ISSN 2007-9737

Lexical Complexity Evaluation based on Context for Russian Language

Aleksei V. Abramov[1], Vladimir V. Ivanov[2], Valery D. Solovyev[1]

[1] Kazan Federal University,
Instiute of Computational Mathematics and Information Technologies,
Russian Federation

[2] Innopolis University,
Institute of Software Development and Software Engineering,
Russian Federation

alvabramov@stud.kpfu.ru, v.ivanov@innopolis.ru, maki.solovyev@mail.ru

Abstract. The task of identifying complex words within a context usually referred to as Complex Word Identification (CWI) or Lexical Complexity Prediction (LCP), is a vital component in Lexical Simplification pipelines. Correctness of complexity estimation depends on presented features, i.e. hand-crafted features, word embeddings, and presence of surrounding context, as well as on exploited rules or models, i.e. manually designed filtering, classic machine learning models, recurrent neural networks, and Transformer-based models. To our knowledge, the majority of existing works in CWI and LCP areas are devoted to investigating properties of English words and texts, accompanied by studies of German, Spanish, French and Hindu languages with little to no attention to Russian. In this paper, we present a study on lexical complexity estimation for the Russian language, by investigating the following topics: how well do morphological, semantic, and syntactic properties of a word represent its complexity; does a surrounding context significantly affect the accuracy of complexity estimation. We provide a brief description of the dataset of lexical complexity in context based on the Russian Synodal Bible and expand it by presenting a dataset of morphological, semantic, and syntactic features for annotated words. Additionally, we present linear regression and RuBERT models as baselines for lexical complexity estimation respectively.

Keywords. Lexical complexity, Russian language, Bible, corpus, Wiktionary.

1 Introduction

The task of Complex Word Identification (CWI) or Lexical Complexity Prediction is considered to be a challenging one not only due to the intricacy of word complexity estimation itself but also due to the ambiguity of annotations and lack of well-annotated and prepared data in various domains and scenarios, which limits our capability to build qualitative models and explore intrinsic dependencies.

Throughout the time several works were presented aiming to investigate different methods of CWI or LCP. Initially, automatic estimation of complexity was used as part of lexical simplification pipelines, by formulating it as a task of binary classification. More recent works suggested using a continuous label for word complexity, i.e., normalized score from the Likert scale. A basic approach to CWI included the creation of special lists of complex words or an approximation of word complexity with its frequency.

More sophisticated methods included basic machine learning models, i.e. Linear Regression, Logistic Regression, Support Vector Machines, Random Forest; intrinsic feature extraction models, i.e. word2vec [62], GloVe [63], fasttext [64]; modern Transformer-based models, i.e. BERT [57], RoBERTa [58], DeBERTa [65], ELECTRA [66], ALBERT [59], ERNIE [60].

In addition, recent works studied multi-lingual (English, Spanish, German, French, Chinese, Japanese, and Hindi) and multi-domain (biblical, biomedical texts, proceedings, or European Parliament) setups in order to provide extensive

research of language and domain impact on complexity evaluation.

This paper aims to extend amount of data available for the Russian language by presenting a supplement to the existing dataset of lexical complexity in context [1] by collecting morphological, semantic, and syntactic features using Russian Wiktionary and validating the correctness of corpora by providing a comparison of simple baselines, such as linear regression and RuBERT [2], trained on several variants of dataset - the one with hand-crafted features, the one with only target words themselves and the last one with surrounding contexts included.

We employ the following methodology: we collected a set of predefined features from available articles from the Russian Wiktionary and filtered features with a high number of missing data. We evaluated the importance of features and employed linear regression as a baseline for setups with various combinations of Wiktionary, fasttext, and handcrafted (e.g., word frequency and length) features. Additionally, for comparison, we evaluated RuBERT in two settings - with and without surrounding contexts for target words. We present a comparison of metrics on the aforementioned baselines.

The gathered results show clear evidence of the importance of complexity evaluation in the presence of surrounding contexts and the non-linear nature of word complexity in relation to word features.

2 Related Works

History of the field of study for CWI and LCP can be traced back to the middle of the XX century. Initially, tasks of complexity assessment were formulated for texts with a focus on readability estimation or text simplification.

In [3] authors presented a formula for predicting text readability and, later, in [4] revisited a previously proposed formula with the updated list of familiar words and criteria. For text simplification purposes, [5] used psycholinguistic features for the detection of simplification candidates; [6] applied proposed rules to develop an automatic system for practical simplification of English newspaper texts and aimed to assist aphasic readers.

With the development of natural language processing tools, modern downstream tasks focused on estimating the complexity of distinct words within texts or separate sentences. Originally, the CWI task was formulated as a ranking task. In LS-2012 [7], participants were asked to build automatic systems for word ranking from the simplest to the most difficult.

Most participants relied on hand-crafted features, such as frequency, n-grams, morphological, syntactic, and psycholinguistic properties [8], [9], [10], [11]. Formulating the CWI task the following way allows us to obtain a higher inter-annotator agreement, thus, leading to the more correct estimation of word complexity with relation to its synonyms and neighbors, but, on the other hand, doesn't provide an absolute complexity score for each word [12].

This formulation was used in the more recent works in Lexical Simplification pipelines, i.e. in [13], where authors combined Newsela corpus [14] with context-aware word embeddings and trained neural ranking model to estimate complex words and suitable substitutions.

In the shared-task CWI-2016 [15], authors and participants addressed a newer formulation of the CWI task as a prediction of binary complexity score, which was originally proposed in [16] and in [17]. In [17] the author presented a dataset of annotated complex words and their simpler substitutions, as well as, a system for the detection and simplification of words with available relevant substitutions.

Though this dataset presented relevant estimations of word complexity, it could not reflect the complexity perception of non-native speakers. In CWI-2016 authors presented a dataset of sentences from Wikipedia, annotated by 400 non-native speakers.

Participants of the shared task from 21 teams presented 42 distinct solutions, mostly based on classic machine learning models, e.g., SVM [18, 19, 20], Decision trees [18, 21, 22], Ensemble methods [19, 23, 24, 25, 26, 27]. The best results were demonstrated by ensemble methods that were able to reflect a non-linear dependency between complexity and word features. Surprisingly, neural networks [34] did not demonstrate outstanding results in this competition.

Table 1. An example of dataset preprocessing

Context and target word (in bold)	Complexity score
So Gad came to David, and told him, and said unto him, Shall seven years of famine come unto thee in thy **land**? or wilt thou flee three months before thine enemies, while they pursue thee? or that there be three days' pestilence in thy land? now advise, and see what answer I shall return to him that sent me.	0.15
The black horses which are therein go forth into the north **country**; and the white go forth after them; and the grisled go forth toward the south country.	0.175
And Moses sent them to spy out the land of Canaan, and said unto them, Get you up **this way** southward, and go up into the mountain:	0.05
And David and his men went up, and invaded the Geshurites, and the Gezrites, and the Amalekites: for those nations were of old the inhabitants of the **land**, as thou goest to Shur, even unto the land of Egypt.	0.025
He shall enter peaceably even upon the fattest **places of the province**; and he shall do that which his fathers have not done, nor his fathers' fathers; he shall scatter among them the prey, and spoil, and riches: yea, and he shall forecast his devices against the strong holds, even for a time.	0.05
Target lemma: **land/country/this way/places of the province** (in the Synodal Bible - **country**)	0.09

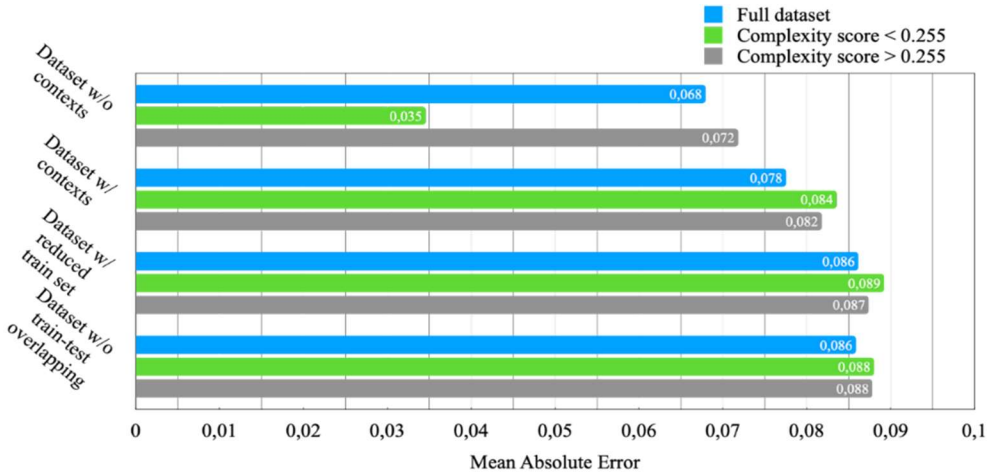

Fig. 1. Mean Absolute Error of complexity score predictions (the lower - the better) for experiments in different setups. Best viewed in color

In the shared ask CWI-2018 [28], the authors addressed a problem of word complexity estimation in multilingual and multi-genre setups.

They collected datasets in four different languages: English, German, Spanish and French and, additionally, presented three datasets within the same domain but with different expected complexity for English - news articles, written by professionals, amateurs, and Wikipedia editors. Moreover, in addition to the task of binary classification with regard to word complexity, the authors presented a track with probability estimation of a word being complex, which introduces a continuous complexity label. English, German and Spanish were used in monolingual tracks, and French was used as a test set in multilingual one. As in CWI-2016 participants mostly used classic machine learning models, e.g., SVM [29, 30] Ensemble algorithms [31, 32, 33, 30, 35], and Neural Networks [32].

An introduction of the continuous complexity label addressed the main issue with binary

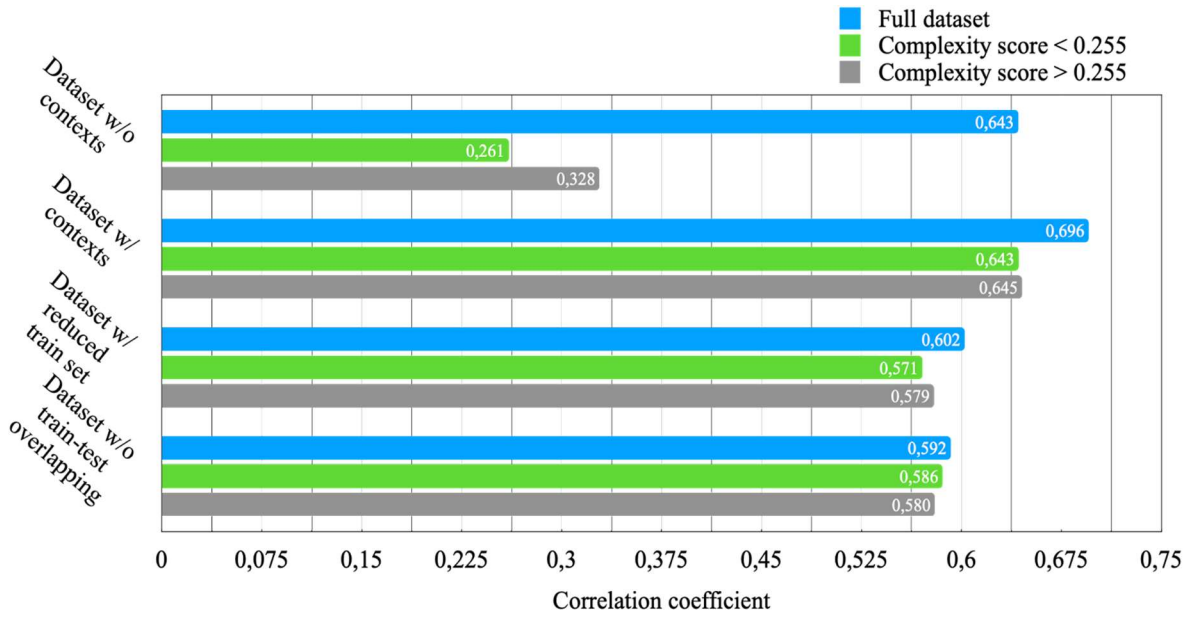

Fig. 2. Pearson Correlation Coefficient of predicted complexity scores with target scores (the higher - the better) for experiments in different setups. Best viewed in color

complexity - its inability to represent a whole spectrum of lexical complexity, which is subjective for every annotator.

In [36, 37] authors demonstrated that the utilization of binary complexity scores might lead to low inter-annotator agreement.

In order to fully investigate this problem, in LCP-2021 [38] authors presented a multi-genre dataset, divided between 3 domains: Bible [39], biomedical articles [40], and Europarl [41]. In addition to single words, authors included multi-word expressions (MWE). Both single words and MWE were presented in context for annotators and participants.

A continuous score was calculated as an average of complexity estimations received from annotators with help of a 5-point Likert scale [42] and normalized into [0,1] interval.

Since this work was published after the rise of Transformer-based architectures, participants mainly relied on solutions based on neural networks [43, 44, 45, 46, 47], but also utilized Ensemble models that proved their efficiency in previous shared tasks [48, 46].

It is also worth mentioning that even though other languages are not so well represented in CWI- or LCP-related works, there are several important works.

In [49] authors presented a dataset of synonyms from the French language ranked with regard to the complexity perceived by an annotator.

For the Spanish language, the authors presented a shared-task ALexS-2020 [50]. Participants were asked to predict a binary complexity score for given data in an unsupervised or semi-supervised way due to the lack of labeled data.

For the Chinese language, authors created and enhanced a system for complexity estimation. In [51], they asked annotators to rank 600 Chinese words with regards to their complexity using a 5-point Likert scale and then translated the obtained continuous score into binary ones.

In [52] authors presented a corpus of Japanese words from the Japanese Education Vocabulary List annotated and divided into 3 groups in terms of complexity: easy, medium, and difficult. In addition, they also estimated the complexity of words from Japanese Wikipedia, the Tsukuba Web Corpus, and the Corpus of Contemporary Written Japanese [53]. Both Chinese and Japanese

Table 2. Pearson correlation coefficient of features, computed over full dataset

All features, except those denoted with RWN, were extracted from Russian Wiktionary

synonyms	1,00															
antonyms	0,30	1,00														
hyperonyms	0,38	0,23	1,00													
hyponyms	0,09	0,16	0,24	1,00												
definitions	0,41	0,15	0,34	0,10	1,00											
grammar tags	0,10	0,03	0,03	0,07	0,15	1,00										
diminutive tags	0,15	0,06	0,21	0,05	0,29	0,02	1,00									
nouns	0,34	0,15	0,33	0,14	0,32	0,04	0,18	1,00								
adjectives	0,30	0,14	0,29	0,11	0,29	0,05	0,22	0,85	1,00							
adverbs	0,18	0,22	0,06	0,01	0,18	0,12	0,12	0,36	0,34	1,00						
idioms	0,15	0,08	0,10	0,06	0,38	0,06	0,25	0,14	0,14	0,12	1,00					
postfixes	0,09	-0,03	0,12	0,00	0,18	0,01	0,17	0,07	0,07	-0,04	0,06	1,00				
hyperonyms (RWN)	0,11	0,10	0,12	0,06	0,04	0,03	0,12	0,08	0,06	0,09	0,02	-0,03	1,00			
hyponyms (RWN)	0,07	0,06	0,01	0,09	0,03	0,02	-0,02	0,06	0,01	0,11	0,05	0,04	0,02	1,00		
PoS-synonyms (RWN)	0,00	0,06	0,00	0,03	-0,04	0,03	-0,04	-0,02	-0,03	0,02	-0,01	-0,11	0,24	0,08	1,00	
complexity	-0,12	-0,15	-0,17	-0,11	-0,24	-0,10	-0,24	-0,16	-0,16	-0,19	-0,21	-0,13	-0,14	-0,11	-0,14	1,00

systems employed an SVM as a baseline for complexity prediction.

For the Swedish language, the author presented two annotated datasets: the first one consisted of words from the Common European Reference Framework (CERF) with labels corresponding to language proficiency levels, e.g., A1 or C2; the second dataset was based on words, extracted from Swedish textbooks and dictionaries and annotated following the same methodology, as the one that was used for the first corpus [54].

As a set of baselines authors utilized a variety of machine learning models, e.g., SVM, Random Forest, Logistic Regression, and Naive Bayes, trained on the set of hand-crafted morphological, syntactic, conceptual, and contextual features. Finally, in [55] authors presented a corpus for CWI in Hindi.

They extracted texts from different novels and short stories and annotated them with the help of annotators with various proficiency levels in Hindi in order to correctly estimate differences in perception. As a baseline model, they employed tree-ensemble classifiers.

In the majority of aforementioned works, authors and participants used traditional machine learning models and, sometimes, recurrent neural networks or NN-based models for feature extraction.

After the significant rise of successful applications of Transformer-based [56] architectures for different tasks, it was expected to see novel methods in CWI and LCP areas.

In LCP-2021, participants of shared-task utilized various models. JUST-BLUE [44] system took inspiration from Ensemble methods and combined BERT [57] and RoBERTa [58] models by giving as an input target words for the first pair of BERT and RoBERTa and contexts with target words for the second pair. Predictions of all 4 models were weighted and averaged.

In DeepBlueAI [43] authors employed the ensemble of different models, such as BERT, RoBERTa, ALBERT [59], and ERNIE [60], and model stacking with 5 steps.

Firstly, they obtained predictions from all base models, then created and fitter a wide set of hyperparameters for models, at the third step they applied 7-fold cross-validation in order to avoid overfitting or correction bias, and then utilized various supplementing techniques, e.g., pseudo- labeling.

As a final estimator, the authors trained a simple linear regressor. Both JUST-BLUE and DeepBlueAI used complicated ensemble and model stacking schemes. In opposition to those works, authors of RG_PA [45] utilized only a single RoBERTa model, showing that a properly trained model is able to perceive word complexity at a relatively high level.

3 Data Collection

In order to create a dataset of semantic, syntactic, and morphological features we parsed the Russian Wiktionary dump dated 01 November 2021 and selected words with corresponding information that matched with words from previously published corpora on lexical complexity for the Russian language [1].

In total, 914 out of 931 distinct words with corresponding articles were present in RuWiktionary. We did not perform any additional filtering of polysemy, since we assumed them to be rare enough and unlikely to affect the results. We chose several features that could reflect word complexity according to the partition of articles from Wiktionary. The following features were selected:

- the number of different meanings;
- the number of word synonyms, antonyms, hypernyms, and hyponyms;
- the number of idioms with target word; morphological features - number of prefixes, suffixes, and declension endings;
- the number of words in different categories from word family - nouns, adjectives, verbs, adverbs;
- the number of Wiktionary tags inside word definition, grouped into 5 relatively large groups.

Additionally, we enriched the dataset with features from Russian WordNet [61]. By its nature, RuWordNet (RWN) contains fewer words than Russian Wiktionary, but provides a precise network of connections between them and, therefore, is able to provide us with more accurate data.

For our corpus, we selected only 3 features: the number of hypernyms, hyponyms, and Part-of-Speech synonyms, and excluded all other parts of speech and multiword expressions, except for single nouns and proper nouns.

In order to reduce the amount of potential noise in data, we excluded several features with 100 or fewer entries, e.g., following features with the number of words from word family were removed: proper nouns, predicates, toponyms, ethnonyms, numerals, surnames, participles, etc.

Since the original dataset contained triplets context - target word - complexity score", additional preprocessing was required to be applied. First of all, we lemmatized each target word from triplets and grouped them by resulting lemmas.

Secondly, we averaged complexity scores for each lemma, and, finally, we excluded surrounding contexts since after averaging it would not be possible to match specific context to the corresponding complexity score.

A resulting dataset contained pairs "target lemma - average complexity score". Table 1 illustrates the preprocessing scheme with examples of different triplets before averaging and a resulting pair.

It is also important to consider that different features have different occurrence rates, which leads to gaps in data. In our work, we have chosen to handle missing values by replacing them with zeroes.

Considering this, we have to notice that using zeroes to fill in missing data might not be an optimal solution - zero could either simply represent a missing value or an absence of some particular feature, i.e. zero number of synonyms could tell us that a word is either a very complex one or a very basic and common one. A proper

Table 3. Pearson correlation coefficient of features, computed over dataset with complexity below 0.255

hyperonyms	1,00							
definitions	0,26	1,00						
style tags	0,18	0,51	1,00					
diminutives	0,14	0,25	0,26	1,00				
adverbs	0,04	0,17	0,17	0,13	1,00			
idioms	0,06	0,38	0,25	0,22	0,09	1,00		
declension endings	0,10	0,17	0,09	0,18	-0,09	0,02	1,00	
complexity	-0,15	-0,22	-0,13	-0,22	-0,11	-0,15	-0,11	1,00

study on the processing of missing data and exploitation of rare features is a part of future work.

To eliminate the influence of multicollinearity and select the most significant features we plotted a correlation heatmap for features and target word complexity. Additionally, we split the dataset by median complexity into easy and difficult-to-comprehend samples and plotted the same correlation heatmaps.

Tables 2-4 contain cross-correlation values as well as the correlation of features with word complexity. For the purpose of clearer representation, we excluded features, for which correlation value with word complexity lies within a range of (-0.1; 0.1).

Since each correlation matrix is symmetrical, we demonstrate them as lower triangular matrices.

4 Experiments

To validate on how well complexity can be estimated with collected features we conducted a set of experiments with linear regression as a baseline.

For our experiments, we selected the following setups: trained on all 21 features; selected for training only 5 most important ones that demonstrated the highest absolute correlation with target score; used all features with added handcrafted (HC) features, such as word length, number of syllables and word frequency; and completed all features with additional 300-dimensional fasttext features.

We used 10-fold cross-validation and estimated model performance with Mean Average Error (MAE) and Pearson's correlation coefficient (PCC).

Suggesting that a non-linear dependency between word complexity and word features might be induced by significant differences between groups of easy and hard words, we split the dataset into two approximately equal parts by median complexity (0.225) and conducted the same experiments with the aforementioned setups.

Table 5 contains aggregated validation metrics for all experiments with linear regression baseline rounded up to the third sign, with results of experiments on RuBERT setup for comparison.

As can be seen from Table 5, the best results for linear regression were achieved with a combination of features extracted from Russian Wiktionary articles and supplemented with HC features. We suggest that this observation is

Table 4. Pearson correlation coefficient of features, computed over dataset with complexity above 0.255

antonyms	1,00												
definitions	0,05	1,00											
style tags	0,00	0,38	1,00										
grammar tags	0,01	0,04	0,02	1,00									
nouns	0,10	0,33	0,14	0,06	1,00								
adjectives	0,11	0,33	0,28	0,08	0,48	1,00							
verbs	0,11	0,23	0,04	0,08	0,46	0,36	1,00						
adverbs	0,18	0,12	0,03	0,00	0,32	0,36	0,24	1,00					
suffixes	0,06	0,03	0,01	0,06	0,10	0,05	0,01	0,02	1,00				
hyperonyms (RWN)	0,20	0,05	0,02	0,05	0,10	0,00	0,04	0,08	0,05	1,00			
hyponyms (RWN)	0,04	0,05	-0,02	-0,01	0,09	-0,01	0,06	0,13	0,00	-0,04	1,00		
synonyms (RWN)	0,10	-0,05	-0,10	0,00	0,08	0,05	0,12	0,08	-0,09	0,23	0,10	1,00	
complexity	-0,15	-0,11	0,21	-0,11	-0,11	-0,14	-0,12	-0,13	-0,17	-0,12	-0,10	-0,19	1,00

mostly based on the strong correlation between word complexity and word frequency and is supported by additional information from morphological and syntactic features.

It is also important to notice that a combination of fasttext features and Wiktionary features demonstrated good results in terms of PCC.

We argue that utilization of any set of implicit semantic features could significantly benefit the complexity estimation quality if it would have been supported by a strong model's induced bias, which is supported by many recent works [43-45].

In order to validate this assumption for Russian corpora we conducted a set of experiments on word complexity estimation with RuBERT as a baseline. We selected the following setups for our experiments: utilized only target words as input data; supplemented target words with surrounding contexts; reduced the size of the train set of samples with contexts to the size of the train set of samples without contexts in order to estimate the significance of context presence; excluded samples from the train set, in which lemma of target word matched lemma of any target word from the test set. We also applied the same splitting strategy by median complexity value and obtained average metric values through 10-fold cross-validation. Figures 1 and 2 represent the results of the experiments.

5 Discussions

The results of conducted experiments with linear regression have supported conclusions, previously

Table 5. Results of word complexity prediction with linear regression in different setups

	Full dataset		**Part of the dataset with complexity score below 0.225**		**Part of the dataset with complexity score above 0.225**	
	MAE	PCC	MAE	PCC	MAE	PCC
Linear regression + full set of Wiktionary features	0.085	0.302	0.036	0.054	0.069	0.365
Linear regression + reduced set of Wiktionary features	0.087	0.138	0.035	0.126	0.074	0.045
Linear regression + full set of Wiktionary features and HC features	0.082	0.341	0.034	0.168	**0.068**	**0.412**
Linear regression + full set of Wiktionary features and fasttext	0.09	0.37	0.068	0.077	0.143	0.214
RuBERT + tokenized target words without context	**0.068**	**0.643**	**0.034**	**0.261**	0.072	0.328

demonstrated in works for different languages. To our knowledge, our work is the first to present a study on the importance of various word features for their complexity estimation and the first to conduct research on word-level complexity prediction in the presence of surrounding context with a modern Transformer-based model for the Russian language.

As it comes from the results of the experiments with the linear regression model, even a complex set of morphological, syntactic, and semantic features is able to reflect word complexity only up to some degree.

A more sophisticated model trained on the well-designed dataset might be able to demonstrate a higher quality, which comes at the cost of a more complicated feature engineering process and is easily matched by providing additional simple features to the model, such as word frequency and length, that highly correlate with estimated Word complexity.

Both the results of experiments with the linear regression model and RuBERT have shown the significance of implicit semantic features that are able to reflect connections between words.

The results of the experiments on the linear regression model, trained on a combination of fasttext and RuWiktionary features, and RuBERT, trained solely on target words, demonstrate the best performance on a full dataset with a great drop in experiments on “easy” and “difficult” parts of the dataset.

We argue that this inconsistency might be explained by the presence of additional semantic relations between easy and difficult words within their groups, which allows us to clearly distinguish between groups themselves but is not enough to discriminate words with similar complexity.

Finally, our experiments with RuBERT have proved the expected conclusions. First of all, the importance of surrounding contexts was evidently demonstrated by a comparison of the results of experiments with and without them.

The presence of relevant contexts helps in correct estimation not only for the full dataset but for its separated parts as well.

Secondly, experiments with the train set that was either randomly reduced to match the size of the dataset with target words only or did not include any samples with target words, which appear in the test set, have clearly displayed an influence of corpus size and its degree of inner diversity on the performance of word complexity estimation.

It is also important to notice that our assumptions regarding a clear inter-group separation for easy and difficult words are not that obvious in these cases due to the additional influence of dataset size.

The main limitation of our work is the choice of a single domain for experiments. Our assumptions are yet to be proven on the inter- and intra-domain setups and we are aiming to overcome this in future work.

6 Conclusion

In this paper, we presented an extension of an existing dataset for predicting lexical complexity in the Russian language formed by collecting word features from Russian Wiktionary articles. The dataset consists of 914 distinct words with each word described by 21 morphological, syntactic and semantic features.

We performed an analysis of baseline models performance, such as linear regression model and RuBERT in various setups. We were able to prove the great significance of implicit semantic features for correct word complexity estimation.

Additionally, we observed a consistent pattern in MAE and PCC metrics for experiments on the full dataset and its split parts. We argue that additional semantic information from surrounding contexts is vital for the correct estimation of complexity within groups of words with similar complexity scores. Our work is dedicated to the investigation of LCP phenomenon solely on the Bible domain, and we aim to conduct a more rigorous analysis of LCP for multi-domain setups.

Acknowledgments

This paper has supported by the Russian Science Foundation, grant #22-21-00334[1].

References

1. **Abramov, A. V., Ivanov, V. V. (2022).** Collection and evaluation of lexical complexity data for Russian language using crowdsourcing. Russian Journal of Linguistics, Vol. 26, No. 2, pp. 409–425, DOI: 10.22363/2687-0088-30118.
2. **Kuratov, Y., Arkhipov, M. (2019).** Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv:1905.07213. DOI: 10.48550/arXiv. 1905.07213.
3. **Dale, D. (1948).** The Dale-Chall formula for predicting readability. Educational Research Bulletin, Vol. 27, pp. 11–20.
4. **Chall, J. S., Dale, E. (1995).** Readability revisited: The new Dale-Chall readability formula. Brookli Books
5. **Devlin, S. (1998).** The use of a psycholinguistic database in the simplification of text for aphasic readers. Linguistic databases.
6. **Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J. (1998).** Practical simplification of English newspaper text to assist aphasic readers. Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology Association for the Advancement of Artificial Intelligence, pp. 7–10.
7. **Specia, L., Jauhar, S. K., Mihalcea, R. (2012).** SemEval-2012 task 1: English lexical simplification. Proceedings of First Joint Conference on Lexical and Computational Semantics (SEM), pp. 347–355.
8. **Amoia, M., Romanelli, M. (2012).** Sb: mmsystem-using decompositional semantics for lexical simplification. Proceedings of First Joint Conference on Lexical and Computational Semantics (SEM), pp. 482–486
9. **Ligozat, A. L., Grouin, C., Garcia-Fernandez, A., Bernhard, D. (2012).** Annlor: A naïve notation-system for lexical outputs ranking. Proceedings of

[1] https://rscf.ru/project/22-21-00334/

First Joint Conference on Lexical and Computational Semantics (SEM), pp. 487–492.

10. **Sinha, R. (2012).** Unt-simprank: Systems for lexical simplification ranking. Proceedings of First Joint Conference on Lexical and Computational Semantics (SEM), pp. 493–496.
11. **Jauhar, S. K., Specia, L. (2012).** Uow-shef: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. Porceedings of First Joint Conference on Lexical and Computational Semantics (SEM), pp. 477–481.
12. **Gooding, S., Kochmar, E., Blackwell, A., Sarkar, A. (2019).** Comparative judgments are more consistent than binary classification for labelling word complexity. Association for Computational Linguistics.
13. **Paetzold, G., Specia, L. (2017).** Lexical simplification with neural ranking. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, pp. 34–40.
14. **Xu, W., Callison-Burch, C., Napoles, C. (2015).** Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics, 3, 283-297.
15. **Paetzold, G., Specia, L. (2016).** Semeval 2016 task 11: Complex word identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval´16), pp. 560–569.
16. **Shardlow, M. (2013).** A comparison of techniques to automatically identify complex words. 51st annual meeting of the association for computational linguistics proceedings of the student research workshop, pp. 103–109.
17. **Shardlow, M. (2013).** The cw corpus: A new resource for evaluating the identification of complex words. Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp. 69–77.
18. **Choubey, P. K., Pateria, S. (2016).** Garuda & Bhasha at SemEval-2016 task 11: Complex word identification using aggregated learning models. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 1006–1010.
19. **Zampieri, M., Tan, L., van Genabith, J. (2016, June).** Macsaar at SemEval-2016 task 11: Zipfian and character features for complex word identification. Proceedings of the 10th International Workshop on Semantic Evaluation SemEval´16, pp. 1001–1005.
20. **Kuru, O. (2016).** Ai-ku at semeval-2016 task 11: Word embeddings and substring features for complex word identification. Proceedings of the 10th International Workshop on Semantic Evaluation SemEval´16, pp. 1042–1046.
21. **Quijada, M., Medero, J. (2016).** Hmc at SemEval-2016 task 11: Identifying complex words using depth-limited decision trees. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 1034–1037.
22. **Malmasi, S., Dras, M., Zampieri, M. (2016).** Ltg at SemEval-2016 task 11: Complex word identification with classifier ensembles. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 996–1000.
23. **Malmasi, S., Zampieri, M. (2016).** Maza at SemEval-2016 task 11: Detecting lexical complexity using a decision stump meta-classifier. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 991–995.
24. **Brooke, J., Uitdenbogerd, A. L., Baldwin, T. (2016).** Melbourne at SemEval 2016 task 11: Classifying type-level word complexity using random forests with corpus and word list features. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 975–981.
25. **Nat, G. (2016).** Sensible at SemEval-2016 task 11: Neural nonsense mangled in ensemble mess. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 963–968.
26. **Paetzold, G., Specia, L. (2016).** Sv000gg at Semeval-2016 task 11: Heavy gauge complex word identification with system voting. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 969–974.
27. **Ronzano, F., Anke, L. E., Saggion, H. (2016).** Taln at Semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16, pp. 1011–1016.
28. **Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Zampieri, M. (2018).** A report on the complex word identification shared task 2018. arXiv preprint arXiv:1804.09132, DOI: 10.48550/arXiv.1804.09132.
29. **Butnaru, A. M., Ionescu, R. T. (2018).** UnibucKernel: A kernel-based learning method for complex word identification. arXiv preprint arXiv:1803.07602. DOI: 10.48550/arXiv.1803.07602.
30. **AbuRa'ed, A. G. T., Saggion, H. (2018).** LaSTUS/TALN at complex word identification (CWI) 2018 shared task. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 159–65.

31. **Alfter, D., Pilán, I. (2018).** SB@ GU at the complex word identification 2018 shared task. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 315– 321.
32. **Hartmann, N., Dos Santos, L. B. (2018).** NILC at CWI 2018: Exploring feature engineering and feature learning. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications pp. 335–340.
33. **Kajiwara, T., Komachi, M. (2018).** Complex word identification based on frequency in a learner corpus. Proceedings of the thirteenth workshop on innovative use of NLP for Building Educational Applications, pp. 195–199.
34. **Bingel, J., Schluter, N., Alonso, H. M. (2016).** CoastalCPH at SemEval-2016 Task 11: The importance of designing your neural networks right. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval´16´, pp. 1028–1033.
35. **Aroyehun, S. T., Angel, J., Alvarez, D. A. P., Gelbukh, A. (2018).** Complex word identification: convolutional neural network vs. feature engineering. Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp. 322–327, DOI: 10.18 653/v1/W18-0538.
36. **Maddela, M., Xu, W. (2018).** A word-complexity lexicon and a neural readability ranking model for lexical simplification. DOI: 10.48550/arXiv.1810. 05754.
37. **Shardlow, M., Evans, R., Paetzold, G. H., Zampieri, M. (2021).** Semeval-2021 task 1: Lexical complexity prediction. DOI: 10.48550/arXiv.2106. 00473.
38. **Christodouloupoulos, C., Steedman, M. (2015).** A massively parallel corpus: the bible in 100 languages. Language resources and evaluation, Vol. 49, No. 2, pp. 375–395, DOI: 10.1007/s10579-014-9287-y.
39. **Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Hunter, L. E. (2012).** Concept annotation in the CRAFT corpus.BMC bioinformatics, Vol.13, 1, pp. 1–20, DOI: 10.1186/14 71-2105-13-161.
40. **Koehn, P. (2005).** Europarl: A parallel corpus for statistical machine translation. Proceedings of machine translation summit x, pp. 79–86.
41. **Likert, R. (1932).** A technique for the measurement of attitudes. Archives of psychology.
42. **Pan, C., Song, B., Wang, S., & Luo, Z. (2021).** DeepBlueAI at SemEval-2021 Task 1: Lexical complexity prediction with a deep ensemble approach. Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval´21, pp. 578–584, DOI: 10.18653/v1/2021.semeval-1.72.
43. **Yaseen, T. B., Ismail, Q., Al-Omari, S., Al-Sobh, E., Abdullah, M. (2021).** JUST-BLUE at SemEval-2021 Task 1: Predicting lexical complexity using BERT and RoBERTa Pre-Trained language models. Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval´21, pp. 661–666, DOI: 10.18653/v1/2021.semeval-1.85.
44. **Rao, G., Li, M., Hou, X., Jiang, L., Mo, Y., Shen, J. (2021).** RG PA at SemEval-2021 Task 1: A contextual attention-based model with RoBERTa for lexical complexity prediction. Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval´21, pp. 623–626, DOI: 10.18653/v1/2021.semeval-1.79.
45. **Rotaru, A. (2021).** ANDI at SemEval-2021 Task 1: Predicting complexity in context using distributional models, behavioural norms, and lexical resources. Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval´21, pp. 655–660, DOI: 10.18653/v1/2021.semeval-1.84.
46. **Taya, Y., Pereira, L. K., Cheng, F., Kobayashi, I. (2021).** OCHADAI-KYOTO at SemEval-2021 Task 1: Enhancing model generalization and robustness for lexical complexity prediction. arXiv preprint arXiv:2105.05535, DOI: 10.48550/arXiv.2105. 05535.
47. **Mosquera, A. (2021).** Alejandro Mosquera at SemEval-2021 Task 1: Exploring sentence and word features for lexical complexity prediction. Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval´21, pp. 554–559, DOI: 10.18653/v1/2021.semeval-1.68.
48. **Billami, M., François, T., Gala, N. (2018).** ReSyf: A French lexicon with ranked synonyms. 27th International Conference on Computational Linguistics, COLING´18, pp. 2570–2581.
49. **Ortiz-Zambranoa, J. A., Montejo-Ráezb, A. (2020).** Overview of alexs 2020: First workshop on lexical analysis at sepln. Proceedings of the Iberian Languages Evaluation Forum, IberLEF´20.
50. **Lee, J. S., Yeung, C. Y. (2018).** Personalizing lexical simplification. Proceedings of the 27th International Conference on Computational Linguistics, pp. 224–232.
51. **Nishihara, D., Kajiwara, T. (2020).** Word Complexity Estimation for Japanese Lexical Simplification. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 3114–3120.

52. Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Den, Y. (2010). Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10.

53. Smolenska, G. (2018). Complex Word Identification for Swedish.

54. Venugopal, G., Pramod, D., Shekhar, R. (2022). CWID-hi: A dataset for complex word identification in Hindi text. Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 5627–5636.

55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems.

56. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. DOI: 10.48550/arXiv.1810.04805.

57. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. DOI: 10.48550/arXiv.1907.11692.

58. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. DOI: 10.48550/arXiv.1909.11942.

59. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223. DOI: 10.48550/arXiv.1904.09223.

60. Loukachevitch, N., Lashevich, G. (2016). Multiword expressions in Russian thesauri RuThes and RuWordnet. Proceeding of IEEE Artificial Intelligence and Natural Language Conference*, AINL,* pp. 1–6.

61. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. DOI: 10.48550/arXiv.1301.378.

62. Pennington, J., Socher, R., Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP, pp. 1532–1543.

63. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, Vol. 5, 135–146, DOI: 10.1162/tacl_a_00051.

64. He, P., Liu, X., Gao, J., Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*. DOI: 10.48550/arXiv.2006.03654.

65. Clark, K., Luong, M. T., Le, Q. V., Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*. DOI: 10.48550/arXiv.2003.10555.

Article received on 11/10 /2022; accepted on 25/112022.
Corresponding author is Valery D. Solovyev.