Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 409–425
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30118

Research article

Collection and evaluation of lexical complexity data for Russian language using crowdsourcing

Aleksei V. ABRAMOV ✉, Vladimir V. IVANOV

Kazan Federal University, *Kazan, Russia*
✉AlVAbramov@stud.kpfu.ru

Abstract
Estimating word complexity with binary or continuous scores is a challenging task that has been studied for several domains and natural languages. Commonly this task is referred to as Complex Word Identification (CWI) or Lexical Complexity Prediction (LCP). Correct evaluation of word complexity can be an important step in many Lexical Simplification pipelines. Earlier works have usually presented methodologies of lexical complexity estimation with several restrictions: hand-crafted features correlated with word complexity, performed feature engineering to describe target words with features such as number of hypernyms, count of consonants, Named Entity tag, and evaluations with carefully selected target audiences. Modern works investigated the use of transforner-based models that afford extracting features from surrounding context as well. However, the majority of papers have been devoted to pipelines for the English language and few translated them to other languages such as German, French, and Spanish. In this paper we present a dataset of lexical complexity in context based on the Russian Synodal Bible collected using a crowdsourcing platform. We describe a methodology for collecting the data using a 5-point Likert scale for annotation, present descriptive statistics and compare results with analogous work for the English language. We evaluate a linear regression model as a baseline for predicting word complexity on handcrafted features, fastText and ELMo embeddings of target words. The result is a corpus consisting of 931 distinct words that used in 3,364 different contexts.

Keywords: *Lexical complexity, Russian language, annotation, corpora, Bible*

For citation:
Abramov, Aleksei V. & Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics* 26 (2). 409–425. https://doi.org/10.22363/2687-0088-30118

Научная статья

Сбор и оценка лексической сложности данных для русского языка с помощью краудсорсинга

А.В. АБРАМОВ ✉, В.В. ИВАНОВ

Казанский (Приволжский) федеральный университет, *Казань, Россия*
✉AlVAbramov@stud.kpfu.ru

Аннотация
Оценка сложности слова с помощью бинарной или непрерывной метки является сложной задачей, изучение которой проводилось для различных доменов и естественных языков. Обычно данная задача обозначается как идентификация сложных слов или прогнозирование лексической сложности. Корректная оценка сложности слова может выступать важным этапом в алгоритмах лексического упрощения слов. Представленные в ранних работах методологии прогнозирования лексической сложности нередко предлагались с рядом ограничений: авторы использовали вручную созданные признаки, которые коррелируют со сложностью слов; проводили детальную генерацию признаков для описания целевых слов, таких как количество согласных, гиперонимов, метки именованных сущностей; тщательно выбирали целевую аудиторию для оценки. В более современных работах рассматривалось применение моделей, основанных на архитектуре Transformer, для извлечения признаков из контекста. Однако большинство представленных работ было посвящено алгоритмам оценки для английского языка, и лишь небольшая часть переносила их на другие языки, такие как немецкий, французский и испанский. В данной работе мы представляем набор данных для оценки лексической сложности слова, основанный на Синодальном переводе Библии и собранный с помощью краудсорсинговой платформы. Мы описываем методологию сбора и оценки данных с помощью шкалы Лайкерта с 5 градациями; приводим описательную статистику и сравниваем ее с аналогичной статистикой для английского языка. Мы оцениваем качество работы линейной регрессии как базового алгоритма на ряде признаков: вручную созданных; векторных представлениях слов fastText и ELMo, вычисленных на основе целевых слов. Результатом является корпус, содержащий 931 словоформу, которые встречались в 3364 различных контекстах.

Ключевые слова: *лексическая сложность, русский язык, разметка, корпус, Библия*

Для цитирования:
Abramov A.V., Ivanov V.V. Collection and evaluation lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 409–425. https://doi.org/10.22363/2687-0088-30118

1. Introduction

This paper introduces a new dataset for lexical complexity prediction in Russian. Automatic predicting of lexical complexity can be useful in many areas such as readability assessment and text simplification (Dale 1948: 37–54, Devlin 1998). Typically, this task is formulated as mapping a word in a context with a complexity score on a certain scale. For instance, a selected word in a sentence may be assigned a binary label (complex/non-complex), or a score on the Likert scale (from 1 to 5). In recent works, this task has been studied in both multiple-domain settings, where lexical complexity depends on a subject domain of a text (e.g., biblical text, biomedical articles and proceedings of the European Parliament) and cross-lingual settings (e.g., English, German, Spanish) (Yimam 2018: 66–78). Basic parameters that can affect lexical complexity include a variety of lexical features, including word length, frequency features, character N-grams, and word embeddings[1]. The features that represent words as vectors can be used for fitting a machine learning model to the existing labeled dataset. A general approach of application machine learning models (such as Random Forest, Neural Network or Support Vector Machines) in Computational Linguistics and Natural Language Processing can be found in numerous monographs, including, but not restricted to (Manning & Schutze 1999, Nitin & Damerau 2010, Clark 2013).

Moreover, with the advances in machine learning and natural language processing (Delvin et al. 2018), pre-trained neural language models can be applied in the task of lexical complexity prediction in context (Shardlow 2021: 1–16). A comprehensive overview of computational linguistics methods applied in complexology can be found in (Solovyev et al. 2022). However, labeled datasets are still needed to fine-tune such models. At the same time, a task for multilingual lexical complexity prediction was studied for a limited number of languages. For instance, cross-lingual features for complexity prediction are studied at the level texts in (Morozov et al. 2022), while an neural approach is analyzed in (Sharoff 2022) Therefore, the main aim of the paper is to leverage existing methodology for the development of a Russian dataset for lexical complexity prediction.

We follow the methodology proposed in (Shardlow 2020: 57–62) which uses crowdsourcing to collect data. We investigate the statistical properties of the dataset to compare it with the English counterpart. The dataset contains 931 distinct words that occurred within 3,364 different contexts. Finally, we carried out a series of experiments for predicting lexical complexity with a simple linear function that uses lexical parameters of words as input and outputs a complexity score (so called linear regression model). The results of the model are close to the results of the same model trained on the English dataset.

2. Related works

In this section, we review the studies of lexical complexity prediction (LCP) focusing on two aspects: (i) dataset construction and (ii) baseline models evaluation. Since 2016, to evaluate methods for the lexical analysis, three shared tasks have been organized (Paetzold 2016: 560–569, Yimam 2018: 66–78, Shardlow 2021: 1–16). The first two initiatives address a very close problem of Complex Word Identification (CWI-2016 and CWI-2018), the latter one deals with the LCP task. In CWI-2016, the goal was to detect a complex English word in a context wherein a word is considered complex if it is difficult to understand for at least one of the annotators – non-native speakers. The training dataset had 2,237 instances, each labeled by 20 annotators, and the test dataset had 88,221 instances. Each word was assigned a binary label, naturally leading to a

[1] Word embedding is a representation of a word in the form of a numerical vector.

classification task. The participants experimented with lexical and statistical features available from the external sources, including Simple Wikipedia, as well as word embeddings. The feature sets served as an input to classifiers leveraging existing machine learning models. The post evaluation done in (Zampieri 2017: 59–63) has shown that the majority of the participating systems performed poorly mostly because of the data annotation flaws and the small size of the training dataset. In CWI-2018, the organizers proposed a new dataset aiming at both multilingual (English, German, French and Spanish) and multi-domain evaluation. In addition to the classification task, the participants of the CWI-2018 were able to solve another task, predicting a probability of the given target word in its particular context being complex (a regression problem).

The LCP-2021 dataset features an augmented version of CompLex, a multi-domain English dataset with texts from annotated using a 5-point Likert scale (1–5) (Shardlow 2020: 57–62) texts represent from three sources/domains: the Bible, Europarl (European Parliament), and biomedicine. The dataset covers 10,800 instances spanning three domains and containing unigrams and bigrams as targets for complexity prediction. The task was to predict the complexity value of words in a context (same tokens may appear in different contexts; on average each token has around 2 contexts). The LCP-2021 Shared task has two sub-tasks: predicting the complexity score for single words; and predicting the complexity score for multi-word expressions. For both subtasks the same performance measures were used to evaluate quality: correlations between human assessments and system results (here, the authors used two measures: Pearson's and Spearman's coefficients that show how well machine ranking corresponds to the human ranking of words)[2], and mean absolute and mean squared errors (MAE and MSE that correspond to average deviation between a score assigned by a machine and a score estimated from human judgements, respectively). The top-performing system (Yaseen 2021: 661–666), which applied modern models, where features are weighted token and context representations derived from very large neural networks that are pre-trained on multi-billion token text corpora, i.e., BERT (Delvin et al. 2018) and RoBERTa (Liu et al. 2019), reported 0.7886 Pearson Correlation in Task 1. However, there are 0.0182 points of Pearson's Correlation separating the systems at ranks 1 and 10. The LCP-2021 dataset has only English contexts, therefore this evaluation has not covered any other language except English. In the present paper, we develop a dataset for Russian using a methodology from (Shardlow 2020: 57–62) as closely as possible, because this can ease further multilingual and multi-domain lexical complexity evaluations. The methodology for data collection includes selecting target words and multi-word expressions (MWE) using predetermined frequency bands to ensure that targets are distributed across different ranges of low to high frequency. Automatic part-of-speech (POS) tagging was used for selecting nouns and MWEs that match certain patterns. In data labeling respect, the methodology leveraged a 5-point Likert scale with the

[2] Pearson's Correlation coefficient was used for final ranking of the results.

following descriptors: “Very easy”, “Easy”, “Neutral”, “Difficult”, “Very Difficult”. Each instance was annotated by 20 workers (annotators from English speaking countries) of a crowdsourcing platform. The labels for each word were transformed into a complexity score on a scale [0,1]. The resulting dataset had the average complexity for words equal to 0.395, with a standard deviation of 0.115. The subset of instances that were extracted from the Bible had the lowest average complexity score (0.387).

In the remaining part of this section, we review studies of CWI and LCP tasks. Most of the works have detailed descriptions of technical details that are more relevant in computer science than in linguistics. Therefore, we decided to focus on the review on the features (or parameters) that different methods use to model word complexity.

In (Yimam et al. 2017), the authors use four different language-independent sets of features: Length and frequency features, Syntactic features, Word embeddings features and Topic features. The authors used three length features: the number of vowels, the number of syllables, and the number of characters in a word; and three sets of frequency features: frequency of the word in Wikipedia, frequency of the word in the Google Web 1T 5-Grams, and frequency of the word in a context. A proper normalization for all the length and frequency features was performed. As syntactic features, the authors use the part of speech (POS) tags of words in different languages and map them into universal POS tags. As word embedding features, they use word2vec representations of content words (both complex and simple), in addition to cosine similarity between the vector representations of the word and its context. To compute topic-related features, the authors use a topic modeling technique LDA (Blei 2003) capable of representing each context as a distribution over topics, which in their turn are represented as distribution over words. The authors used 100 topics and computed cosine similarity between the word-topic vector and the document vector. The best classifiers trained on the described sets of features outperformed baseline results; however, feature analysis was not the primary goal in (Yimam 2017).

In (Kajiwara & Komachi 2018), the authors present their system that participated in the CWI-2018. They experimented with length features (Number of characters and Number of words for MWE) as well as with frequency features extracted from several corpora (Wikipedia, WikiNews, Lang-8). The authors evaluated the importance of features using ablation study on a classification task and found that the frequency features can yield better performance in comparison to probabilistic features extracted from the same corpora. The Lang-8 corpus seems to be more useful for their system than Wikipedia.

In (Aroyehum et al. 2018), the authors compared two approaches: feature engineering and a deep neural network. Both approaches achieved comparable performance on the English test set. The features sets used for training can be divided into several groups: Morphological Features, Syntactic and Lexical Features, Psycholinguistic Features, Word Embedding Distances that served as Features.

In (Malmasi et al. 2016), the authors of the LTG system focused on the use of contextual language model features and the application of ensemble classification methods. Both versions of their systems achieved good performance (second and third place in CWI-2016). They leveraged a core set of features based on estimating n-gram probabilities using web-scale language models from the Microsoft Web N-Gram Service. These probabilities fall into three groups: Word Probability (how likely it is that the word is present in the corpus), Conditional Probability (how likely it is that the word is present in the corpus given the immediate previous word), and Joint Probability (how likely it is that the pairs and triples of words are in the corpus). All of these probabilities help a system modeling the context in which a word appears. In later works on CWI and LCP such information was represented using word embeddings. In addition, the authors use the length of a word as a feature.

A number of novel features were proposed in (Aprosio 2020). Their approach based on the user's native language identifies complex terms by automatically detecting cognates and false friends, using distributional similarity computed from fastText (Bojanowski 2017: 135–146) word embeddings. Similar types of features are used in (Zaharia 2020). To calculate similarity measures between words, the authors apply a technique presented in (Conneau 2017) to learn a linear mapping of two vector spaces that represent monolingual fastText word embeddings (e.g., between Spanish and German) into the same vector space.

The MacSaar (Zampieri 2016) system presented in CWI-2016 based on a simple idea – observing Zipfian frequency distributions computed from text corpus – helps to determine whether a word is complex or simple. The authors calculate the *Zipfian frequency* feature by taking the inverse of the rank of a word. Additionally, word length, normalized sum probability of the character trigrams in a word, sentence length and sum probability of the character trigrams of the sentence were used in their experiments.

In 2021, a number of models and features were evaluated in the new LCP-2021 Shared task in (Shardlow 2021: 1–16). First, we should mention that top-performing systems for lexical complexity prediction used context by means of contextualized pre-trained language models. Those systems, as mentioned above, use deep learning models that make use of the Transformer architecture which in recent years has disrupted the field of natural language processing (Vaswani 2017). During pre-training, such language models are forced to use context in order to reconstruct missing words in a large corpus (usually, multi-billion tokens corpora). In the LCP-2021, the participants used BERT-based models: BERT (Delvin 2018), RoBERTa (Liu 2019), ELECTRA (Clark 2020), ALBERT (Zhenzhong 2019), DeBERTa (He 2020) to encode (i.e., to represent in the form of vectors) both a target word and the input context of the word. Other systems used a variety of features, including lexical frequency and length features, psycholinguistic features that represent human perception of words, semantic features from WordNet to represent word ambiguity or abstractness. The third group of systems combined the deep learning models with the models trained on engineered feature sets. An extensive exploration of sentence and word features are presented in (Mosquera 2021), where the author investigates feature engineering methods for predicting the complexity of English words in a context using regression models. A substantial set of 51 features was studied, including Word and Lemma lengths, Syllable count, Morpheme length (a number of morphemes for the target word), Google frequency (the frequency of the target word based on Google ngram corpus), two Wikipedia-based word frequencies (one based on the target word occurrences and the other based on the number of documents in Wikipedia where the target word appears), Complexity score taken from a complexity lexicon (Maddela & Xu 2018), Zipf frequency, two Kucera-Francis frequencies: for a target word and for the target word lemma, binary features (is_stopword and is_acronym), Average age of acquisition, Average concreteness, Word and Lemma frequencies in COCA, WordNet-related features (Number word senses, synonyms, hypernyms, hyponyms), Minimum and maximum distances to the root hypernym in WordNet for the target word, Number of Greek or Latin affixes, Year of appearance (the first year when the target and its preceding word appeared in the Google Books Ngram Dataset), as well as a number of SUBTLEX-based features and various readability scores (such as SMOG index, Dale-Chall index, Gunning-Fog, etc.). A list of top ten important features includes age of acquisition, Dale-Chall index, Zipf frequency, average concreteness and lemma frequency.

3. Methodology

Following the methodology proposed for the English language in (Shardlow 2020: 57–62), we chose a Russian parallel translation of the Bible from (Christodouloupoulos 2015: 375–395), based on the Russian Synodal Bible, as the initial corpus. For annotation we selected nouns listed in the Frequency dictionary of modern Russian language (Lyashevskaya 2009), that fall within the following frequency intervals (ipm, instances per million): (2-4), (5-10), (11-50), (51-250), (251-500), (501-1400), (1401-3100). Such restrictions on the choice of part of speech and specific frequency intervals provide us with a basis for a fair comparison with the original methodology. The selection of suitable nouns was performed in such a way that the number of words in each frequency interval was approximately the same for the first four intervals and decreased with the growth of frequency for the rest. We selected 931 distinct words that occurred within 3,364 different contexts. Each word was provided with a surrounding context, such as a Bible verse.

The assessors were asked to estimate the lexical complexity of a highlighted word in a given context using five-level Likert scale with the following items:

1. Very easy: the meaning of the highlighted word is clear;
2. Easy: the meaning of the highlighted word is obvious and the context supplements it;

3. Average[3]: the meaning of the highlighted word is familiar, but it becomes clear only after taking into account the surrounding context;

4. Difficult: the meaning of the highlighted word is not evident, but might be understood after considering the context;

5. Very difficult: the meaning of the highlighted word is unclear or the word itself is unfamiliar.

Compared to the data labeling procedure described for CompLex, we decided to present a more detailed description for each item of the scale, particularly, in terms of impact of the context on the understanding of the word meaning. A detailed explanation for each item could simplify the lexical complexity evaluation for the assessors and the subsequent analysis of the answers.

The words and their surrounding contexts were grouped into samples as in the following sample: "*Их конец – погибель, их бог –* **чрево***, и слава их – в сраме, они мыслят о земном*" ("*Whose end is destruction, whose God is their* **belly***, and whose glory is in their shame, who mind earthly things*"), where the target word is bold type and its context is marked with italics. The collected samples were shuffled and divided into batches of 10 samples each to ensure that every batch had samples with different lexical complexity. Additionally, we split batches into 12 task pools with 30 batches each, except for the last one with 7 batches. Every batch was shown to 10 distinct annotators, so that every word with a corresponding context was evaluated 10 times. We selected assessors from Russia, Ukraine, Belarus and Kazakhstan to introduce speakers with different language skills. A more detailed information about their native language and experience of using Russian could be useful, but unfortunately we were not able to collect such data from the crowdsourcing platform (Yandex.Toloka). To filter assessors with reliable assessments and to gather various opinions, we used the following automatic rules:

- Limited daily earnings: if the assessor completed five tasks per day, he (she) would be suspended for 24 hours;
- The number of skipped assignments: if the assessor skipped more than two assignments in a row, he (she) would be banned for three days;
- Captcha: if at least three out of five last captchas were not recognized, the assessor would be banned for seven days;
- Limit on response time: if at least two out of five latest assignments were completed in less than 15 seconds, the assessor would be banned for seven days;
- Majority vote: if more than five out of the last ten assignments were completed with responses different from the majority (minimum five similar responses), the assessor would be banned for seven days.

We selected the top 10% of the available assessors and paid 10 cents for each evaluated batch. All the gathered evaluations were transformed into [0,1] range and averaged per sample. Examples of words in a context, corresponding complexities and score variance are listed in Table 1 above.

[3] In Russian we use the descriptor "Средняя сложность" (moderate, medium) that better corresponds to the original descriptor "Neutral".

Table 1. **Samples from corpus; target words are in *bold type***

Samples	Complexity	Variance
При **выходе** их из Иудейской синагоги язычники просили их говорить о том же в следующую субботу. (And when the Jews **were gone out** of the synagogue, the Gentiles besought that these words might be preached to them the next sabbath).	0.075	0.11
Моисей весьма огорчился и сказал Господу: не обращай ***взора*** Твоего на приношение их; я не взял ни у одного из них осла и не сделал зла ни одному из них. (And Moses was very wroth, and said unto the Lord, ***Respect*** not thou their offering: I have not taken one ass from them, neither have I hurt one of them).	0.28	0.175
Никакое гнилое слово да не исходит из уст ваших, а только доброе для ***назидания*** в вере, дабы оно доставляло благодать слушающим. (Let no corrupt communication proceed out of your mouth, but that which is good to the use of ***edifying***, that it may minister grace unto the hearers).	0.4	0.19
Услышав об этом, все бывшие в башне Сихемской ушли в башню ***капища*** Ваал-Верифа. (And when all the men of the tower of Shechem heard that, they entered into an hold of the ***house of the god*** Berith).	0.63	0.26

It took 60.4 seconds on average to annotate one batch of samples and 135 assessors on average to complete the task pool. Each assessor annotated 2.19 batches of samples. We did not use any training or control tasks due to the following reasons: 1) the evaluation of lexical complexity is subjective and depends on various factors, such as education, occupation, overall erudition, age (some modern words might be more familiar to a younger audience), and language proficiency (in our research we also included annotations gathered from non-native speakers); thus we cannot reliably provide "correct" answers for tasks to estimate one's accuracy; 2) the use of averaged or majority's answers as ground truth could narrow down the amount of available assessors to those who have similar views on lexical complexities of different words; therefore, we would not be able to estimate the true distribution of lexical complexities performed by people with different background.

4. Analysis

We conducted the distribution analysis of the obtained lexical complexities by estimating their distribution and connection with the word frequency. Figure 1 contains histograms of lexical complexity scores from (Shardlow 2020: 57–62) and our work.

It can be observed that there was a median complexity score equal to 0.225 (denoted as a vertical blue dashed line), wherein the majority of given evaluations are equal to either "Very easy" or "Easy", according to the aforementioned scale. This is consistent with the well known dependency between lexical complexity and word frequency; uncommon words tend to have a higher complexity; therefore, truly rare and difficult words are harder to obtain and less likely to fit in our frequency ranges.

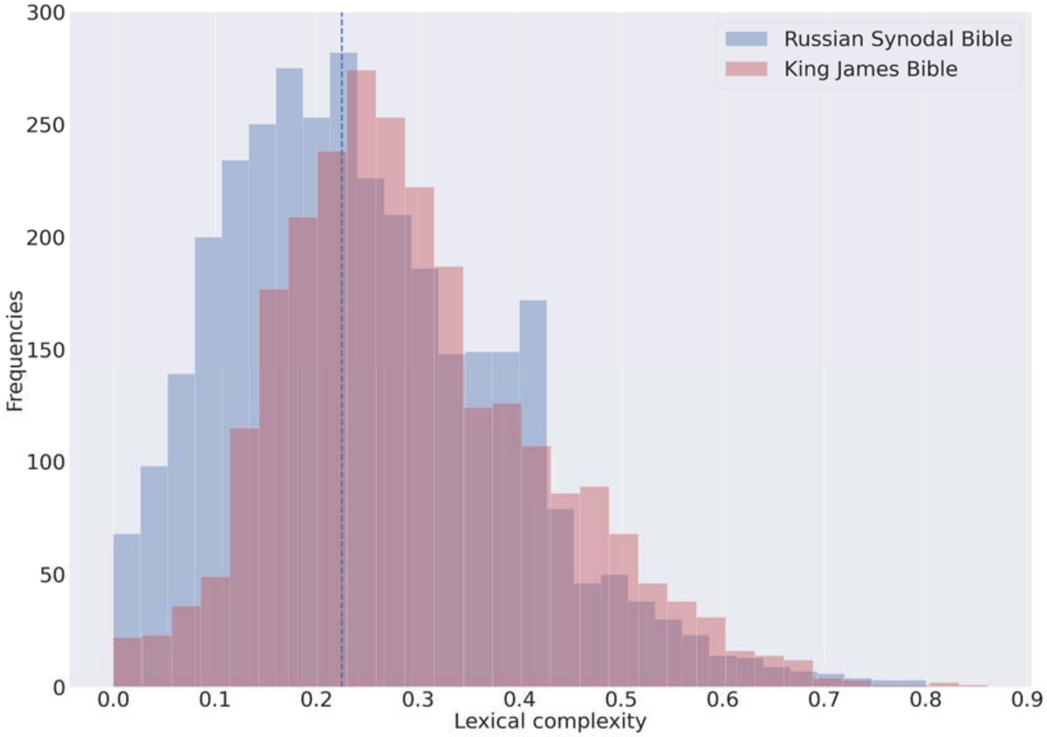

Figure 1. Lexical complexity scores distribution for words selected from King James Bible and Russian Synodal Bible

We also noticed a non-linear dependency between lexical complexity and word frequency; an estimated lexical complexity remains mostly the same almost among all frequency ranges, but starts to rise when close to the lowest frequencies. Indeed, the Pearson correlation between lexical complexity and word frequency is moderately low ($\boldsymbol{r_f}$ = –0.32), albeit significant. Additionally, we observed a weak positive correlation between lexical complexity and word length: $\boldsymbol{r_w}$ = 0.14. A similar dependency can be observed for the CompLex dataset with a weak negative correlation between lexical complexity and word frequency ($\boldsymbol{r_f}$ = –0.24) and slightly stronger positive correlation between lexical complexity and word length ($\boldsymbol{r_w}$ = 0.28).

Figures 2 and 3 contain randomly sampled subsets of the corpuses that illustrate such phenomena; the x-axis is depicted in log-scale, lexical complexities are averaged per lemma. This shows that lexical complexity depends on other word features as well, such as length, number of syllables, morphological structure, context, meaning ambiguity, etc.

We also noticed a dependency between word’s frequency and variance of lexical complexity scores from different annotators. Figure 3 illustrates this observation, i.e., the variance of complexity scores within certain frequency ranges decreases as range boundaries increase. These results can be explained by the following reason – the less frequent (and, hence, more complex) a word is, the fewer annotators are familiar with it, which translates into higher complexity scores from annotators who are unfamiliar with the meaning of the word or unable to derive it from the context. But we did not observe the same dependency for the CompLex dataset as shown in Figure 4.

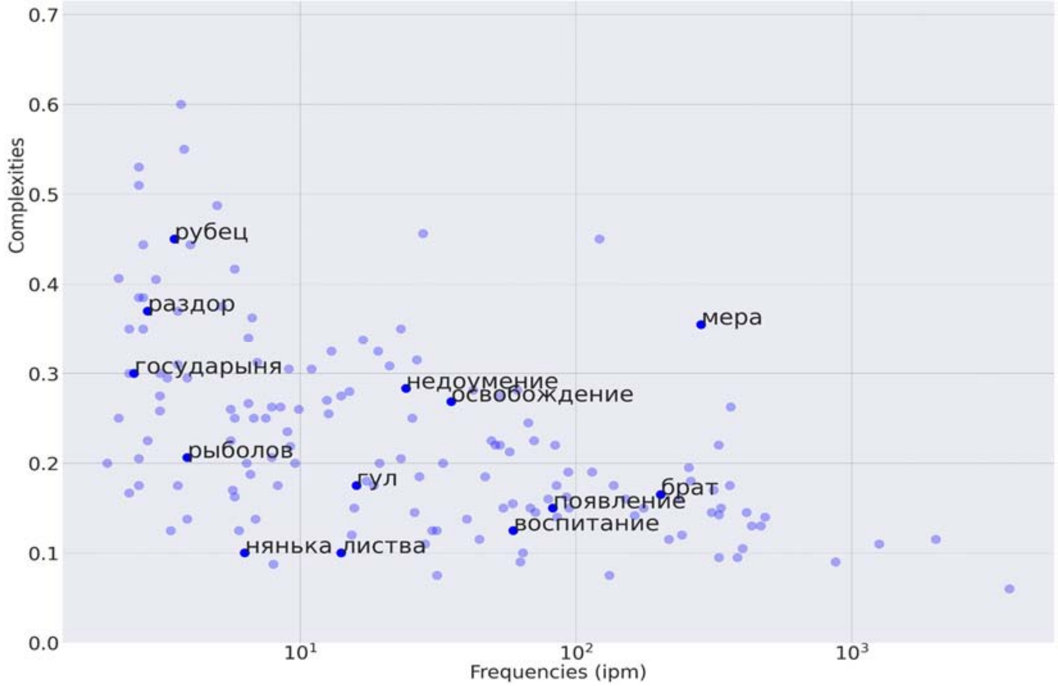

Figure 2. Dependency between word frequency and lexical complexity for words from Russian Synodal Bible

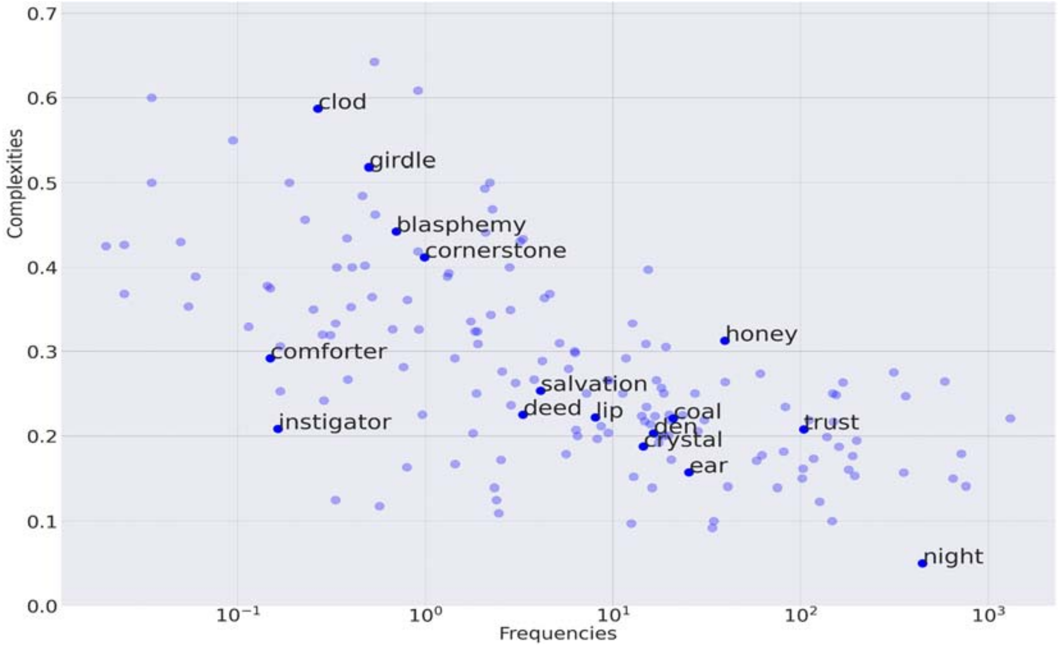

Figure 3. Dependency between word frequency and lexical complexity for words from King James Bible

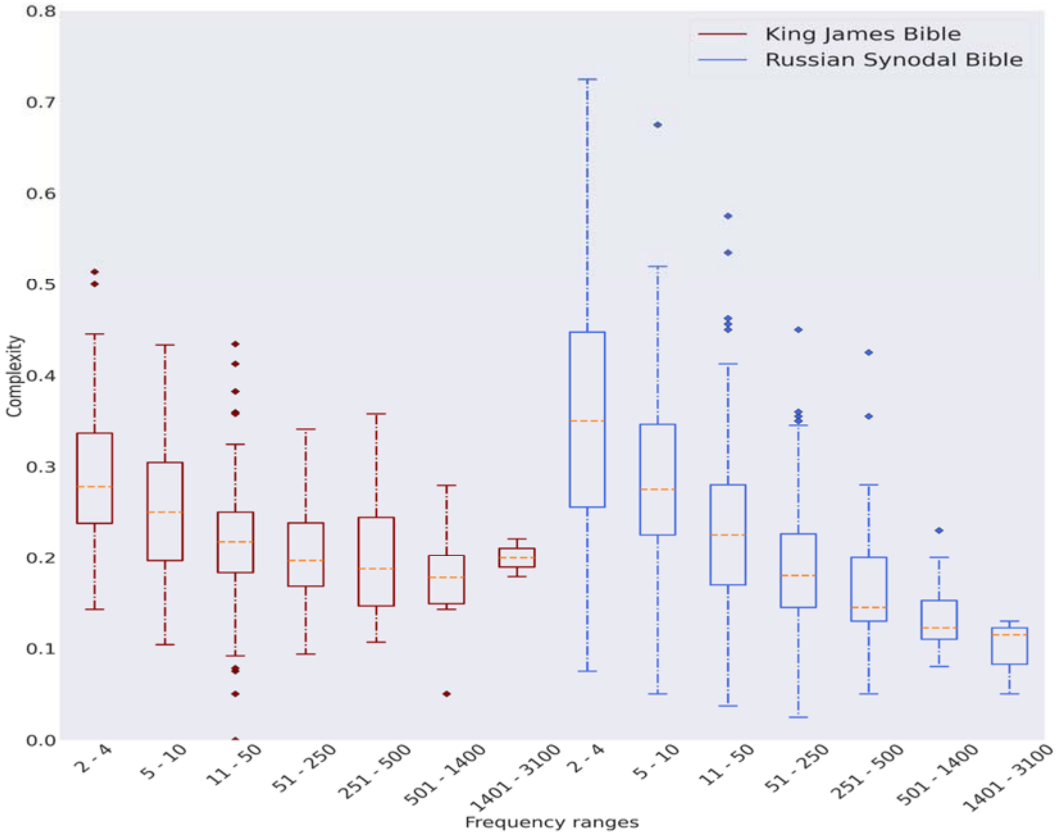

Figure 4. Box plots showing the distribution of lexical complexity scores of words grouped by their frequency from King James Bible (*left*) and Russian Synodal Bible (*right*)

Table 2. **The result of linear regression on handcrafted features (HC), Fasttext and ELMo embeddings and concatenated features**

	Handcrafted	Fasttext	ELMo	Fasttext+HC	ELMo+HC
MAE	0.102	0.084	0.099	0.084	0.099
Pearson correlation	0.342	0.614	0.498	0.619	0.501

Table 3. **Pearson correlation between handcrafted features**

	Frequency	Word length	Number of syllables
Frequency	1	-0.206	-0.172
Word length	-0.206	1	0.819
Number of syllables	-0.172	0.819	1

5. Experiments

To investigate how simple features, such as word frequency and its length, affect lexical complexity, we created a simple baseline. For comparison with CompLex, we selected a linear regression as our model with the following three features: (i) word length, (ii) word frequency according to Lyashevskaya's dictionary and (iii) number of syllables. These features were mentioned by (Shardlow 2020: 57–62) as hand-crafted (HC) features. On target words from the

Complex, linear regression achieved Mean Absolute Error of 0.0888 (for the same set of features). In Table 3, we provide the pairwise correlations between the three HC features in our dataset. In addition, we fitted linear regression using fastText and ELMo embeddings – N-dimensional vector representations of words trained on a large unannotated corpus (Bojanowski 2017: 135–146, Peters 2018: 2227–2237).

FastText embeddings were pretrained on the joint Russian Wikipedia and Lenta.ru corpora; ELMo embeddings were pretrained on the Russian WMT-News corpora; both were taken from DeepPavlov repository (Burtsev M. 2018: 122–127). In our case, we used 300-dimensional embeddings from fastText and 1024-dimensional embeddings from ELMo. For a complete comparison we concatenated embeddings and handcrafted features and applied linear regression as well. As evaluation metrics, we selected Mean Absolute Error and Pearson correlation. Final results were averaged using 10-fold cross-validation.

6. Discussion

The main novel contribution of the work is a new dataset for word-level complexity evaluation in Russian. At present we are not aware of any other resources with a comparable size or coverage. We claim that the dataset also has a comparable quality to its English counterpart. This claim can be supported by a comparison of the complexity scores distribution and the experiments we carried out with the baseline models for lexical complexity prediction. Indeed, this was expected because we applied the same principles to collect and label the data, which led to very similar results. For instance, Figures 4 and 5 illustrate similar behavior for variance of complexity scores, which decay when the word frequency grows. Moreover, experiments with the linear regression model trained on the similar feature sets show similar results (Table 2): on the English dataset MAE value for hand-crafted features was 0.089, while for Russian it is 0.100; training with word embeddings as features provides almost identical results.

Despite these positive findings, we need to mention a few substantial differences between Russian and English datasets. First, complexity score histograms for Russian and English are shifted relative to each other (see Fig. 1); overall, the Russian version contains simpler words. Second, the correlation between word frequency and complexity in the Russian dataset (–0.32) differs from its English counterpart, wherein the correlation coefficient is slightly weaker (–0.24). This histogram shift and the discrepancy in correlation coefficients can be explained by the fact that the King James Bible was published long before the Russian Synodal edition of the Bible and contains more deprecated words and expressions compared to the Russian edition. Hence, the Russian data have simpler labels than the English data.

Our dataset has a few limitations, including a coverage restricted to a single domain (Bible texts) and only single words, without multi-word expressions. We are aiming to overcome the first limitation in our future work, as the methodology that we made use of is already well-studied and has proved to be successful. The second limitation (lack of MWEs) seems to be important, but less urgent. The LCP-2021 evaluation shows that thye prediction of single word complexity in a context is harder than the MWE complexity prediction.

7. Conclusion

In this paper, we presented a novel dataset for predicting lexical complexity in the Russian language. The dataset has 931 distinct words that occurred within 3,364 different contexts. It was labeled using a crowdsourcing platform (Yandex Toloka). During data collection and labeling we followed a well-studied methodology previously applied in English. We compared our dataset with its English counterpart by two means: 1) we analyzed statistical properties of both datasets; 2) we trained a linear regression model on Russian data and compared its outcomes to its English analog. We found a few discrepancies between datasets which are viewed as potential targets of our further investigation. In our future experiments with the dataset, we expect to develop better models and study extensive feature sets for predicting lexical complexity, which might be important in a broader context of text and discourse complexity studies, as well as the development of automatic complexity analyzers (Solnyshkina et al. 2022).

Acknowledgments

This paper has been supported by the Russian Science Foundation, grant # 22-21-00334, https://rscf.ru/project/22-21-00334/.

REFERENCES

Aprosio, Alessio P., Stefano Menini & Sara Tonelli. 2020. Adaptive complex word identification through false friend detection. *In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization.* 192–200. https://doi.org/10.1145/3340631.3394857

Aroyehun, Segun Taofeek, Jason Angel, Daniel Alejandro Pérez Alvarez & Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. *In Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications.* 322–327. https://doi.org/10.18653/v1/W18-0538

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146.

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 Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva & Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. *Proceedings of ACL 2018, System Demonstrations*. 122–127. https://doi.org/10.18653/v1/P18-4021

Christodouloupoulos, Christos & Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation* 2(49). 375–395. https://doi.org/10.1007/s10579-014-9287-y

Clark, Alexander, Chris Fox & Shalom Lappin (eds.). 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le & Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *In Proceedings of the International Conference on Learning Representations.*

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer & Hervé Jégou. 2017. Word translation without parallel data. *In Proceedings of the International Conference on Learning Representations.*

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin* 27. 37–54.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. 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* (1). 4171–4186. https://doi.org/10.18653/v1/N19-1423

Devlin, Siobhan & John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*. 161–173.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *In Proceedings of the International Conference on Learning Representations.*

Kajiwara, Tomoyuki & Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. *In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 195–199.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma & Radu Soricut.2019. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019).

Lyashevskaya, Olga N. & Sergey A. Sharoff. 2009. *The Frequency Dictionary of Modern Russian Language.* Moscow: Azbukovnik. (In Russ.)

Maddela, Mounica & Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* 3749–3760. https://doi.org/10.18653/v1/D18-1410

Malmasi, Shervin, Mark Dras & Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 996–1000. https://doi.org/10.18653/v1/S16-1154

Manning, Christopher & Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing.* MIT press.

Morozov, Dmitry, Anna Glazkova & Boris Iomdin. 2022. Text Complexity and Linguistic Features: their correlation in English and Russian. *Russian Journal of Linguistics* 26 (2). 425-447.

Mosquera, Alejandro. 2021. Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 554–559. https://doi.org/10.18653/v1/2021.semeval-1.68

Nitin, Indurkhya & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd edn. Boca Raton: CRC Press.

Paetzold, Gustavo & Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 560–569. https://doi.org/10.18653/v1/S16-1085

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Shardlow, Matthew, Michael Cooper & Marcos Zampieri. 2020. CompLex – A New corpus for lexical complexity prediction from Likert Scale Data. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI).* 57–62.

Shardlow, Matthew, Richard Evans, Gustavo Henrique Paetzold & Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 1–16. https://doi.org/10.18653/v1/2021.semeval-1.1

Sharoff, Serge. 2022. What neural networks know about linguistic complexity? *Russian Journal of Linguistics.* 26(2). 370–389.

Solnyshkina, Marina, Mcnamara Danielle & Zamaletdinov Radif. 2022. Natural language processing and discourse complexity studies. *Russian Journal of Linguistics.* 26(2). 317–341.

Solovyev, Valery, Marina Solnyshkina & Mcnamara Danielle. 2022. Computational linguistics and Discourse complexology. *Russian Journal of Linguistics.* 26(2). 275–316.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems.* 5998–6008.

Yaseen, Tuqa Bani, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh & Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 661–666. https://doi.org/10.18653/v1/2021.semeval-1.85

Yimam, Seid Muhie, Sanja Stajner, Martin Riedl & Chris Biemann. 2017. Multilingual and cross-lingual complex word identification. *In Proceedings of the International Conference Recent Advances in Natural Language Processing.* 813–822. https://doi.org/10.26615/978-954-452-049-6_104

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack & Marcos Zampieri. 2018. A report on the complex word identification shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA).* 66–78. https://doi.org/10.18653/v1/W18-0507

Zaharia, George-Eduard, Dumitru-Clementin Cercel & Mihai Dascalu. 2020. Cross-lingual transfer learning for complex word identification. *In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI).* 384–390. https://doi.org/10.1109/ICTAI50040.2020.00067

Zampieri, Marcos, Liling Tan & Josef van Genabith. 2016. Macsaar at semeval-2016 task 11: Zipfian and character features for complex word identification. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 1001–1005. https://doi.org/10.18653/v1/S16-1155

Zampieri, Marcos, Shervin Malmasi, Gustavo Paetzold & Lucia Specia. 2017. Complex word identification: Challenges in Data Annotation and System Performance. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017).* 59–63.

Article history:
Received: 27 October 2021
Accepted: 21 January 2022

Bionotes:
Aleksei V. ABRAMOV is a PhD student at Kazan (Volga region) Federal University (Russia). His research interests comprise applied linguistics, use of computer technologies in the creation of corpora and methods of assessment texts and words complexity.
Contact information:
Kazan (Volga region) Federal University
18 Kremliovskaya str., Kazan, 420008, Russia
e-mail: AlVAbramov@stud.kpfu.ru
ORCID: 0000-0002-5509-9680

Vladimir V. IVANOV is Assistant Professor at Innopolis University (Russia). His scientific interests lie in the field of applied computational linguistics (text complexity analysis, information extraction from text), development and application of knowledge bases and knowledge graphs, as well as the application of machine learning methods in software engineering.
Contact information:
Innopolis University
1 Universitetskaya st., Innopolis, 420500, Russia
e-mail: v.ivanov@innopolis.ru
ORCID: 0000-0003-3289-8188

Сведения об авторах:
Алексей Валерьевич АБРАМОВ – аспирант Казанского (Приволжского) федерального университета. К сферам его научных интересов относятся прикладная лингвистика и применение компьютерных технологий в создании корпусов и методов, посвященных оценке сложности понимания отдельных слов и текстов.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: AlVAbramov@stud.kpfu.ru
ORCID: 0000-0002-5509-9680

Владимир Владимирович ИВАНОВ – доцент Университета Иннополис (Россия). Его научные интересы включают прикладную компьютерную лингвистику (анализ сложности текста, извлечение информации из текста), разработку и применение баз знаний и графов знаний, применение методов машинного обучения в разработке программного обеспечения.
Контактная информация:
Университет Иннополис
Россия, 420500, Иннополис, ул. Университетская, д. 1
e-mail: v.ivanov@innopolis.ru
ORCID: 0000-0003-3289-8188