Estimating Lexical Complexity in Multi-Domain Settings for the Russian Language

Aleksei Abramov, Vladimir Ivanov and **Valery Solovyev**
Kazan Federal University, Kazan, Russian Federation
Innopolis University, Innopolis, Russian Federation
alvabramov@stud.kpfu.ru
v.ivanov@innopolis.ru
maki.solovyev@mail.ru

Abstract

During the last few years, the tasks of Complex Word Identification (CWI) and Lexical Complexity Prediction (LCP) have received growing attention from researchers who studied the quality of complexity estimation with modern Transformer-based models in different setups and for various languages: English, German, French, Spanish, Chinese, Japanese, Swedish and Russian. The crucial step in creating a large and robust modern language model for any language is a collection of representative data. Supporting the trend of providing the researchers with the ability to study the task of LCP for various languages and aiming to increase the representativity for the Russian language, we extend the number of available domains with two new ones: biomedical and sociopolitical. The data was collected and annotated in accordance with the original methodology, originally described for the CompLex dataset. We conduct a comparative study on data from both domains and perform a cross-domain study. We present the results of experiments with the RuBERT-based model in mono-domain and cross-domain settings for both domains. The results of our work are two novel corpora consisting of 756 distinct words for the biomedical domain with 3050 contexts and 669 distinct words for the sociopolitical domain with 3025 contexts.

1 Introduction

The task of Lexical Complexity Prediction (LCP) has received growing interest from various research groups around the world. Originating from works on text readability assessment and text simplification with the help of manually created formulas and dictionaries, it eventually evolved into the task of evaluating word complexity in the presence of surrounding context with already well-established Transformer-based models, as can be seen from recent works presenting results of experiments conducted with the following models: BERT, RoBERTa, DeBERTa, ELECTRA, ALBERT, ERNIE (Devlin et al., 2019; Zhuang et al., 2019; Lan et al., 2019; Pan et al., 2021; Yaseen et al., 2021; Rao et al., 2021).. An important trend in modern research is the spreading of the research field into previously unexplored setups, such as LCP for Arabic, French, Spanish Chinese, Japanese, Swedish, Hindu, and Russian languages, or estimation of word complexity for multiple domains.

Another essential trend related to word complexity estimation is increasing attention to problems of representativity. Original works on text readability assessment were devoted exclusively to performing estimation for students studying in 4-12 grades of native English-speaking school and college students. More recent works presented improved studies on the task of Complex Word Identification (CWI), by, firstly, including various sources of data employed in collection and annotation processes, such as articles from Wikipedia, Simple Wikipedia, news articles, written by both professional and amateurs, abstracts of scientific papers contemporary literature, etc., and, secondly, utilizing crowdsourcing for data annotation with an additional focus on non-native speakers.

Strongly supporting these trends, we provide a basis for future research of a task of LCP for the Russian Language by presenting two novel datasets in biomedical and sociopolitical domains. Our research on the aforementioned domains follows the original methodology, initially presented for the CompLex dataset (Shardlow et al., 2021), and later exploited in works on word

complexity estimation for the data from the Russian Synodal Bible (Abramov et al., 2022; Abramov et al., 2023). We present an analysis of collected complexity scores comparing statistics for both domains, evaluating statistical significance, and exploring intersected words. Additionally, we conduct experiments with the RuBERT-based (Kuratov and Arkhipov, 2019) model in two setups: for experiments within a single domain we evaluate Mean Average Error and Pearson Correlation Coefficient with cross-validation on 5 folds; for cross-domain experiments, we estimate how well complexity estimation ability transfers from one domain to another by training on data solely from one domain and evaluating on data from another domain. Our results demonstrate an influence of domain on perceived complexity and a possibility of word complexity knowledge transfer between domains. By presenting the new dataset we believe to create a strong basis for future research on LCP for the Russian language, and, hopefully, for multi-language setups.

2 Related works

The early works related to the area of LCP described text simplification or readability assessment methods. For the English language, the Dave-Chall formula, exploited as a tool for text readability assessment, was presented and later revised by (Dale and Chall, 1948) and (Chall and Dale, 1995). Systems for text simplification used the detection of candidates for simplification as a part of their pipelines (Devlin, 1998; Caroll et al., 1998).

Later the tasks of CWI or LCP were introduced as standalone ones. In the Lexical Simplification task at SemEval-2012 (Specia et al., 2012) the CWI task was presented as a ranking task in which the participants were required to build a system for ranking words in terms of complexity. In (Shardlow, 2013), (Shardlow, 2013), and later in CWI-2016 (Paetzold and Specia, 2016) a representation of the CWI task evolved into a prediction of binary score “not complex”/” complex”. In (Shardlow, 2013) authors presented a dataset with pairs of “annotated complex word” - “simple substitution”. Organizers of CWI-2016 focused on the representativity of the data by presenting a dataset of sentences from Wikipedia, annotated by 400 non-native speakers. In CWI-2018 (Yimam et al., 2018) authors conducted a survey on complexity estimation in multi-genre and multilingual settings by presenting datasets for 4 languages: English, German, Spanish, and French. An additional dataset of news articles, written by professionals, amateurs, and Wikipedia, was used to evaluate how well models can estimate word complexity for sources with different initial implicit complexity. Alongside a track of binary complexity estimation, the authors introduced a track with probability estimation of a word being complex. In LCP-2021 (Shardlow et al., 2021) authors eventually addressed the task of word complexity estimation as a prediction of continuous labels. They presented a multi-genre dataset for 3 domains: Bible (Christodouloupoulos and Steedman, 2015), biomedical data (Bada et al., 2012), and sociopolitical data (Koehn, 2005) with annotated single nouns and multi-word expressions (MWE). A continuous score was calculated as an average of the annotator's scores normalized into [0,1] intervals.

Even though the English language is the most represented in CWI and LCP tasks, there are several works devoted to solving the same problem for other languages. In a shared-task ALexS-2020 (Ortiz-Zambrano and Ráez, 2020) authors presented an annotated dataset for the Spanish language and required participants to predict binary complexity scores. In (Ortiz-Zambrano et al., 2022) authors studied how well different Transformer-based models combined with regression algorithms can estimate word complexity. For the Chinese language authors (Lee and Yeung, 2018) presented a dataset with binary complexity scores for 600 words. For the Swedish language, the author presented two datasets with words annotated with language proficiency levels from the Common European Reference Framework (CERF) as labels, ranging from A1 to C2 (Smolenska, 2018). For the Japanese language authors annotated words from the Japanese Education Vocabulary List and split them into 3 categories - easy, medium, and difficult (Maekawa et al., 2010). Another work for the Japanese language was focused on creating an LCP dataset JaLeCoN with single words and MWE taken from 2 different sources: news and government texts, and annotated with a focus on non-native speakers with L1 proficiency in Chinese/Korean (Ide et al., 2023). For the Hindi language, authors collected a vast dataset of words from novels and short stories and asked annotators with different proficiency

levels to annotate them to make the corpus more representative (Venugopal et al., 2022). For the French language authors of (Billiami et al., 2018) presented a dataset of synonyms ranked by perceived complexity. In another work dedicated to both English and French, the authors created a dataset and trained multilingual models to classify words according to CERF levels (Aleksandrova and Pouliot, 2023). For the Russian language, in (Solovyev, 2019) authors studied the readability of textbooks for native and non-native speakers to determine the differences in estimating lexical complexity for both groups. In (Abramov and Ivanov, 2022) and (Abramov et al., 2023) authors created a dataset of annotated words in the domain of the Bible and studied estimation of lexical complexity for Russian texts with regression and Transformer-based methods.

In modern works the authors heavily utilized Transformer-based models for estimation of lexical complexity, e.g. BERT, RoBERTa, ALBERT, ERNIE (Devlin et al., 2019; Zhuang et al., 2019; Lan et al., 2019; Pan et al., 2021; Yaseen et al., 2021; Rao et al., 2021).

3 Data collection

An original methodology of data preparation presented in (Shardlow et al., 2021) consisted of several steps: data collection from open sources, such as the Bible, biomedical texts, and Europarl data; sampling of words that fall into the set of predetermined frequency intervals; data annotation with crowdsourcing, where each word was annotated using a 5-point Likert scale (1-5). Following this methodology we constructed our data collection and annotation pipeline similarly: firstly, we performed an initial preparation of data sources for sampling; secondly, we sampled words and corresponding surrounding contexts for annotation utilizing a set of predefined frequency ranges; thirdly, we employed the crowdsourcing platform Yandex.Toloka for data annotation, and, finally, we collected annotated data and created a full corpus.

To follow the original methodology as closely as possible, we selected the following parallel corpora as sources of data: firstly, we chose a part of Medline corpus as a source of biomedical data, namely, a training set of WMT 2020 with parallel English-Russian pairs of abstracts from scientific papers (Bawden et al., 2020); and, secondly, we utilized a part of United Nations Parallel Corpus as

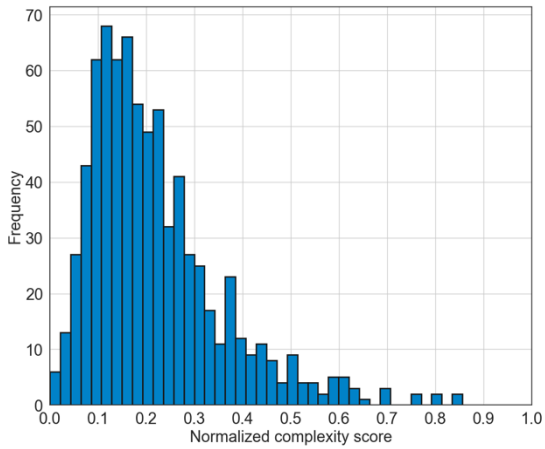

Figure 1: Complexity scores distribution of biomedical data.

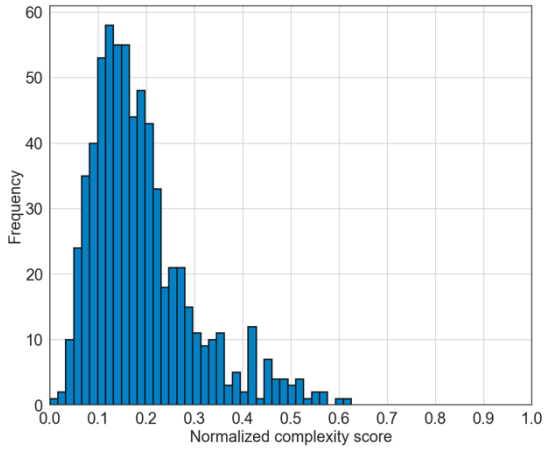

Figure 2: Complexity scores distribution of sociopolitical data.

a source of sociopolitical data, in particular, parallel English-Russian pairs of official records recorded in 2014 (Ziemski et al., 2016).

We have developed several simple heuristics to exclude too simple words or too short surrounding contexts from being sampled:

- As the United Nations Parallel Corpus is composed of official records, it contains a significant number of either very short sentences (e.g. headings) or very long ones (e.g. lists of countries or very detailed explanations). Therefore, we set a minimum length of a sampled sentence to 15 words, and a maximum length of 30 words to exclude these sentences and provide annotators with reasonably sized contexts;

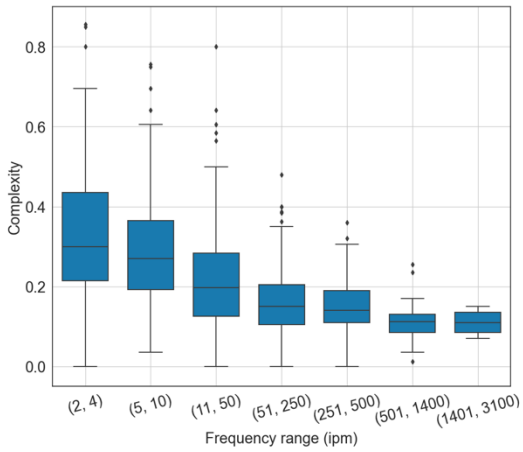

Figure 3: Statistics of lexical complexity scores of words grouped by their frequency for biomedical data.

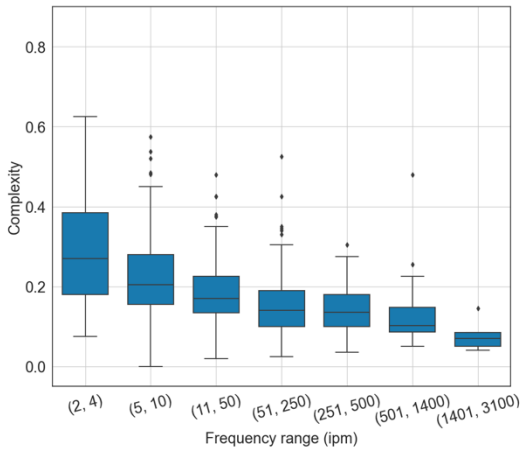

Figure 4: Statistics of lexical complexity scores of words grouped by their frequency for sociopolitical data.

- For both corpora, we set a minimum length of a sampled word to 4 letters to exclude too short and simple samples;

- To ensure the diversity among sampled words we constrain a number of possible distinct contexts for each sampled lemma by 5(max).

We exploited the same word frequency ranges (ipm, instances per million) to sample words with different expected lexical complexity: (2-4), (5-10), (11-50), (51-250), (251-500), (501-1400), (1401-3100). As there are only a few words that fall into the minimum and maximum frequency ranges, we sampled as many unique lemmas as possible, and for the rest frequency intervals, we sampled a roughly equal number of unique lemmas. Finally, in our work, we did not include any other parts of speech and multi-word expressions (MWE), except for nouns only.

For word sampling we utilized the Frequency dictionary constructed from the Google Books NGram data (Solovyev et al., 2019). As both source corpora were created with the help of modern texts, we exploited a smaller version constructed from the data collected in the period between 1992 and 2019, instead of a full version with the data from the period between 1920 and 2019.

During the annotation process, each annotator was presented with at least 10 different contexts, where each context contained only a single highlighted word.

The annotators were provided with the following description of complexity levels:

- Very Easy: the meaning of the highlighted word is clear;

- Easy: the meaning of the highlighted word is obvious and the context supplements it;

- Moderate: the meaning of the highlighted word is familiar, but it becomes clear only after taking the surrounding context into account;

- Difficult: the meaning of the highlighted word is not evident, but might be inferred after considering the context;

- Very Difficult: the meaning of the highlighted word is unclear or the word itself is unfamiliar.

To ensure diversity and representativity among collected complexity scores during the annotation process in Yandex.Toloka we created a set of rules for annotators to be used in the annotation of both corpora:

- If an annotator has earned more than 0.5$ in the last 24 hours, he/she would be suspended in the annotation pool for 1 day;

- If an annotator has skipped more than 2 tasks in a row, he/she would be banned from the project for 3 days;

- If an annotator has spent less than 15 seconds on the annotation of 2 out of 5 last tasks, he/she would be banned from the project for 7 days;

- If the answers of an annotator have not matched with the answers of at least 5 out of 10 annotators in more than 5 out of 10 last tasks, he/she would be suspended in the annotation pool for 1 day.

As our target auditory, we selected annotators from Russia, Ukraine, Belarus, and Kazakhstan, who were ranked by Yandex.Toloka to be Top 10%. We set the price of 0.1$ per task suite with 10 tasks. Our resulting biomedical corpus consisted of 756 distinct lemmas with 3050 surrounding contexts, and the sociopolitical corpus consisted of 669 distinct lemmas with 3025 surrounding contexts.

4 Data analysis

We conducted an analysis of collected complexity scores by computing scores statistics, observing their distribution, and comparing complexity scores for data from the biomedical domain with scores for data from the sociopolitical domain.

Firstly, in order to evaluate whether one domain could be considered to be more complex than another, we calculated and compared mean values and standard deviation for complexity scores: for the biomedical domain mean value is 0.218, and the standard deviation is 0.139; for sociopolitical domain mean value is 0.19 and standard deviation is 0.107. The difference between domains can be clearly seen from the complexity scores distributions in Figure 1 and Figure 2. The distribution for the biomedical domain has a heavier right tail with several complexity scores exceeding 0.8; the distribution of the scores for the sociopolitical domain is more concentrated around its peak.

Secondly, we evaluated how complexity scores are distributed among selected frequency intervals. Figure 3 and Figure 4 highlight the same pattern for both domains. Mean complexity and standard deviation simultaneously become smaller with the growth of word frequency. This observation could be explained in the following way - the degree of familiarity with simpler words is higher than with rare ones, therefore, annotators tend to agree on the same score more often. To validate our assumptions, we measured a degree of agreement

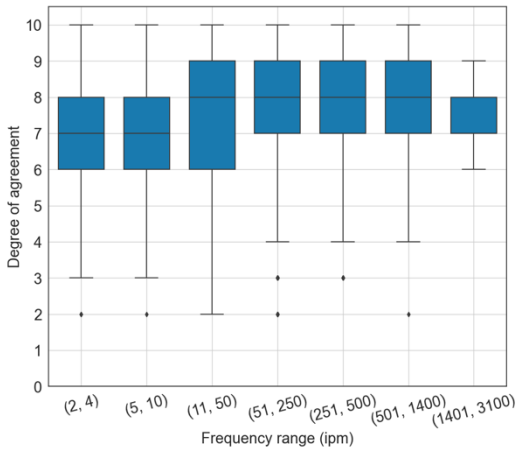

Figure 5: Degrees of agreement for annotated data from the biomedical domain.

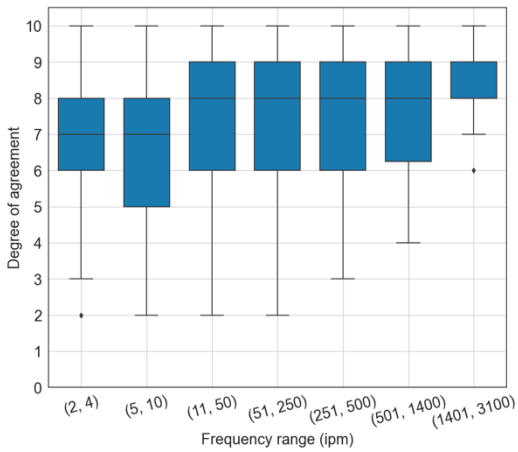

Figure 6: Degrees of agreement for annotated data from the sociopolitical domain.

between annotators following (Iavarone et al., 2021), where the degree of agreement is defined as the number of annotators who gave a complexity score within the range of standard deviation around the mean value. Figure 5 and Figure 6 demonstrate a slightly lower level of agreement for rare words than for the more frequent ones with an inverted dependency between ranges of levels of agreement and frequency ranges. It is also important to note that for all frequency ranges there is a high level of agreement among annotators for both domains, showing a typical degree of agreement to be in the range of 7-9.

Thirdly, we evaluated how contexts from different domains influence perceived complexity. We collected 198 words and their corresponding contexts that appear in both domains and estimated how well their complexity scores correlate. A higher degree of correlation would mean a lower

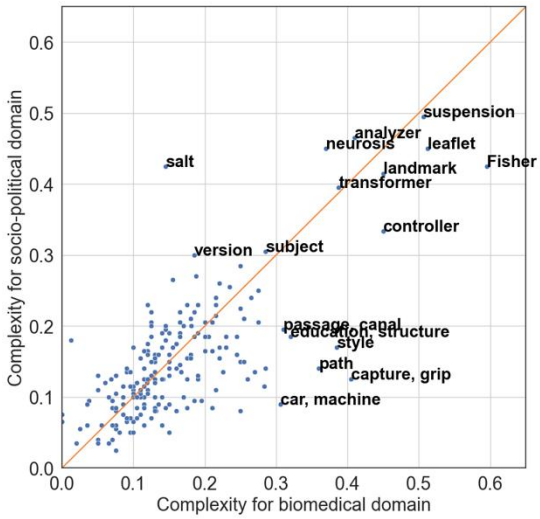

Figure 7: Degrees of agreement for annotated data from the sociopolitical domain. Annotated words were translated from Russian to English.

influence of the source domain and the opposite. First of all, we computed Pearson Correlation and Spearman Correlation Coefficients. They were equal to 0.72 and 0.66 respectively with a p-value significantly lower than 10^{-6} for both metrics. As both Correlation Coefficients are significantly high, but still far below 1.0, it supports previous results proving that a major part of word complexity is conditioned on the word itself, yet context still plays an important part.

Figure 7 illustrates our findings - as simple words have, on average, the same estimated complexity, the complexity of harder words tends to deviate from the linear trend. It can be noticed that words with contents from biomedical data are usually perceived as more complex by annotators. To make it clear, we denoted words that have estimated complexity ≥ 0.3 for both domains.

To validate our assumptions, we performed two paired tests of (non-)equivalence for two dependent paired samples with alternative hypothesis $low \leq mean \leq upper$, where *low* and *upper* were equal 0 and 0.1 respectively in our settings and *mean* denotes an average difference between complexities of given data of common words from biomedical and sociopolitical domains. For all 198 common words, the resulting p-value for the lower boundary was 0.018 and for the upper boundary, the resulting p-value was significantly below 10^{-6}, which resulted in an overall p-value of 0.018. An additional paired test was performed for words with perceived complexity ≥ 0.2. For these 100 words, p-values for the both boundaries were lower than 10^{-6}. These results demonstrate that even though we are unable to observe statistically significant differences in complexity for all samples, it begins to appear as words become more complex and the significance of the surrounding context rises.

Metric	Baseline	Optimal Hyperparameters
MAE	0.074	0.073
PCC	0.759	0.775

Table 1: The results of experiments for biomedical domain

Metric	Baseline	Optimal Hyperparameters
MAE	0.067	0.065
PCC	0.677	0.714

Table 2: The results of experiments for sociopolitical domain

5 Experiments

For our baseline experiments we employed a pre-trained RuBERT (Kuratov and Arkhipov, 2019) model and trained with the following setting for all experiments: batch size of 128, weight decay of 0.01, the learning rate was 10^{-5},the optimizer was AdamW (Loshchilov and Hutter, 2017), and the number of fine-tuning epochs was 20. Freezing these parameters allowed us to exclude their influence on the final metrics, making them dependent completely on the quality of the data. Following previous works on LCP, we selected Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) as metrics.

Additionally, we performed a hyperparameter search with the help of Optuna (Akiba et al., 2019) to investigate whether there is a potential for improvement of our model's performance. We chose 4 hyperparameters to search for learning rate, batch size, number of warmup steps, and weight decay. The total number of trials was 50; the minimized objective for trial evaluation was the sum of negative PCC and positive MAE. The exact range of values for each parameter was: [10^{-6}; 0.01] for learning rate; [16, 32, 64, 128] for batch size, [1, 5, 10, 15, 20] for number of warmup steps and [10^{-5}; 0.1] for weight decay.

5.1 Experiments within a single domain

For our experiments within a single domain, we performed cross-validation on 5 folds for each

Metric	Baseline	Optimal Hyperparameters
MAE	0.076	0.070
PCC	0.592	0.630

Table 3: The results of multi-domain experiments with after training on biomedical data.

Metric	Baseline	Optimal Hyperparameters
MAE	0.083	0.080
PCC	0.662	0.701

Table 4: The results of multi-domain experiments with after training on sociopolitical data.

domain. The resulting metrics were computed as an average of metrics for each fold.

Hyperparameter search within a single domain was performed in the following way. We selected the first 256 samples from the dataset as a small training dataset; the validation dataset for the search consisted of 256 random samples, chosen among those, that were not included into the training part. After the searching process was completed, we applied found parameters to a freshly initialized model for each cross-validation fold. We were able to obtain the following optimal hyperparameters for biomedical domain: the learning rate was 3.53 * 10^{-5}, batch size of 16, the number of warmup steps was 20, weight decay of 1.13 * 10^{-3}; for sociopolitical domain: the learning rate was1.99 * 10^{-5}, batch size of 32, the number of warmup steps was 1, weight decay of 6.89 * 10^{-2}.

The results of experiments for both domains showed a substantial improvement after hyperparameter search. Table 1 and Table 2 demonstrate the resulting metrics of experiments for biomedical and sociopolitical domains respectively.

5.2 Experiments in multi-domain settings

In addition to experiments within a single domain, we validated how well a model can transfer its knowledge about word complexity from one domain to another domain. We estimated this by performing training solely on data from one domain and validating data from another domain. We did not perform any filtering of words that appear in both domains since the surrounding contexts for them are unique.

Similarly to the experiments within a single domain, we performed a hyperparameter search for experiments within multi-domain settings. For both small training and validation datasets we

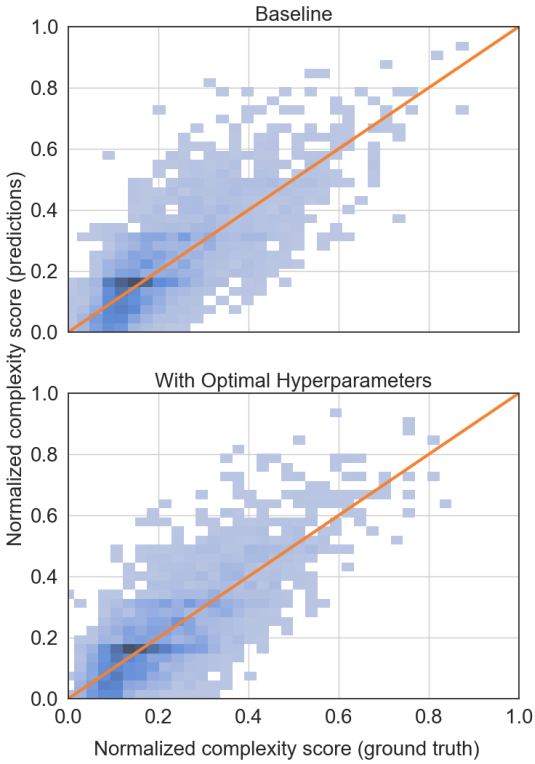

Figure 8: Joint distribution of ground truth and predicted scores in mono-domain experiments on biomedical data.

selected the first 256 samples from each domain. The optimal hyperparameters for the model trained on biomedical domain: the learning rate was 2.99 * 10^{-5}, batch size of 32, the number of warmup steps was 10, weight decay of 7.34 * 10^{-3}; for the model trained on sociopolitical domain: the learning rate was 4.55 * 10^{-5}, batch size of 16, the number of warmup steps was 1, and weight decay of 1.11 * 10^{-5}.

We were able to observe an improvement in metrics value after hyperparameter search for experiment in multi-domain settings as well. Table 3 and Table 4 demonstrate the resulting metrics of multi-domain experiments after training on biomedical and sociopolitical domains respectively.

6 Discussion

To investigate the reasons for metrics improvement in all experiments, we visualized joint distributions of ground truth and predicted complexity scores.

As the datasets for both domains are imbalanced the main contribution is conditioned by an improvement in complexity estimation for relatively simple words. It can be seen in Figure 8

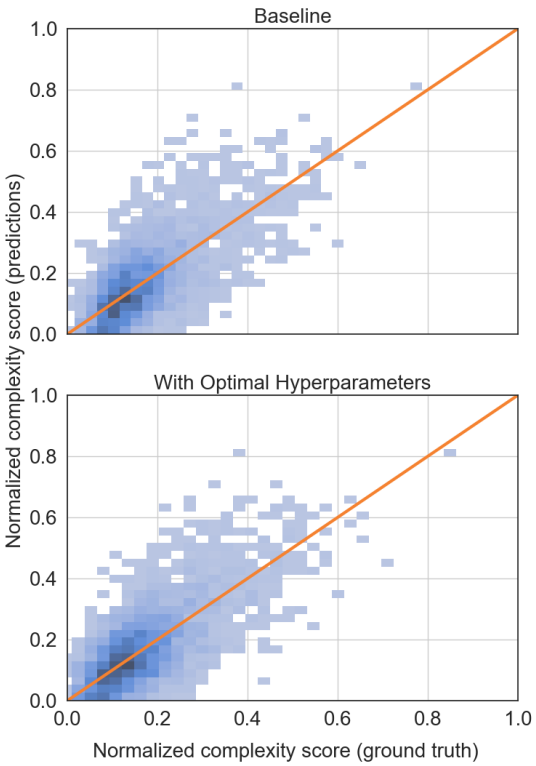

Figure 9: Joint distribution of ground truth and predicted scores in mono-domain experiments on sociopolitical data.

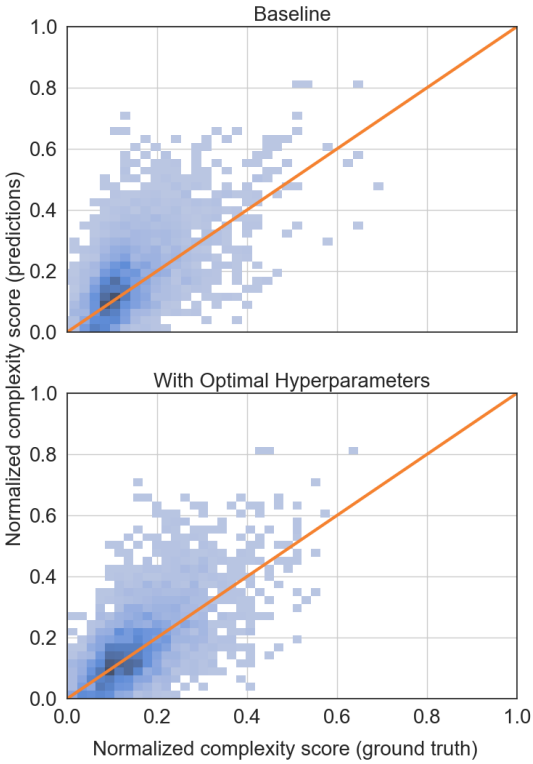

Figure 10: Joint distribution of ground truth and predicted scores in multi-domain experiments after training on biomedical data.

and Figure 9 for mono-domain experiments, and in Figure 10 and Figure 11 for multi-domain experiments, that the joint distribution of ground truth and predicted complexity scores for words with complexity ≤ 0.4 stretches along the main diagonal in experiments after the hyperparameters search. Nevertheless, there are still many examples of the high discrepancy between ground truth and predicted scores even for simple words. Additionally, it should be noted that in all cases joint distributions are slightly shifted towards the upper left part of the image, which might indicate that the model underestimates the complexity of some difficult words. We argue that for the full investigation of a model's capabilities in complexity estimation, the construction of a more balanced dataset is necessary.

7 Conclusion

In this paper, we presented two datasets for predicting lexical complexity for the Russian language with 756 distinct words for the biomedical domain with 3050 contexts and 669 distinct words for the sociopolitical domain with 3025 contexts. It was constructed following the

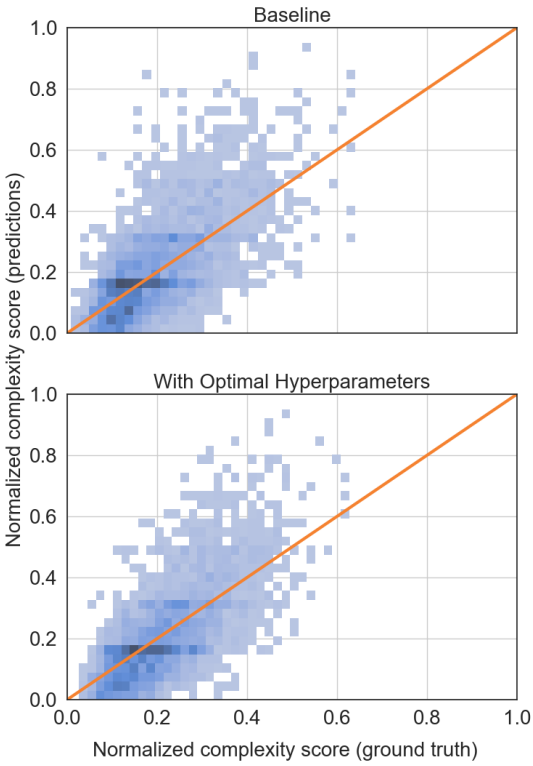

Figure 11: Joint distribution of ground truth and predicted scores in multi-domain experiments after training on sociopolitical data.

methodology of the CompLex dataset and labeled using a crowdsourcing platform Yandex.Toloka. We conducted a study on the distribution of perceived complexity scores for both domains, validated our hypothesis on the dependence between annotators agreement and word familiarity, and performed a comparison of scores for words that appear in both datasets. Additionally, we conducted a series of experiments with a RuBERT-based model in different setups: we performed a 5-fold cross-validation for experiments within a single domain and we performed a multi-domain experiment to study the possibility of word complexity knowledge transfer, which we were able to prove. An additional analysis of experiment results demonstrated a necessity for more balanced datasets. Our work is dedicated to studying the complexity phenomena in only two domains within monolingual settings, and we aim to conduct an additional comparison and analysis in multilingual settings. The presented corpus will be freely accessible for the international research community in Zenodo repository.

Acknowledgments

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

References

Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin*, 27(1):11–20+28.

Jeanne S. Chall and Edgar Dale. 1995. Readability *Revisited: The New Dale-Chall Readability Formula.* Brookline Books.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words.. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and 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*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375-395.

Michael Bada, Miriam R. Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):1-20.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Jenny Alexandra Ortiz-Zambrano and Arturo Montejo Ráez. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), pages 1–6.

John Lee and Chak Yan Yeung. 2018. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Greta Smolenska. 2018. Complex Word Identification for Swedish.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gayatri Venugopal, Dhanya Pramod, and Ravi Shekhar. 2022. CWID-hi: A Dataset for Complex Word Identification in Hindi Text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5627–5636, Marseille, France. European Language Resources Association.

Mokhtar B. Billami, Thomas François, and Núria Gala. 2018. ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Zhenzhong Lan, Mingda Chen, Seth Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Computing Research Repository,* arXiv: 1909.11942. Version 6

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584, Online. Association for Computational Linguistics.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and 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)*, pages 661–666, Online. Association for Computational Linguistics.

Gang Rao, Maochang Li, Xiaolong Hou, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. RG PA at SemEval-2021 Task 1: A Contextual Attention-based Model with RoBERTa for Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 623–626, Online. Association for Computational Linguistics.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese Lexical Complexity for Non-Native Readers: A New Dataset. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 477–487, Toronto, Canada. Association for Computational Linguistics.

Desislava Aleksandrova and Vincent Pouliot. 2023. CEFR-based Contextual Lexical Complexity Classifier in English and French. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527, Toronto, Canada. Association for Computational Linguistics.

Jenny Alexandra Ortiz-Zambrano, César Espin-Riofrio, and Arturo Montejo-Ráez. 2022. Transformers for Lexical Complexity Prediction in Spanish Language. Procesamiento del Lenguaje Natural, 69:177–188.

Aleksei V. Abramov and Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics*, 26(2):409–425.

Aleksei V. Abramov, Vladimir V. Ivanov, and Valery D. Solovyev. 2023. Lexical Complexity Evaluation based on Context for Russian Language. *Computación y Sistemas*, 27(1):127–139.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de-Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel

Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence Complexity in Context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199, Online. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. *Computing Research Repository,* arXiv: 1905.07213. Version 1

Valery D. Solovyev, Vladimir V. Bochkarev, Yulia S. Maslennikova, Anna V. Shevlyakova. 2019. Diachronic frequency dictionary of Russian vocabulary.

Ilya Loshchilov and Frank Hutter. 2017. *Decoupled weight decay regularization*. *Computing Research Repository,* arXiv: 1711.05101. Version 3

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. – Association for the Advancement of Artificial Intelligence:7–10.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. Linguistic databases

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pages 2623-2631*.

Valery D. Solovyev, Marina Solnyshkina, Vladimir Ivanov, and Ildar Z. Batyrshin. 2019. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy Systems*, 36(5):4553–4563.