all corpora. Readability features exceeded the baseline on most models for most datasets except the BR corpus. Punctuation, traditional, and syntactic features showed a performance growth at least for two models on each corpus. Frequency and topic modeling features produced mixed results. On the one hand, topic modeling features improved the performance of all classifiers on two corpora. Nevertheless, the score on the OSE corpus increased for only RF classifier. This could be because the corpus contains parallel versions of the same papers. Although frequency features improved the performance in some cases, they demonstrated

higher MAE in most classifiers on the CL dataset. Probably, it reflects the fact that short texts normally lack word frequency and context information because of word sparsity (Yan et al. 2013, Xun et al. 2016).

Table 4. F1 (%) and MAE for each type of features. For F1 the best result is the highest, for MAE — the lowest. BERT refers to the BERT-base in English corpora and to RuBERT in the Russian ones

Model	BR	RL	Fic	CC	CL	OSE
BERT	45.23	62.74	80.96	42.18	0.453	70.99
LSVC	32.31	63.16	76.66	28.22	0.673	70.41
RF	30.94	48.21	78.87	30.03	0.627	68.21
FNN	34.22	63.26	66.34	33.73	0.533	54.00
CNN	39.82	58.12	80.12	33.60	0.593	70.64
+ readability						
LSVC	32.12	63.16	76.67	<u>32.43</u>	<u>0.663</u>	70.49
RF	29.19	<u>49.89</u>	78.45	27.77	<u>0.599</u>	<u>70.11</u>
FNN	<u>40.89</u>	63.62	<u>68.23</u>	<u>37.56</u>	<u>0.502</u>	<u>56.07</u>
CNN	<u>45.90</u>	<u>61.35</u>	80.52	<u>35.89</u>	0.590	68.59
+ traditional						
LSVC	<u>33.15</u>	62.67	77.14	<u>29.3</u>	<u>0.666</u>	69.89
RF	30.03	46.53	78.26	28.57	<u>0.609</u>	73.01
FNN	32.12	69.76	<u>70.51</u>	<u>34.7</u>	<u>0.482</u>	<u>58.76</u>
CNN	<u>44.32</u>	<u>65.19</u>	80.68	45.98	0.604	64.82
+ morphological						
LSVC	32.55	63.22	77.03	<u>31.99</u>	<u>0.662</u>	<u>71.75</u>
RF	30.36	46.63	76.20	29.56	<u>0.611</u>	<u>70.67</u>
FNN	<u>35.63</u>	<u>69.12</u>	<u>72.04</u>	<u>37.42</u>	<u>0.504</u>	<u>62.00</u>
CNN	<u>42.29</u>	<u>68.63</u>	80.75	<u>37.12</u>	<u>0.573</u>	69.02
+ punctuation						
LSVC	32.26	62.87	76.73	<u>30.44</u>	<u>0.664</u>	70.41
RF	30.30	47.25	78.20	28.39	0.629	<u>68.92</u>
FNN	<u>35.21</u>	<u>66.54</u>	<u>68.70</u>	32.51	<u>0.505</u>	<u>55.79</u>
CNN	<u>40.74</u>	<u>67.95</u>	80.86	<u>43.68</u>	<u>0.580</u>	64.33
+ syntactic						
LSVC	<u>32.66</u>	61.91	76.88	<u>29.27</u>	0.674	70.54
RF	28.84	46.70	77.41	33.97	<u>0.617</u>	<u>72.59</u>
FNN	32.10	<u>69.41</u>	<u>68.31</u>	36.48	<u>0.476</u>	<u>56.68</u>
CNN	<u>45.49</u>	<u>65.35</u>	<u>81.01</u>	<u>36.19</u>	0.592	58.71
+ frequency						
LSVC	32.52	63.07	76.84	<u>33.08</u>	<u>0.662</u>	<u>71.34</u>
RF	30.01	45.87	77.76	26.02	0.640	67.63
FNN	31.45	<u>67.46</u>	<u>67.58</u>	<u>35.33</u>	0.729	<u>63.01</u>
CNN	46.97	<u>65.08</u>	81.11	<u>38.65</u>	0.597	56.38
+ topic modeling						
LSVC	<u>35.36</u>	62.14	76.92	<u>29.97</u>	0.669	67.00
RF	<u>34.09</u>	<u>49.44</u>	77.65	27.15	0.623	66.10
FNN	<u>38.85</u>	62.01	<u>77.30</u>	<u>34.08</u>	<u>0.516</u>	<u>59.46</u>
CNN	<u>43.93</u>	<u>65.78</u>	80.91	<u>41.28</u>	0.588	64.95

Table 5. Features types with positive impact for N classifiers on each corpus.
1 — readability indices, 2 — traditional features, 3 — morphological features, 4 — punctuation,
5 — syntactic features, 6 — frequencies, 7 — topic modeling.

Improvement	BR	RL	Fic	CC	CL	OSE
N=4	7	-	-	5	1, 3, 7	-
N=3	3	1, 3	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 6, 7	2, 4, 5	1, 3, 5
N=2	1, 2, 4, 5, 6	2, 4, 5, 6, 7	-	4	-	2, 4, 6
N=1	-	-	-	-	6	7

Tables 6 and 7 illustrate the performance growth as a percentage averaged over all classifiers for Russian and English corpora (**RQ2**). The averaged results demonstrate that the models trained on Russian texts benefit more from topic modeling and frequency features in comparison with the models trained on English corpora. On the other hand, the results on the CC corpus indicate that this superiority is rather due to text length than language properties. Readability and punctuation features present similar results for both languages. Although morphological, traditional, and syntactic features show better performance on English texts, the results on specific corpora are strongly determined by the source of texts and the type of markup. Thus, any influence of syntactic features for the OSE corpus could not be identified during our experiments. However, there was a noticeable increase for the CC corpus containing fiction texts that characterized English as an analytic language. Overall, these results indicate that the impact of all feature types is mainly attributable to specific circumstances of a corpus. This enables one to use transfer-learning algorithms for cross-lingual analysis of text corpora having similar characteristics.

Table 6. Averaged performance growth for Russian corpora, %

Features	BR	RL	Fic	Avg Rus
Readability	7.13	2.4	0.71	3.41
Traditional	1.21	4.54	1.71	2.49
Morphological	2.3	6.04	1.62	3.32
Punctuation	0.75	4.91	0.93	2.2
Syntactic	0.58	4.26	0.63	1.83
Frequency	1.88	3.4	0.48	1.92
Topic modeling	10.87	3.04	4.07	5.99

Table 7. Averaged performance growth for English corpora, %

Features	CC	CL	OSE	Avg Eng
Readability	6.39	3.05	0.96	3.47
Traditional	9.67	3.04	-1.33	4.74
Morphological	8.3	3.19	4.51	5.33
Punctuation	7.2	2.06	-1.14	2.71
Syntactic	8.18	3.04	-1.33	3.3
Frequency	5.91	-9.54	-0.76	-1.46

The performance of the models trained on feature combinations per dataset is presented in Table 8. The results are given only for those models the performance

of which was increased by two and more types of features. We enriched the baseline models with the concatenation of all features that showed a positive impact for the relevant models and datasets. The combination of features increased the F1 of RF on the OSE corpus outperforming all the results obtained for this dataset. This result is marked with an asterisk (*). Moreover, FNN trained on feature combinations showed the best result among all the feature-enriched models on the CL corpus. Taken together, the results presented in Table 4 and Table 8 demonstrate that feature-enriched models outperformed BERT on five out of the six corpora (**RQ3**). In some cases, significant increases were obtained, including 7.02% for the RL corpus and 3.8% for the CC corpus. By contrast, the performance of feature-enriched models depends on the features used and data specifics. Simultaneously, in some cases, models trained on feature combination showed a worse result, than those trained on the one type of features.

Table 8. F1 (%) and MAE for feature combinations

Model	BR	RL	Fic	CC	CL	OSE
LSVC	34.50	-	78.09	33.12	0.633	71.44
RF	-	49.38	-	-	0.568	76.44*
FNN	40.88	62.99	78.70	39.71	0.466	74.24
CNN	43.85	65.29	81.06	43.58	0.541	-
BERT	45.23	62.74	80.96	42.18	0.453	70.99

5. Conclusion

We have presented the first comparative analysis of the impact of seven types of linguistic features on the performance of text complexity models. We provided the results of large-scale experiments on six text corpora. Each feature type was evaluated in four representative ML models. Our research demonstrated the advantage of some features over others. For example, morphological features improved the performance of our models in almost all cases. At the same time, topic modeling features did not show any positive impact on the corpus containing parallel versions of the same papers. We also identified performance categories based on the scores obtained and estimated the impact of feature combinations. According to our study, the results depend more on the specific characteristics of the dataset than on language. This provides an opportunity for exploring cross-lingual transfer learning and multilingual models for text complexity assessment. Finally, experimental results on five out of the six corpora showed that feature-enriched models can achieve significant improvements in comparison with the state-of-the-art ones. Here, future research may focus on evaluating more complex semantic and narrative features, such as plot characteristics and the features related to named entity analysis, on including BERT-based features, and on explaining text complexity in terms of each feature type.

Acknowledgements

The article was funded by RFBR, project number 19-29-14224.

REFERENCES

- Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022. <https://doi.org/10.1016/B978-0-12-411519-4.00006-9>
- Burtsev, Mikhail, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lyman, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhрева & Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*. 122–127. <https://doi.org/10.18653/v1/P18-4021>
- Cantos, Pascual & Ángela Almela. 2019. Readability indices for the assessment of textbooks: A feasibility study in the context of EFL. *Vigo International Journal of Applied Linguistics* 16. 31–52. <https://doi.org/10.35869/VIAL.V0116.92>
- Chollet, Francois. 2015. Keras. Github. <https://github.com/fchollet/keras> (accessed 31.01.2022).
- Coleman, Meri & Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60(2). 283.
- Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin* 27. 11–20, 37–54.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186. Minneapolis, Minnesota. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Deutsch, Tovly, Masoud Jasbi & Stuart Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics*. 1–17. <https://doi.org/10.18653/v1/2020.bea-1.1>
- Feng, Lijun, Martin Jansche, Matt Huenerfauth & Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*. 276–284.
- Glazkova, Anna, Yury Egorov & Maksim Glazkov. 2021. A comparative study of feature types for age-based text classification. *Analysis of Images, Social Networks and Texts*. 120–134. Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-72610-2_9
- Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Iomdin, Boris L. & Dmitry A. Morozov. 2021. Who Can Understand “Dunno”? Automatic Assessment of Text Complexity in Children’s Literature. *Russian Speech* 5. 55–68. <https://doi.org/10.31857/S013161170017239-1>
- Isaeva, Ulyana & Alexey Sorokin. 2020. Investigating the robustness of reading difficulty models for Russian educational texts. In *AIST 2020: Recent Trends in Analysis of Images, Social Networks and Texts*. 65–77. https://doi.org/10.1007/978-3-030-71214-3_6
- Ivanov, Vladimir, Marina Solnyshkina & Valery Solovyev. 2018. Efficiency of text readability features in Russian academic texts, In *Komp'yuternaya Lingvistika I Intellektual'nye Tehnologii*. 284–293.
- Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Technical Training Command Millington TN Research Branch. <https://doi.org/10.21236/ada006655>

- Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.
- Korobov, Mikhail. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*. 320–332. Springer. https://doi.org/10.1007/978-3-319-26123-2_31
- Kuratov, Yuri & Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *Komp'uternaya Lingvistika i Intellektual'nye Tehnologii*. 333–339.
- Kutuzov, Andrey & Elizaveta Kuzmenko. 2016. Web-vectors: A toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts*. 155–161. Springer. https://doi.org/10.1007/978-3-319-52920-2_15
- Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Routledge.
- Loper, Edward & Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 63–70.
- Loshchilov, Ilya & Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Lyashevskaya, Olga & Serge Sharoff. 2009. *The Frequency Dictionary of the Modern Russian Language (Based on the Materials of the Russian National Corpus)*. Moscow: Azbukovnik.
- Martinc, Matej, Senja Pollak & Marko Robnik-Sikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics* 47. 1–39. https://doi.org/10.1162/coli_a_00398
- McLaughlin, G. Harry. 1969. Smog grading – a new readability formula. *Journal of reading* 12(8). 639–646.
- Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhresch & Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12. 2825–2830.
- Rehurek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
- Reimers, Nils & Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 3982–3992. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Reimers, Nils & Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 4512–4525. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- Senter, R. J. & E. A. Smith. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*. 1–14.
- Solnyshkina, Marina, Vladimir Ivanov & Valery Solovyev. 2018. Readability formula for Russian texts: A modified version. In *Mexican International Conference on Artificial Intelligence*. 132–145. Springer. https://doi.org/10.1007/978-3-030-04497-8_11

- Templin, Mildred C. 1957. *Certain Language Skills in Children; Their Development and Interrelationships*. Minneapolis: University of Minnesota Press.
- Vajjala, Sowmya & Ivana Lucic. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 297–304. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0535>
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Xun, Guangxu, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao & Aidong Zhang. 2016. Topic discovery for short texts using word embeddings. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 1299–1304. IEEE.
- Yan, Xiaohui, Jiafeng Guo, Yanyan Lan & Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*. 1445–1456. <https://doi.org/10.1145/2488388.2488514>

Internet sources

- Chapter 699a. *Readable language in insurance policies*. URL: https://www.cga.ct.gov/current/pub/chap_699a.htm#sec_38a-29 (accessed 29.05.2022).
- Readability. 2021. URL: <https://github.com/morozowdmitry/readability> (accessed 29.05.2022).
- Readability 0.3.1. 2019. URL: <https://pypi.org/project/readability/> (accessed 29.05.2022).

Appendix A. Books Read by Students corpus

If the original text was published in a language other than English, the translated title is followed by the title in the original language.

Table 9. Books included into the Books Read by Students corpus

Category	Title (original title)	Author(s)
1-2	Winnie-the-Pooh	Alan Alexander Milne
1-2	The Wizard of the Emerald City (Волшебник Изумрудного города)	Alexander Volkov
1-2	Urfin Jus and his Wooden Soldiers (Урфин Джус и его деревянные солдаты)	Alexander Volkov
1-2	The Smart Dog Sonya (Умная собачка Соня)	Andrey Usachev
1-2	Grandma and Eight Children in the Forest (Mormor og de åtte ungene i skogen)	Anne-Catharina Vestly
1-2	Eight Children and a Truck (Åtte små, to store og en lastebil)	Anne-Catharina Vestly
1-2	Pippi Longstocking (Pippi Långstrump)	Astrid Lindgren
1-2	Emil of Lönneberga (Emil i Lönneberga)	Astrid Lindgren
1-2	The Lion, the Witch and the Wardrobe	Clive Staples Lewis
1-2	Harry Potter and the Half-Blood Prince	Joanne Rowling

Category	Title (original title)	Author(s)
1-2	Harry Potter and the Chamber of Secrets	Joanne Rowling
1-2	Harry Potter and the Philosopher's Stone	Joanne Rowling
1-2	Alice's Adventures in Wonderland	Lewis Carroll
1-2	Waffle Hearts (Vaffelhjarte)	Maria Parr
1-2	Dunno in Sun City (Незнайка в Солнечном городе)	Nikolay Nosov
1-2	The Adventures of Dunno and his Friends (Приключения Незнайки и его друзей)	Nikolay Nosov
1-2	The Little Witch (Die kleine Hexe)	Otfried Preußler
1-2	Treasure Island	Robert Louis Stevenson
1-2	The Wonderful Adventures of Nils (Nils Holgerssons underbara resa genom Sverige)	Selma Lagerlöf
1-2	Pancakes for Findus (Pannkakstårten)	Sven Nordqvist
1-2	When Findus was Little and Disappeared (När Findus var liten och försvann)	Sven Nordqvist
1-2	The Mechanical Santa (Tomtemaskinen)	Sven Nordqvist
1-2	Findus and the Fox (Rävjakten)	Sven Nordqvist
1-2	Findus Plants Meatballs (Kackel i grönsakslandet)	Sven Nordqvist
1-2	Findus Goes Fishing (Stackars Pettson)	Sven Nordqvist
1-2	Findus Goes Camping (Pettson tältar)	Sven Nordqvist
1-2	Findus at Christmas (Pettson får julbesök)	Sven Nordqvist
1-2	Findus Moves Out (Findus flyttar ut)	Sven Nordqvist
1-2	The Rooster's Minute (Tuppens minut)	Sven Nordqvist
1-2	Finn Family Moomintroll (Trollkarlens hatt)	Tove Jansson
1-2	The Adventures of Dennis (Денискины рассказы)	Viktor Dragunsky
3-4	Seven Underground Kings (Семь подземных королей)	Alexander Volkov
3-4	Ronia, the Robber's Daughter (Ronja rövardotter)	Astrid Lindgren
3-4	Robinson Crusoe	Daniel Defoe
3-4	Pollyanna	Eleanor H. Porter
3-4	Three Jolly Fellows (Naksitrallid)	Eno Raud
3-4	Harry Potter and the Deathly Hallows	Joanne Rowling
3-4	Harry Potter and the Goblet of Fire	Joanne Rowling
3-4	Harry Potter and the Prisoner of Azkaban	Joanne Rowling
3-4	The Mysterious Island	Jules Verne
3-4	One hundred years ahead (Сто лет тому вперёд)	Kir Bulychev
3-4	The Adventures of Tom Sawyer	Mark Twain
3-4	The Little Water Sprite (Der kleine Wassermann)	Otfried Preußler
3-4	The Little Ghost (Das kleine Gespenst)	Otfried Preußler
3-4	Comet in Moominland (Kometjakten)	Tove Jansson
5-6-7	The Three Musketeers (Les Trois Mousquetaires)	Alexandre Dumas
5-6-7	The Captain's Daughter (Капитанская дочка)	Alexander Pushkin
5-6-7	The Adventure of the Final Problem	Arthur Conan Doyle
5-6-7	The Hound of the Baskervilles	Arthur Conan Doyle
5-6-7	A Study in Scarlet	Arthur Conan Doyle
5-6-7	The Hobbit, or There and Back Again	John Ronald Reuel Tolkien
5-6-7	Harry Potter and the Order of the Phoenix	Joanne Rowling
5-6-7	In Search of the Castaways (Les Enfants du capitaine Grant)	Jules Verne
8-9	The Time Is Always Good (Время всегда хорошее)	Andrey Zhvaleyevsky and Evgeniya Pasternak

Category	Title (original title)	Author(s)
8-9	Monday Begins on Saturday (Понедельник начинается в субботу)	Arkady and Boris Strugatsky
8-9	The Lost World	Arthur Conan Doyle
8-9	His Last Bow	Arthur Conan Doyle
8-9	The Sign of the Four	Arthur Conan Doyle
8-9	The Adventure of the Empty House	Arthur Conan Doyle
8-9	The Adventure of the Six Napoleons	Arthur Conan Doyle
8-9	Ivanhoe	Walter Scott
8-9	The Two Captains (Два капитана)	Veniamin Kaverin
8-9	The Invisible Man	H.G. Wells
8-9	The Lord of the Rings	John Ronald Reuel Tolkien
8-9	George's Secret Key to the Universe	Lucy Hawking, Stephen Hawking, Christophe Galfard
8-9	The Master and Margarita (Мастер и Маргарита)	Mikhail Bulgakov
8-9	Dandelion Wine	Ray Bradbury
10-11	The Catcher in the Rye	J. D. Salinger
10-11	1984	George Orwell
10-11	Fathers and Sons (Отцы и дети)	Ivan Turgenev
10-11	Brave New World	Aldous Huxley
10-11	Fahrenheit 451	Ray Bradbury
10-11	Lord of the Flies	William Golding
10-11	Crime and Punishment (Преступление и наказание)	Fyodor Dostoevsky
10-11	To Kill a Mockingbird	Harper Lee

Appendix B. Evaluated Features

Readability indices

- 1 Flesch–Kincaid readability test (Kincaid et al. 1975).
- 2 Coleman–Liau index (Coleman and Liau 1975).
- 3 Automated readability (ARI) index (Senter and Smith 1967).
- 4 SMOG grade (McLaughlin 1969).
- 5 Dale-Chall index (Dale and Chall 1948).

Traditional features

- 1 Average and mean sentence length.
- 2 Average and mean word length.
- 3 Long words (>4 syllables) proportion.
- 4 Type/token ratio (TTR) (Templin 1957).
- 5 NAV: TTR for Nouns only plus TTR for Adjectives only divided by TTR for Verbs only (Solnyshkina et al. 2018).

Morphological features

- 1 Percentages of lexical categories.
- 2 Percentage of grammatical cases.
- 3 Proportion of animated nouns.
- 4 Proportion of grammatical aspects for verbs.
- 5 Proportion of grammatical tenses for verbs.
- 6 Proportion of transitive verbs.

Punctuation

- 1 Punctuation/token ratio.
- 2 Semicolons/token ratio.

Syntactic features

Three features were extracted from each of the following characteristics: average, mean, and maximum.

- 1 Syntactic tree depth.
- 2 Distance between a node and its descendant.
- 3 Number of clauses.
- 4 Number of adverbial clause modifiers.
- 5 Number of adnominal clauses.
- 6 Number of clausal complements.
- 7 Number of open clausal complements.
- 8 Number of nominal modifiers.
- 9 Length of nominal modifiers sequence.

Frequencies

For evaluating frequencies of Russian and English words we used dictionaries based on Russian National Corpus (Lyashevskaya & Sharoff 2009) and British National Corpus (Leech et al. 2001) respectively.

- 1 Average and mean frequency.
- 2 Proportion of words, which are in the list of the most 100/200/.../1000 popular words, and similar features for nouns, verbs, adverbs, and adjectives separately.

Appendix C. Overall Results

In the tables below we use the following notation: F — F1-score weighted, P — precision weighted, R — recall weighted, MAE — mean absolute error, MSE — mean squared error.

MSE measures the average of the squares of the errors:

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}, \quad (3)$$

where Y_i is the vector of true values, \hat{Y}_i is the vector of predicted values.

Russian corpora

Table 10. Results for the Books Read By Students corpus

Model	F		P		R	
BERT	45.23		54.06		41.32	
LSVC	32.31		35.74		34.28	
RF	30.94		32.73		37.18	
FNN	34.22		39.06		31.75	
CNN	39.82		57.34		33.66	
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	32.12	35.5	34.2	33.15	36.79	35.2

RF	29.19	26.87	36.04	30.03	32.49	36.34
FNN	40.89	61.23	31.83	32.12	44.37	27.08
CNN	45.9	66.18	37.8	44.32	64.88	36.27
	+ morphological			+ punctuation		
LSVC	32.55	37.52	36.5	32.26	35.79	34.35
RF	30.36	37.52	36.5	30.3	37.94	36.57
FNN	35.63	42.75	31.68	35.21	39.54	33.05
CNN	42.29	55.72	37.26	40.74	57.25	33.44
	+ syntactic			+ frequency		
LSVC	32.66	36.02	34.66	32.52	35.77	34.28
RF	28.84	31.26	34.74	30.01	32.14	35.88
FNN	32.1	40.95	28.46	31.45	37.54	28.39
CNN	45.49	67.47	36.57	46.97	69.57	38.41
	+ topic modeling			Combined		
LSVC	35.36	38.63	36.88	34.5	37.36	35.88
RF	34.09	37.74	38.18	-	-	-
FNN	38.85	45.77	35.96	40.88	55.03	35.96
CNN	43.93	62.93	36.65	43.85	62.93	37.18

Table 11. Results for the Recommended Literature corpus

Model	F			P			R		
BERT	62.74			65.71			61.86		
LSVC	63.16			63.54			64.98		
RF	48.21			61.8			59.92		
FNN	63.26			79.19			53.76		
CNN	58.12			58.23			58.99		
	F	P	R	F	P	R			
	+ readability			+ traditional					
LSVC	63.16	63.22	64.89	62.67	62.83	64.56			
RF	49.89	63.88	60.68	46.53	55.2	58.82			
FNN	63.62	81.66	52.91	69.76	93.52	56.03			
CNN	61.35	66.33	59.49	65.19	66.22	64.64			
	+ morphological			+ punctuation					
LSVC	63.22	63.11	64.98	62.87	63.07	64.73			
RF	46.63	58.54	59.07	47.25	62.9	59.58			
FNN	69.12	92.34	55.78	66.54	87.1	54.43			
CNN	68.63	72.84	66.58	67.95	71.19	66.33			
	+ syntactic			+ frequency					
LSVC	61.91	61.88	63.88	63.07	62.93	64.64			
RF	46.7	57.58	58.9	45.87	57.89	58.65			
FNN	69.41	93.01	55.78	67.46	89.35	54.6			
CNN	65.35	69.58	63.21	65.08	66.22	64.64			
	+ topic modeling			Combined					
LSVC	62.14	62.71	64.22	-	-	-			
RF	49.44	65.98	60.68	49.38	62.93	60.34			
FNN	62.01	65.98	59.66	62.99	68.99	58.99			
CNN	65.78	67.24	64.89	65.29	68.18	63.54			

Table 12. Results for the Fiction Previews corpus

Model	F		P		R	
BERT	80.96		81.83		80.82	
LSVC	76.66		77.89		76.87	
RF	78.87		79.67		78.99	
FNN	66.34		72.31		65.01	
CNN	80.12		80.87		80.04	
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	76.67	77.84	76.87	77.14	78.29	77.34
RF	78.45	78.85	78.51	78.26	78.86	78.36
FNN	68.23	72.54	67.36	70.51	70.61	70.49
CNN	80.52	81.9	80.37	80.68	81.74	80.56
	+ morphological			+ punctuation		
LSVC	77.03	78.27	77.24	76.73	77.94	76.94
RF	76.2	77.16	76.38	78.2	78.93	78.32
FNN	72.04	72.09	72.04	68.7	68.75	68.69
CNN	80.75	81.73	80.65	80.86	81.84	80.75
	+ syntactic			+ frequency		
LSVC	76.88	78.08	77.09	76.84	78	77.04
RF	77.41	78.21	77.54	77.76	78.4	77.86
FNN	68.31	68.41	68.29	67.58	67.59	67.57
CNN	81.01	81.97	80.9	81.11	82.08	81.01
	+ topic modeling			Combined		
LSVC	76.92	78.18	77.12	78.09	79.3	78.27
RF	77.65	78	77.71	-	-	-
FNN	77.3	78.28	77.17	78.7	79.06	78.66
CNN	80.91	82.07	80.78	81.06	82.17	80.93

English corpora

Table 13. Results for the Common Core State Standards corpus

Model	F		P		R	
BERT	42.18		64.57		33.77	
LSVC	28.22		30.13		30.61	
RF	30.03		30.38		34.65	
FNN	33.73		37.93		32.9	
CNN	33.6		58.04		26.92	
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	32.43	33.55	35.59	29.3	31.37	31.5
RF	27.77	26.95	31.95	28.57	28.68	32.88
FNN	37.56	42.34	36.53	34.7	38.48	34.28
CNN	35.89	56.83	29.25	45.98	78.2	36.12
	+ morphological			+ punctuation		
LSVC	31.99	35.29	33.33	30.44	32.07	33.32
RF	29.56	29.53	34.26	28.39	27.23	34.65
FNN	37.42	46.15	34.7	32.51	37.2	32

CNN	37.12	57.32	30.62	43.68	60.51	37.95
	+ syntactic			+ frequency		
LSVC	29.27	29.45	31.95	33.08	35.74	34.67
RF	33.97	34.75	38.33	26.02	23.17	31.55
FNN	36.48	41.42	35.64	35.33	40.79	34.27
CNN	36.19	62.18	28.3	38.65	54.04	32.45
	+ topic modeling			Combined		
LSVC	29.97	31.38	32.42	33.12	35.21	34.67
RF	27.15	29.19	30.15	-	-	-
FNN	34.08	38.34	32.91	39.71	47.55	37.94
CNN	41.28	65.93	33.85	43.58	44.09	39.44

Table 14. Results for the CommonLit corpus

Model	MAE		MSE	
BERT	0.4532		0.3159	
LSVC	0.6728		0.695	
RF	0.6266		0.6199	
FNN	0.533		0.4421	
CNN	0.5926		0.555	
	MAE	MSE	MAE	MSE
	+ readability		+ traditional	
LSVC	0.6627	0.6742	0.6664	0.6819
RF	0.5986	0.5743	0.609	0.5831
FNN	0.5024	0.4045	0.4823	0.3832
CNN	0.5896	0.5496	0.6041	0.5813
	+morphological		+ punctuation	
LSVC	0.6621	0.6775	0.664	0.6785
RF	0.6113	0.5917	0.6288	0.6204
FNN	0.5042	0.4002	0.5053	0.4102
CNN	0.5728	0.5269	0.5803	0.5307
	+ syntactic		+ frequency	
LSVC	0.6741	0.6924	0.6619	0.6703
RF	0.6167	0.5853	0.6401	0.643
FNN	0.4759	0.3705	0.7293	0.7627
CNN	0.5923	0.5566	0.5973	0.5602
	+ topic modeling		combined	
LSVC	0.6686	0.6861	0.6334	0.6166
RF	0.623	0.5986	0.568	0.5174
FNN	0.5156	0.4149	0.4658	0.3542
CNN	0.5882	0.5403	0.5408	0.4726

Table 15. Results for the OneStopEnglish corpus

Model	F	P	R
BERT	70.99	78.15	69.34
LSVC	70.41	72.15	72.03
RF	68.21	70.44	69.85
FNN	54	56.34	52.83
CNN	70.64	84.44	65.23

	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	70.49	72.17	72.02	69.89	71.76	71.69
RF	70.11	71.63	71.83	73.01	74.89	74.45
FNN	56.07	59.02	54.59	58.76	62.86	57.18
CNN	68.59	76.29	67.37	64.82	77.32	60.71
	+ morphological			+ punctuation		
LSVC	71.75	73.65	73.39	70.41	72.15	72.03
RF	70.67	72.22	72.25	68.92	70.24	70.4
FNN	62	65.37	60.19	55.79	57.56	54.8
CNN	69.02	78.87	66.33	64.33	75.55	60.33
	+ syntactic			+ frequency		
LSVC	70.54	72.61	72.37	71.34	73.1	73.04
RF	72.59	73.67	73.82	67.63	68.8	69.89
FNN	56.68	77.87	49.85	63.01	65.63	61.68
CNN	58.71	73.85	54.88	56.38	68.41	53.15
	+ topic modeling			Combined		
LSVC	67	68.9	69.14	71.44	72.96	73.07
RF	66.1	68.1	66.45	76.44	77.18	77.37
FNN	59.46	61.84	58.38	74.24	75.71	74.17
CNN	64.95	76.98	62.17	-	-	-

Article history:

Received: 20 October 2021

Accepted: 21 January 2022

Bionotes:

Dmitry A. MOROZOV is Junior Researcher at the Laboratory of Applied Digital Technologies, International Mathematical Center of Novosibirsk State University, Novosibirsk, Russia, and Developer at the Russian National Corpus. His spheres of interest include corpus linguistics, discrete math, math modeling, and machine learning.

Contact information:

Novosibirsk National Research State University

1 Pirogova, Novosibirsk, 630090, Russia

e-mail: morozowdm@gmail.com

ORCID: 0000-0003-4464-1355

Anna V. GLAZKOVA is Doctor of Sc. (Technology), Associate Professor of the Department of Software at the Institute of Mathematics and Computer Science of Tyumen University, Russia. Her current research interests include natural language processing and text mining, with a focus on text classification and deep learning.

Contact information:

University of Tyumen

6 Volodarsky, Tyumen, 625003, Russia

e-mail: a.v.glazkova@utmn.ru

ORCID: 0000-0001-8409-6457

Boris L. IOMDIN holds a Ph.D. in Philology and is a Leading Researcher at Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia. He is one of the authors and editors of the Active Dictionary of the Russian Language and a co-author of the project “Semantic Complexity of Words”.

Contact information:

Vinogradov Russian Language Institute of the Russian Academy of Sciences
18/2 Volkhonka, Moscow, 119019, Russia
e-mail: iomdin@ruslang.ru
ORCID: 0000-0002-1767-5480

Сведения об авторах:

Дмитрий Алексеевич МОРОЗОВ – младший научный сотрудник Лаборатории прикладных цифровых технологий Международного математического центра НГУ, Новосибирск, Россия. Сотрудник Национального корпуса русского языка. Область научных интересов: корпусная лингвистика, дискретная математика, математическое моделирование, машинное обучение.

Контактная информация:

Новосибирский национальный исследовательский государственный университет
Россия, 630090, Новосибирск, ул. Пирогова, д. 1
e-mail: morozowdm@gmail.com
ORCID: 0000-0003-4464-1355

Анна Валерьевна ГЛАЗКОВА – кандидат технических наук, доцент кафедры программного обеспечения Института математики и компьютерных наук Тюменского государственного университета, Тюмень, Россия. Область научных интересов: обработка естественного языка, интеллектуальный анализ текста, классификация текстов, глубокое обучение.

Контактная информация:

Тюменский государственный университет
Россия, 625003, Тюмень, ул. Володарского, д. 6
e-mail: a.v.glazkova@utmn.ru
ORCID: 0000-0001-8409-6457

Борис Леонидович ИОМДИН – кандидат филологических наук, ведущий научный сотрудник Института русского языка им. В.В. Виноградова Российской академии наук, Москва, Россия. Один из авторов и редакторов Активного словаря русского языка, соавтор проекта «Семантическая сложность слов».

Контактная информация:

Институт русского языка им. В. В. Виноградова РАН
Россия, 119019, Москва, ул. Волхонка, д. 18/2
e-mail: iomdin@ruslang.ru
ORCID: 0000-0002-1767-5480