Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

A new dataset for sentence-level complexity in Russian

Vladimir Ivanov
Kazan Federal University
Kazan, Russia;
Innopolis University,
Innopolis, Russia
v.ivanov@innopolis.ru

Elbayoumi Mohamed Gamal
Innopolis University
Innopolis, Russia
m.elbayoumi@innopolis.university

Abstract

Text complexity prediction is a well-studied task. Predicting complexity sentence-level has attracted less research interest in Russian. One possible application of sentence-level complexity prediction is more precise and fine-grained modeling of text complexity. In the paper we present a novel dataset with sentence-level annotation of complexity. The dataset is open and contains 1,200 Russian sentences extracted from SynTagRus treebank. Annotations were collected via Yandex Toloka platform using 7-point scale. The paper presents various linguistic features that can contribute to sentence complexity as well as a baseline linear model.

Keywords: sentence complexity, crowdsourcing, readability
DOI: 10.28995/2075-7182-2023-22-181-190

Набор данных с оценками сложности предложений на русском языке

Иванов Владимир
Казанский федеральный университет
г. Казань, Россия
Университет Иннополис
г. Иннополис, Россия
v.ivanov@innopolis.ru

Эльбайоуми Мохамед Гамаль
Университет Иннополис
г. Иннополис, Россия
m.elbayoumi@innopolis.university

Аннотация

Прогнозирование сложности текста — хорошо изученная задача. Предсказание уровня сложности отдельного предложения привлекает несколько меньший исследовательский интерес. В статье представлен новый набор данных с аннотацией сложности на уровне предложений. Набор данных открытый и содержит 1,200 предложений на русском языке, извлеченных из корпуса SynTagRus. Аннотации собирались через платформу Яндекс Толока по 7-бальной шкале. В статье представлены различные лингвистические признаки, которые могут быть использованы при оценке сложности предложений, а также предложена простая линейная модель.

Ключевые слова: сложность текста на уровне предложения, читабельность, краудсорсинг

1 Introduction

Text complexity prediction is a task studied at various levels of linguistic units ((Crossley et al., 2008; Collins-Thompson and Callan, 2005; Heilman et al., 2008; Shardlow et al., 2021; Shardlow et al., 2020)). The sentence-level complexity evaluation (SCE) subtask takes an intermediate position between the text fragment level (i.e., several coherent sentences) and the level of an individual word/phrase complexity prediction.

Recent works study sets of features that can be used in SCE, including lexical, syntactical features from the target sentence, and contextual features from surrounding sentences (Schumacher et al., 2016; Iavarone et al., 2021). One possible application of sentence-level complexity prediction is more precise modeling of text complexity beyond the passage-level. For longer texts readability measures such as Flesch-Kinkaid formula (Flesch, 1948) (as well as many others) make use of statistics and typically provide a robust solution. However, statistics such as average sentence length and average word length tend to vary a lot when one analyze individual sentence which may produce less robust predictions. Therefore, in such cases a fine-grained model for sentence complexity prediction might be useful.

The SCE task presents issues, at the levels of interpretation of the model's results and feature selection. One of the state-of-the-art approaches is deep neural networks capable to explore a wide range of features and combine them in a hierarchical and non-linear manner. What is more, deep neural networks have been applied in SCE before. For instance, (Schicchi et al., 2020) evaluated the long short-term memory (LSTM) model with attention mechanism in a binary classification of Italian sentences.

Datasets with manual annotations of sentence complexity were created for a number of languages. (Brunato et al., 2018) present a detailed analysis of features that affect human perception of sentence complexity. The authors study the contribution of a set of lexical, morphosyntactic, and syntactic features. The most important features are sentence length, maximum dependency length in a dependency syntax tree, etc.; for sentences with the same length, the most important factors include average word length and lexical density. Analysis of text complexity in Russian academic text was performed in (Solovyev et al., 2018; Solnyshkina et al., 2018), where the main focus is modeling text complexity of a whole text or a passage.

In this paper, we address two issues. First, we present a new dataset with sentence-level complexity annotations on a scale from 1 to 7. The dataset contains 1,200 sentences with more than 23,000 individual complexity judgments. To the best of our knowledge, this is the first dataset of this kind in Russian. Second, we analyze several types of features and evaluate linear models for predicting sentence-level complexity.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents an approach to collect data. Section 4 describes the dataset and its features; Section 5 presents the experimental results on sentence complexity prediction.

2 Related Work

Here, we focus only on works that are closely related to the present study and consider sentence-level complexity datasets and evaluations. In (Inui and Yamamoto, 2001), authors study the relative complexity of sentences in the readability context for deaf people. Authors collected a corpus with pairs of sentences with paraphrases. Modeling complexity was targeted on the classification of paraphrases into three levels/classes ('left', 'right', 'same'). Inui and Yamamoto developed a rule-based method and compared it to the SVM classifier. Later, in (Vajjala and Meurers, 2014), authors evaluated an SVM classifier to predict relative complexity of pairs of complex and simplified sentences. The study (Maqsood et al., 2022) compares different algorithms for SCE in English dataset with seven categories. Classification of sentence difficulty in Arabic language is addressed in (Khallaf and Sharoff, 2021).

(Schumacher et al., 2016) studied models for predicting the reading difficulty of sentences, with and without the surrounding context. They binned sentences according to grade levels (e.g., a sentence from grade 1 was paired with sentences from grades 3-4, 5-6, 7-8, 9-10, 11-12). Authors studied lexical and grammatical features to train a logistic regression classifier and Bayesian ranker. These authors show that considering the context improves predicting sentence readability. The simplest model has only the AoA-based features, which allows to achieve higher score on the dataset. For Russian language sentence-level complexity prediction was addressed in (Ivanov, 2022), but that study used automatically generated complexity scores for sentences extracted from school textbooks.

(Brunato et al., 2018) applied crowdsourcing to model human perception of single-sentence difficulty in Italian and English. These authors investigate a wide set of linguistic features and their importance for human perception of sentence complexity. Brunato et al. analyzed few tens of features, such as 'char_tok' (average number of characters per word) and 'n_tokens' (average number of words per sentence). In their experiments, authors show that syntactic features can play important role in defining the sentence complexity, but 'char_tok' and 'n_tokens' features are always in the top important features as well. What is more, to explicitly control for sentence length, authors applied binning, i.e. sentences were grouped by length (e.g. 10, 15, 20, etc.) up to 35 tokens.

Finally, deep neural networks for sentence complexity classification were proposed in (Lo Bosco et al., 2021). Their model uses the TreeTagger to extract syntactic features, two LSTM layers, and a linear layer. The last layer outputs the probability of a sentence belonging to the easy or complex class. The experimental results show the increased approach effectiveness for both Italian and English, compared with several baselines such as Support Vector Machine, Gradient Boosting, and Random Forest.

3 Data Collection and Annotation

Our approach to dataset collection consists of two parts: selection of sentences and annotating them using the crowdsourcing platform (Yandex Toloka). We sampled sentences from the SynTagRus corpus. This Syntactically Tagged Russian text corpus contains more than 87,000 sentences (https://universaldependencies.org/treebanks/ru_syntagrus/index.html). The Universal Dependency version of SynTagRus is a comprehensive Russian dependency treebank that was developed by the Institute for Information Transmission Problems of the Russian Academy of Sciences (Lyashevkaya et al., 2016; Marneffe et al., 2021). It is a revision of the original SynTagRus treebank that uses the Universal Dependency annotation scheme. The Universal Dependency annotation scheme is a standard annotation scheme for dependency treebanks that is used in many different languages. The treebank covers a wide range of genres, including news articles, fiction books, and academic papers. It is annotated with a variety of linguistic features, including part-of-speech, morphology, syntax, and semantics. We chose the Universal Dependency version of SynTagRus because it is a high-quality treebank that covers a wide range of genres. We also believe that the linguistic features that are annotated in SynTagRus are relevant to the study of sentence-level complexity.

For extracting a sample of sentences for our dataset, we followed the methodology presented in (Brunato et al., 2018). Authors proposed reducing the influence of lexicon by pruning the sentences containing low-frequency lemmas using a lemma frequency list. In our study we use the frequency list developed by Sharov and Lyashevskaya (Lyashevskaya and S.A., 2009).

All the sentences contained in the SynTagRus corpus were grouped into 6 bins based on a different sentence length, i.e. 10, 15, 20, 25, 30, 35 tokens. Sentences in each subset were then ranked according to the average frequency of their lemmas. We extracted for each bin the first 200-top ranked sentences. Therefore, the dataset for annotation contains 1,200 sentences.

Assessments were collected via crowdsourcing of human judgments in the following way. Sentences were randomly shuffled and divided into task pages (one sentence per page). Each assessor should have passed a test for knowledge of Russian language. Out of all (approximately 10,000) such native speakers available at the Yandex Toloka platform, we admitted 30% of assessors with the highest score (according to the platform).

We used several mechanisms to ensure the quality of the data. First, each sentence was evaluated by multiple participants (each sentence got scores from ten assessors), which allowed us to calculate an average complexity score for each sentence and to estimate the level of agreement among the participants. Second, we used "gold standard" sentences in the task. These were sentences for which we already had reliable complexity ratings. The participants were not aware which sentences were the gold ones. Their ratings for these sentences were used to monitor their performance and to adjust their trust scores. If a participant consistently rated the gold standard sentences incorrectly, their future responses were given less weight in the final calculation of the sentence complexity scores.

Assessors were asked to read a sentence and rate how difficult it was on a 7-point scale where 1 means “very easy” and 7 “very difficult”. We chose a 7-point scale because we wanted to have a granular range of complexity ratings. We also wanted to avoid using a binary scale (e.g., easy vs. difficult), as we believe that sentence complexity is a spectrum. There are a number of theoretical bases for using a 7-point scale to measure sentence complexity. One theory is that sentence complexity is a continuous variable, rather than a discrete variable (Gernsbacher, 1999). This means that there are an infinite number of possible levels of complexity, rather than just a finite number of levels. Another theory is that sentence complexity is a multidimensional concept (Fletcher et al., 1986). This means that there are multiple factors that contribute to complexity, such as syntactic complexity, semantic complexity, and lexical complexity. A

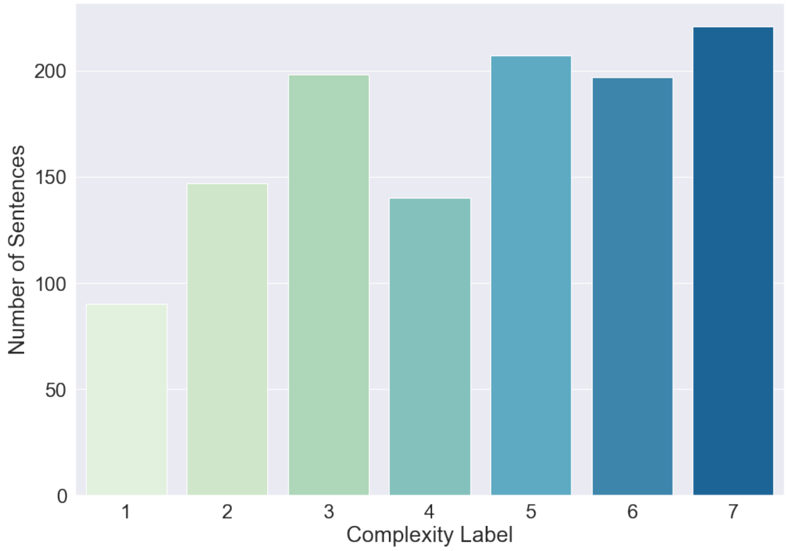

Figure 1: Distribution of scores in the corpus.

7-point scale can be used to capture these multiple factors. Last, but not least, a 7-point scale was applied in a similar previous work done for English and Italian (Brunato et al., 2018); which enables comparison between the datasets in future. In the following section we analyze the collected dataset.

4 Data Analysis

4.1 Analysis of annotations and agreement

The analysis of inter-annotator agreement is an important aspect of dataset validation, as it provides insights into the quality and reliability of the annotations. First, we make use of the Toloka's aggregation method (Dawid-Skene model) that provides a confidence score for each sentence. The mean score for each of seven label categories is above 99%. The distribution of aggregated labels are presented in Figure 1. One can see that overall the dataset is slightly imbalanced towards difficult sentences. The simplest score ('1') has only 90 examples.

Next, for each sentence we calculated the maximum number of assessors who agreed about some category for that sentence. On average, 4.3 assessors per sentence have agreed about a complexity label (with standard deviation of 1.2). Finally, in Figure 2 we plot the deviation of scores with respect to sentence length (bin). This plot clearly shows the correlation between sentence length and complexity score.

Our analysis suggests that the new dataset can provide a useful resource for studying sentence-level complexity in Russian, but caution should be exercised when interpreting the scores, especially for longer sentences. In the following subsection we analyze a set of features, including syntactical.

4.2 Exploring features and correlation

For our study, we extracted features that reflect various facets of sentence complexity, such as:

- **Average_path_length,** which represents the average dependency distance between words in a sentence. Dependency distance is defined as the number of words between two words that have a dependency relationship.
- **Maximum_path_length,** this feature represents the maximum dependency distance between words in a sentence.
- **Num_clauses,** represents the number of clauses in each sentence.
- **Num_phrases,** represents the number of phrases in each sentence.

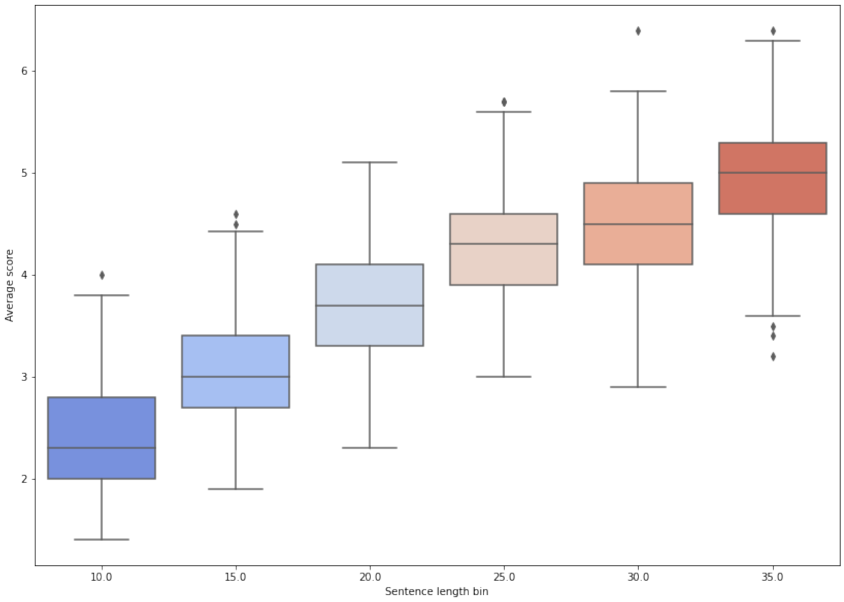

Figure 2: Distribution of complexity scores with respect to sentence length.

- **Num_subordinating_conjunctions** represents the number of subordinating conjunctions in each sentence. It is a measure of syntactic complexity and indicates the degree of subordination in the sentence.
- **Prop_nouns,** represents the proportion of nouns in each sentence based on their POS-tag.
- **Prop_verbs,** represents the proportion of verbs in each sentence.
- **Prop_adjectives,** represents the proportion of adjectives in each sentence.
- **Prop_pronouns,** represents the proportion of pronouns in each sentence.
- **Average_freq,** represents the average frequency of words in the sentence.
- **Avg_token_length,** represents average letters per word.
- **sen_len** is the sentence length measured in characters.

The following examples describe how these features can contribute to complexity.

Dependency distance is the length of the dependency path between a word and its head. A dependency path is the sequence of words that connect a word to its head. For example, in the sentence "The cat that sat on the mat was black," the dependency path between the word "black" and its head "cat" is "cat-sat-on-the-mat-black." The dependency distance between "black" and its head "cat" is 4. We found that dependency distance was positively correlated with sentence complexity. This means that sentences with longer dependency distances were more complex than sentences with shorter dependency distances. One reason why dependency distance is correlated with complexity is that it is a measure of the syntactic complexity of a sentence. Sentences with longer dependency distances have more complex syntax, which makes them more difficult to understand. This observation is supported by other studies of text complexity both at sentence level (Brunato et al., 2018) and at the passage level (Solovyev et al., 2023).

Number of Clauses, a clause is a group of words that has a subject and a verb. A sentence can have one or more clauses. For example, the sentence "The cat that sat on the mat was black" has two clauses: "The cat sat on the mat" and "The cat was black." We found that the number of clauses in a sentence was positively correlated with sentence complexity. This means that sentences with more clauses were more complex than sentences with fewer clauses. One reason why the number of clauses is correlated with complexity is that it is a measure of the semantic complexity of a sentence. Sentences with more clauses have more complex semantics, which makes them more difficult to understand.

Proportion of Nouns and Phrases, the proportion of nouns and phrases in a sentence is a measure of the lexical complexity of a sentence. Nouns and phrases are lexical items, which are words or groups of words that have meaning. We found that the proportion of nouns and phrases in a sentence was positively correlated with sentence complexity. This means that sentences with a higher proportion of nouns and phrases were more complex than sentences with a lower proportion of nouns and phrases. One reason why the proportion of nouns and phrases is correlated with complexity is that it is a measure of the vocabulary load of a sentence. Sentences with a higher proportion of nouns and phrases have a higher vocabulary load, which makes them more difficult to understand.

Table 1: Average values of linguistic features within different bins.

Feature Name	**L10**	**L15**	**L20**	**L25**	**L30**	**L35**
average_path_length	1.65	2.01	2.32	2.44	2.48	2.62
maximum_path_length	8.04	10.80	14.71	18.31	20.84	24.04
num_clauses	0.14	0.24	0.42	0.59	0.69	0.85
num_phrases	2.63	3.13	4.03	4.81	5.37	5.79
num_subord._conjunctions	0.15	0.22	0.33	0.47	0.46	0.67
prop_nouns	0.24	0.24	0.24	0.24	0.25	0.25
prop_verbs	0.13	0.11	0.11	0.11	0.10	0.10
prop_adjectives	0.08	0.08	0.10	0.10	0.10	0.10
prop_pronouns	0.02	0.02	0.02	0.03	0.03	0.02
avg_token_len	5.59	5.57	5.81	5.89	5.96	6.06
avg_freq	6333.51	5837.55	5548.08	5036.07	4687.93	4151.05
sen_len	55.90	83.53	116.18	147.33	178.76	212.18
score	2.43	3.09	3.73	4.27	4.52	4.94
std_score	1.05	1.11	1.22	1.28	1.36	1.31

Table 2: The correlation between linguistic features and sentence complexity within different bins.

Feature Name	**L10**	**L15**	**L20**	**L25**	**L30**	**L35**	**All**
average_path_length	0.15	0.02	0.12	-0.08	-0.10	-0.10	0.39
maximum_path_length	0.05	0.08	0.09	-0.09	-0.12	-0.15	0.58
num_clauses	0.07	0.13	0.08	-0.11	-0.00	-0.01	0.32
num_phrases	-0.14	-0.03	-0.05	-0.12	-0.04	-0.01	0.40
num_subordinating_conjunctions	0.09	0.01	0.03	-0.04	-0.07	-0.08	0.20
prop_nouns	-0.17	-0.01	0.00	0.20	0.16	0.15	0.08
prop_verbs	0.02	0.11	0.03	-0.16	-0.07	0.04	-0.10
prop_adjectives	0.02	-0.05	0.09	0.16	0.14	0.06	0.14
prop_pronouns	0.04	-0.03	0.09	-0.10	0.03	-0.04	0.02
avg_token_len	0.16	0.26	0.30	0.31	0.49	0.34	0.33
avg_freq	0.04	-0.06	-0.15	-0.04	-0.08	-0.19	-0.51
sen_len (in characters)	0.80	0.64	0.59	0.53	0.70	0.72	0.83

To analyze the correlation between linguistic features and sentence complexity, we first calculated the average complexity judgments for six bins of sentences with the same length (10, 15, 20, 25, 30, and 35 tokens). Pearson correlation coefficient is presented in Table 2. As anticipated, the feature with the strongest correlation to sentence complexity is sentence length (measured in characters).However, as indicated in Figure 3, exceptions exist where short sentences have high complexity scores and long sentences have low complexity scores.

Our analysis revealed that some features had a stronger correlation with sentence complexity than others. For example, we observed that the correlation coefficients for various features differ depending on the sentence length bin (see Table 2). Overall, features with the highest correlations are those related to

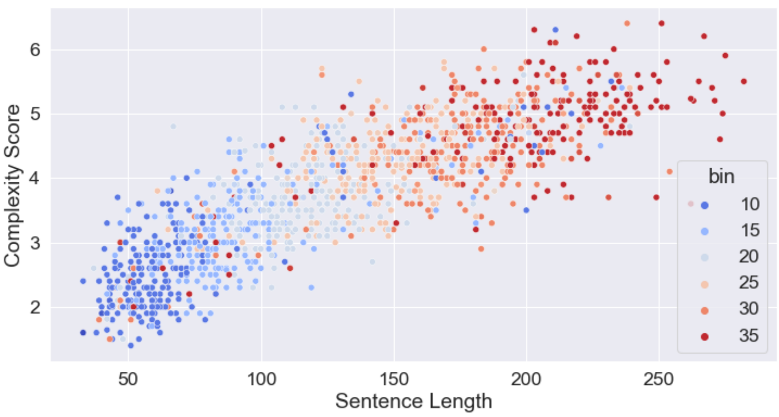

Figure 3: Complexity scores strongly correlate with sentence length. The plot also shows a substantial number of exceptions.

path length, proportions of nouns and phrases, and frequency as well as sentence/token lengths. These findings imply that specific linguistic features substantially influence sentence complexity in Russian. Our results provide insights into the linguistic factors that contribute to sentence-level complexity in Russian and highlight the importance of considering multiple features when assessing sentence complexity.

4.3 Comparison with English / Italian dataset

(Brunato et al., 2018) investigated the correlation between different linguistic features and human judgments of sentence difficulty, using Spearman's rank correlation coefficient. In contrast, our study explores the relationship between linguistic features and sentence complexity using Pearson correlation.

Comparing the findings of the two studies, some similarities can be observed. Both studies found that sentence length (in characters) has a strong positive correlation with sentence complexity. Additionally, the two studies identified similar linguistic features that are significantly correlated with sentence complexity, such as average token length and number of clauses.

In conclusion, while there are some differences in the correlation coefficients between the two studies, the overall findings suggest that certain linguistic features are consistently associated with sentence complexity.

5 Linear Regression Model for Sentence Complexity

Given the correlations coefficients (Table 2), we first train and evaluate a linear regression model. Feature selection shows that the best linear regression model can use three parameters, sentence length in characters (SLC), average path length (APL), and the number of clauses (NCL). The model presented below has MSE=0.32 (±0.03), MAE=0.45 (±0.02) and R^2 value of 0.71, while a model with a single parameter (SLC) has MSE=0.33 and MAE=0.46.

$$Compl.Score = -1.61 + 0.014 * SLC + 0.146 * APL + 0.057 * NCL$$

To confirm SLC is a strong predictor, we run two experiments. First, the Linear regression without the SLC parameter achieves only 0.61 (MSE). Second, we fine-tuned the pre-trained RuBERT model on 80% of the data. The performance of RuBERT is 0.54(MSE) and 0.57(MAE). It is worth noting that the linear model with three parameters systematically underestimates sentences with higher scores (close to 6) and overestimates the complexity of simple sentences with low scores (Fig. 4). Our analysis of such errors shows that the most errors are coming from relying on the SLC value. Therefore, we propose and evaluate models that make use of stratified by sentence length. To this end, we compare performance of

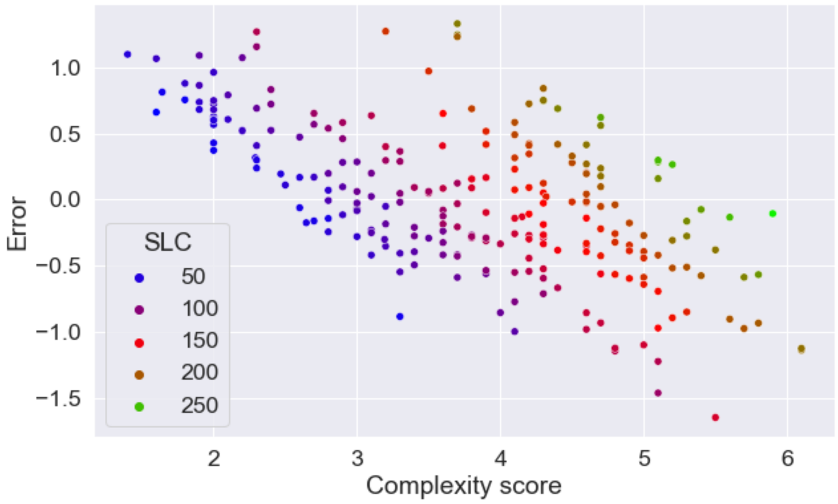

Figure 4: Negative correlation between linear model's errors and the true values of sentence complexity (Error = predicted value - true value).

linear models that either use or not use the SLC feature in each of the six bins. The results are provided in the Table 3.

Table 3: A comparison of models trained in a specific length range with and without sentence length parameter.

	with SLC		w/o SLC	
bin	MSE	MAE	MSE	MAE
L10	**0.254**	0.405	0.257	0.408
L15	0.272	0.417	**0.269**	0.414
L20	0.310	0.443	**0.293**	0.430
L25	0.295	0.438	**0.294**	0.435
L30	**0.295**	0.435	0.298	0.439
L35	0.313	0.413	**0.312**	0.421

6 Conclusion

In this paper, we present a dataset of 1,200 Russian sentences annotated for complexity, collected through crowdsourcing using the Yandex Toloka platform. The analysis of the dataset shows that it is slightly unbalanced towards difficult sentences, with a correlation between sentence length and complexity score. The paper also presents various linguistic features that contribute to sentence complexity in Russian, such as dependency distance, number of clauses and subordinating conjunctions, and proportion of nouns and phrases. The study found that certain features had a stronger correlation with sentence complexity than others. These findings provide insights into the linguistic factors that contribute to sentence-level complexity in Russian, and the dataset can be a useful resource for further research on this topic. The dataset is available at `https://zenodo.org/record/7937828#.ZGJEHC9ByZA`.

Acknowledgments

This work is funded by Russian Science Foundation, grant # 22-21-00334.

References

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? // *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, P 2690–2699, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the american society for information science and technology*, 56(13):1448–1462.

Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Garth Fletcher, Paula Danilovics, Guadalupe Fernandez, Dena Peterson, and Glenn Reeder. 1986. Attributional complexity. an individual differences measure. *Journal of Personality and Social Psychology*, 51:875–884, 10.

Morton Gernsbacher. 1999. Comprehension: A paradigm for cognition. *American Scientist*, 87, 11.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. // *Proceedings of the third workshop on innovative use of NLP for building educational applications*, P 71–79.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. // *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, P 186–199, Online, June. Association for Computational Linguistics.

Kentaro Inui and Satomi Yamamoto. 2001. Corpus-based acquisition of sentence readability ranking models for deaf people. // *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*, P 159–166.

Vladimir Ivanov. 2022. Sentence-level complexity in russian: An evaluation of bert and graph neural networks. *Frontiers in Artificial Intelligence*, 5.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. // *Workshop on Arabic Natural Language Processing*.

Giosué Lo Bosco, Giovanni Pilato, and Daniele Schicchi. 2021. Deepeva: A deep neural network architecture for assessing sentence complexity in italian and english languages. *Array*, 12:100097.

Olga Lyashevkaya, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, and Elena Shakurova. 2016. Universal dependencies for russian: A new syntactic dependencies tagset. *SSRN Electronic Journal*, 01.

Olga Lyashevskaya and Sharov S.A. 2009. *Frequency dictionary of the modern Russian language (the Russian National Corpus)*. 01.

Shazia Maqsood, Abdul Shahid, Muhammad Tanvir Afzal, Muhammad Roman, Zahid Khan, Zubair Nawaz, and Muhammad Haris Aziz. 2022. Assessing english language sentences readability using machine learning models. *PeerJ Computer Science*, 7:e818.

Marie-Catherine Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47:1–52, 03.

Daniele Schicchi, Giovanni Pilato, and Giosué Lo Bosco. 2020. Deep neural attention-based model for the evaluation of italian sentences complexity. // *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, P 253–256. IEEE.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, P 1871–1881, Austin, Texas, November. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and 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)*, P 57–62, Marseille, France, May. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. // *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, P 1–16, Online, August. Association for Computational Linguistics.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. Readability formula for russian texts: a modified version. // *Advances in Computational Intelligence: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II 17*, P 132–145. Springer.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of intelligent & fuzzy systems*, 34(5):3049–3058.

Valery Solovyev, Marina Solnyshkina, Vladimir Ivanov, and Svetlana Timoshenko. 2023. Complexity of russian academic texts as the function of syntactic parameters. // *Computational Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, March 18–24, 2018, Revised Selected Papers, Part I*, P 168–179. Springer.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. // *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, P 288–297, Gothenburg, Sweden, April. Association for Computational Linguistics.