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

June 14–16, 2023

Text complexity as a non-discrete value: Russian L2 text complexity dataset annotation based on Elo rating system

Laposhina Antonina Nikolaevna
Pushkin State Russian Language Institute, Moscow, Russia
antonina.laposhina@gmail.com

Abstract

The task of assessing text complexity for L2 learners can be approached as either a classification or regression problem, depending on the chosen scale. The primary bottleneck in such research lies in the limited availability of appropriate data samples. This study presents a combined approach to create a dataset of Russian texts for L2 learners, placed on a continuous scale of complexity, involving expert pairwise comparisons and the Elo rating system. For this pilot dataset, 104 texts from Russian L2 textbooks, TORFL tests, and authentic sources were selected and annotated. The resulting data is useful for evaluation of the automated models for assessing text complexity.

Keywords: text complexity; Russian as a foreign language; Elo ratings; text complexity dataset; pairwise annotation
DOI: 10.28995/2075-7182-2023-22-278-286

Сложность текста как недискретная величина: экспертная разметка сложности текстов по РКИ на основе рейтингов Эло

Лапошина Антонина Николаевна
Государственный институт русского языка им. А. С. Пушкина, Москва
antonina.laposhina@gmail.com

Аннотация

В исследовании представлен подход к созданию коллекции текстов, аннотированных по сложности для изучающих русский язык как иностранный, на непрерывной шкале, базирующейся на уровнях CEFR. Подход основан на попарной экспертной оценке текстов и системе рейтингов Эло. Исследование выполнено на 104 текстах из специализированных пособий по РКИ и аутентичных источников. Полученные данные полезны для оценки предсказательных моделей уровня сложности текста для изучающих русский язык как иностранный.

Ключевые слова: сложность текста; русский язык как иностранный; рейтинги Эло; попарное сравнение

1 Introduction

The crucial initial step in text complexity studies is to establish a complexity scale and acquire a collection of text samples marked with this scale. The model is then developed and tested based on this data. Depending on the chosen scale, the task of text complexity evaluation can be resolved as a classification problem (resulting in the anticipated class, grades, levels) (Karpov et al. 2014; Francois, Fairon 2012; Reynolds 2016) or a regression problem (yielding any decimal number on a specified scale) (Kate et al. 2010; Seiffe et al. 2022). Hence, not only does the algorithm's design depend on the selection of the scale, but the researcher's fundamental perspective on the concept of text complexity as discrete levels or as a continuum of difficulty.

The primary bottleneck in such research lies in the limited availability of appropriate training data. Most existing datasets consist of discrete complexity levels, such as school materials annotated by grade (Solovyov et al. 2018), age or abstract units (Pitler, Nenkova 2008), or «easy-difficult» binary scale (Sharoff et al. 2008). Regarding the assessment of the text complexity for L2 learners, the Common European Framework of Reference (CEFR) is the preferred choice for the majority of researchers

(Reynolds 2016; Karpov et al. 2014; Schwarm and Ostendorf 2005; Laposhina et al. 2018; Corlatescu et al. 2022).

1.1. CEFR levels as a complexity scale

The Common European Framework of Reference for Language Proficiency (CEFR) establishes universal standards that are utilized worldwide to determine language proficiency levels and serve as a means to acknowledge qualifications obtained from diverse educational systems. In its current version, the 2018 descriptors, the CEFR scale comprises 7 levels ranging from pre-A1 to C2. Nevertheless, even the CEFR descriptor`s authors acknowledge the conventional nature of the proposed scale. «All categories in the humanities and liberal arts are in any case conventional, socially constructed concepts. Like the colors of the rainbow, language proficiency is actually a continuum. Yet, as with the rainbow, despite the fuzziness of the boundaries between colors, we tend to see some colors more than others. Yet to communicate, we simplify and focus on six main colors» (Common European Framework 2018: 34). Private initiatives, such as the sub-level system shown in Figure 1, based on the main CEFR scale material and used in the Polyskills institutes network, further support the practical necessity for a more detailed level scale.

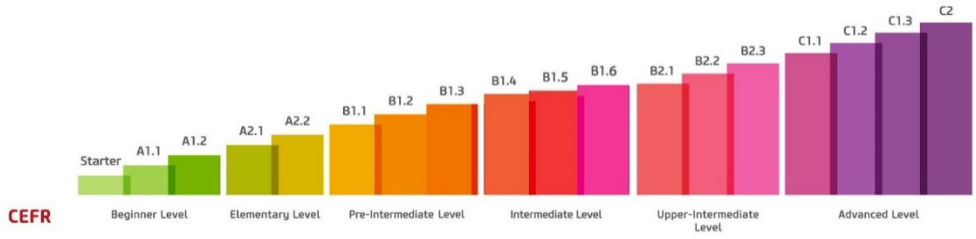


Figure 1: Detailed visualization of CEFR levels for ease of use in teaching practice

In studies on Russian L2 materials, it has been found that teachers commonly use unofficial terms to specify the placement of a particular text within a CEFR level, such as «beginning of B1» «end of B1» «B1+» etc (Laposhina, 2018). Consequently, the formal presentation of text complexity as a conventional scale of levels does not always suit the users' needs and requires more precise information about the place of the text on it.

1.2. Datasets for L2 text complexity assessment task

An automated approach to the complexity assessment of the Russian L2 texts has several examples, most of them are based on datasets with discrete levels, such as the corpus of textbooks annotated by publishers on the CEFR scale can be used (Reynolds 2016; Karpov et al. 2014; Batinic et al. 2016; Laposhina et al. 2018; Corlatescu et al. 2022). However, to create a non-discrete scale, expert annotation is necessary, which can be time-consuming and expensive. Besides, some studies report a low level of expert agreement in the direct task of assigning a text to one of the levels (Laposhina, 2018). To optimize this step, the problem of text complexity annotation can be modified to a pairwise comparison problem of the complexity of two texts (De Clercq et al., 2014; Chen et al.,2013).

Consider the scenario where the creation of a needed dataset is attempted not de novo, but instead utilizing an existing dataset with discrete levels and just refining them by pairwise comparison. The Elo rating system, originally designed to assess the relative strength of chess players (Elo, 1978), has been applied to rank various types of data, including the complexity of the educational content difficulty (Mangaroska et al. 2019), compiling a set of lexical and grammatical topics for Russian L2 learners and evaluating a student's proficiency level in these topics (Jue Hou et al. 2019).

In this article, we examine a combined approach to the ranking of Russian texts for L2 learners on a continuous scale of complexity, which involves expert pairwise comparisons and the Elo rating system. The resulting data is useful for the creation and evaluation of automated models for assessing text complexity for Russian L2 learners.

2 Materials and methods

2.1. Data

For this pilot study, 104 texts from Russian L2 textbooks, reading sections of TORFL tests, and different samples of authentic sources - news sites, blogs, and other media were selected. As an initial complexity level, we used the information about the CEFR level indicated by textbook editors; texts from authentic sources got initial levels C1 (blogs, news, and non-fiction notes) and C2 (academic and official text fragments). The length of the text samples for levels A2-C2 varies from 98 to 127 tokens to save a relatively complete idea of the text fragment. Texts for A1 level are usually shorter, so their length varies from 55 to 103 tokens. The composition of the text sample is shown in Table 1.

Text source\ Number of texts	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
Russian L2 textbooks	13	15	15	13	3	0	**55**
TORFL test reading section	2	2	3	3	2	0	**12**
Authentic sources (news, blogs, online magazines)	0	0	0	0	20	13	**33**
Total	**15**	**17**	**18**	**16**	**25**	**13**	**104**

Table 1: Number of text samples per initial levels and text sources

2.2. Annotation process

To collect data, we have developed a special web interface for pairwise comparison of texts. First, the window with instructions is demonstrated, and a test comparison of two texts with obvious results (elementary and very complex text) so that we could make sure that the expert understood the task correctly. After successfully passing the instruction part, experts were asked to compare texts in pairs, having 3 options: «Left text is more difficult», «Right text is more difficult» and «It's difficult to answer». Pairs of text samples for the main annotation track were generated randomly. In order to save annotators' resources and not show too obvious text pairs the main annotation track, we set an additional rule that a pair of texts should have an equal initial text level or +- one level (e.g. B1 vs B1; B2 vs C1, etc.). An example of an interface is shown in Figure 2.

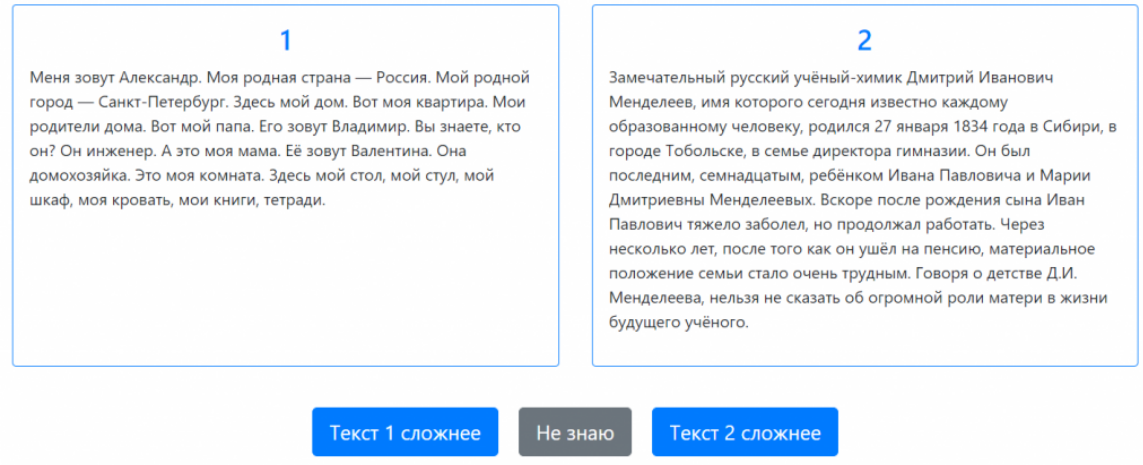


Figure 2. Interface for pairwise annotation of text complexity

The annotation process of this study can be called «expert crowdsourcing». On the other hand, the project was open to all potential annotators who successfully passed the validation stage (as described below). However, the promotion of the annotation project was limited to professional communities of Russian as foreign language teachers. The sample size consisted of 102 anonymous annotation sessions, each of which entailed 10 pairwise comparisons.

To ensure the adequacy and accuracy of the expert's work, we implemented an algorithm for primary verification. Specifically, pairs of texts with obvious right answers (text with initial level 1 or 2 VS 5 or 6) were shown twice during each annotation session. The annotation sessions were considered reliable if both verification tasks were completed accurately. Of the 102 annotation sessions, two had errors in both verification tasks, 17 made mistakes in one of the tasks, and 83 passed both verification tasks and were thus deemed acceptable for analysis.

2.3. Elo ratings

The concept behind the ratings introduced by the Hungarian mathematician Arpad Elo is to calculate the value of a player's victory based on the predictability of his victory. Using a chess analogy, the initial level of the text is «the rating of the player in the preliminary tournament table», which identifies who are the «grandmasters», «beginners», and «players of comparable skill». «Match» is a pairwise comparison of texts by an expert, where the «winner» is the text marked as more difficult.

The initial level of a text is a CEFR level declared by authors and editors, converted to a numeric format (A1 = 1, A2 = 2, B1 = 3, etc.); texts from authentic sources received starting levels 5 and 6 depending on the genre. As a result of each comparison session, texts received points: 1 - if the text annotated was more difficult; 0 - if the text annotated was simpler than the other; 0.5 - if the expert found it difficult to answer.

To adapt this idea to the text complexity task, we used formulas proposed in (Pelanek, 2016). According to them, the probability of «win» for text *i* in a «match» with text *j* is calculated with the following formula:

$$M_{ij} = \frac{1}{1+\exp(L_j - L_i)}$$

where L_i is the level of the text *i* and L_j is the level of the text *j* at the moment of comparison. New level of text *i* as a result of its comparison with text *j* was calculated with the formula:

$$L'_i = L_i + K(P_{ij} - M_{ij})$$

Where L'_i is the new assessment of the text, L_i is the level of the text at the time of comparison, P_{ij} is the score that gets *i* in a «match» (comparison) with text *j*, M_{ij} is the mathematical expectation that the *i*-th text will be more difficult than the *j*-th one. The factor K controls the maximum level adjustment that is possible at one round of comparison, so we set it to 0.25, following the (Ontaelio, 2016). The L values of both texts are updated after each comparison session.

The step-by-step example of comparing two texts is presented below: text 1 «Mailman» (example 1) is a fragment of an authentic text of an interview with a starting level of 5, text 2 «Burglary» (example 2) is a fragment of a journalistic text from the Russain L2 textbook with a level declared by authors of B2, i.e. with initial level 4.

(1) *До того как я сюда устроилась, думала, что почта — это уже прошлый век, вроде городского телефона: мало кто ей пользуется. Но, оказывается, на почту приходит множество людей! Конечно, загрузка у всех отделений разная, но наше, например, находится недалеко от метро, и народу здесь всегда хватает. У меня бывает больше 150 человек в день, всего работают три окошка, то есть получается порядка 500 человек ежедневно. Норма обслуживания на каждого клиента — восемь минут, и это очень мало, конечно. Такого, что скучно и не знаешь, чем себя занять, у нас не бывает. Всегда много клиентов, запросы у всех разные, только и успевай шевелиться.*

(2) *По статистике больше всего квартирных краж совершается в новых районах, так как новосёлы ещё плохо знакомы с соседями. Большая часть краж совершается с 9 до 12 часов (больше половины краж) и с 12 до 15 часов (четверть случаев). Воры предпочитают квартиры на первых и последних этажах: часто залезают в квартиры с крыш. Открытое окно или балкон – серьёзная ошибка. Неважно, на каком этаже вы живёте. Нередко воры залезают в понравившуюся им квартиру из квартиры этажом выше или ниже через балкон. Пытаясь узнать, дома хозяева или уехали, воры придумывают нехитрые манипуляции: периодически звонят в дверь (если хозяева откроют, то всегда можно представиться сотрудником компании, устанавливающей спутниковую связь или продавцом, предлагающем купить картошку, сахар и так далее).*

Annotator N felt that the text 2 about burglaries was more difficult. Therefore, according to the outcome of the comparison, text 1 gets 0 points, and text 2 gets 1 point. Based on the initial levels (5 and 4, respectively), the mathematical expectation of such an outcome will be:

$$M_{mailman} = \frac{1}{1+\exp(4-5)} = 0.73$$

$$M_{burglary} = \frac{1}{1+\exp(5-4)} = 0.27$$

In other words, based on the initial level of these two texts, with a probability of 73% text 1 should have «win» (be marked as more difficult). The annotator, on the contrary, chose text 2 as more difficult, although the probability of such an outcome was only 27%. The new levels for the two given texts will be equal to:

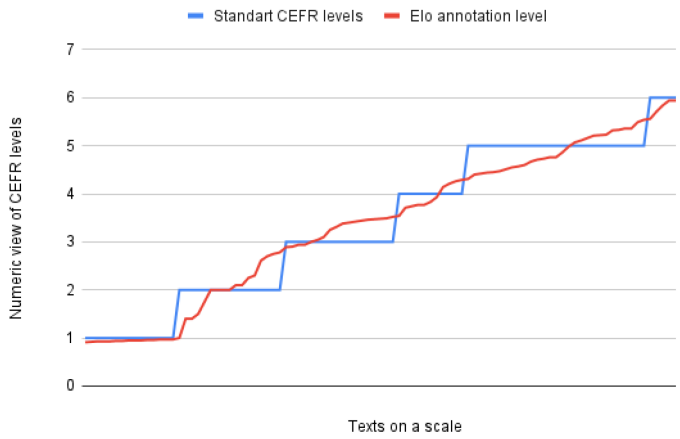
$$NewL_{mailman} = 5+0.25(0-0.73) = 4.82$$

$$NewL_{burglary} = 4+0.25(1-0.27) = 4.18$$

As a result of this comparison, the level of text 1 «Mailman» decreased, and the level of text 2 «Burglary» increased. Then the next comparison takes place, where the initial levels will be considered to be a new value. In total, text 1 participated in 24 comparisons with different texts, and as a result, its level decreased from 5 to 4.6.

3 Results

As a result of the pairwise annotation and calculations described above, we have obtained a collection of 104 texts smoothly distributed along the text complexity CEFR-based scale. Figure 3 illustrates a comparison of the distribution of texts on a scale and their initial CEFR levels.



Following the expert annotation process, the minimum and maximum values of the difficulty level were altered. Whereas the initial collection was marked on a scale of 1 (A1) to 6 (C2), the minimum level value decreased to 0.9, and the maximum increased to 6.8. Consequently, the study generated samples of texts that even native speakers find challenging. Interestingly, the most difficult text in the

collection turned out to be a fragment of an education bill. Figure 4 shows a detailed example of how texts with initial level 3 (B1) were distributed after assessments by expert annotators, from 2.5 to 3.7.

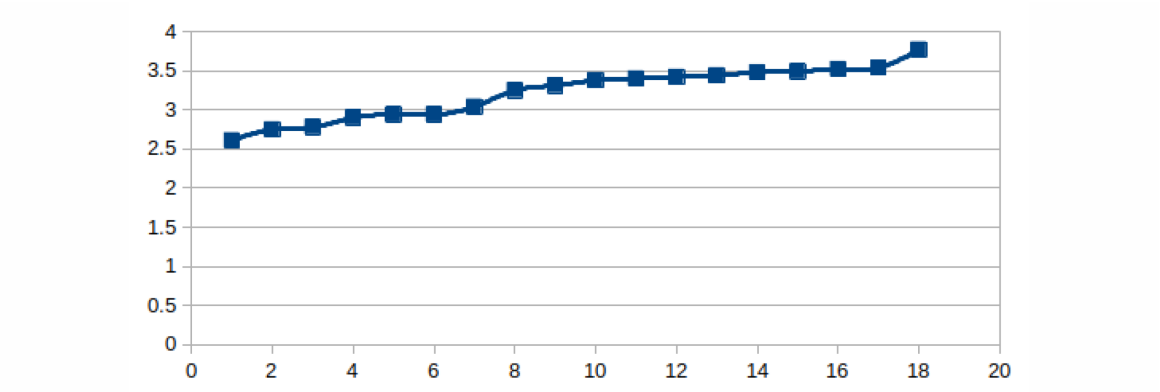


Figure 4: Texts with initial level 3 (B1) after annotation: x-axis: text number in the ranked list, y-axis: the new level value after the annotation, where 2 is equal to A2, 3 is equal to B1, etc.

As illustrated in the figure, the changes in text levels were not revolutionary; level 3 texts were smoothly distributed on a difficulty scale ranging from 2.5 to 3.7. However, some texts shifted to the end of the previous level 2, which is an important finding for evaluating the quality of the text complexity assessment model. Additionally, such a view of the level of text complexity aligns naturally with the idea of language acquisition as a gradual progression from simple to complex.

3.1. Assessment of the validity of expert answers

Annotation design using Elo rating system protects data to some extent from inconsistent markup: even if an expert N made an unexpected decision, next experts and next comparisons will be able to «shift» given text on the scale, thus creating an average expert opinion about the right place of this text on the complexity scale.

For additional verification, we inserted one specific pair of texts into each session, on the basis of which it became possible to calculate the agreement of the annotators. The percent agreement was found to be 79%, indicating an acceptable level of agreement.

3.2. The resulting data as a test set

One of the main purposes of this dataset was to be a test set for the algorithm of the text complexity assessment for Russian L2 learners. In the previous study we developed the ML system trained on 800 texts from Russian L2 textbooks and a set of linguistic features, including lexical, morphological, grammatical, and syntactic ones (Laposhina et al. 2018). The examples of linguistic features are shown in Table 2.

Group of features	Examples of features
Lexical	average word length; percentage of words longer than 4 syllables; lexical diversity (TTR); lexical diversity (MLTD TTR); lexical density; text coverage with a frequency list of 1000, 5000 and 10000 of the most common words from a frequency dictionary; text coverage with vocabulary lists for L2 learners; percentage of abstract words
Grammatical	percentage of each POS in text; percentage of words in the genitive case in text; percentage of verbs in finite forms in text; percentage of words with 1st person tag in text
Syntactic	average sentence length; number of adversarial conjunctions per text; number of coordinating conjunctions per text; average number of punctuators per sentence; text coverage with a list of the 500 most frequent POS trigrams

Table 2: Linguistic features for model training

We have experimented with two linear regression algorithms: ordinary least squares Linear Regression and Ridge Regression (linear least squares with l2 regularization, alpha=1.0) from the scikit-learn library. The best result was achieved by Ridge regression trained on 44 best correlation linguistic features. For the model evaluation, we implemented a twenty-fold cross-validation test that showed an accuracy of 0.82 (±0.05).

However, using standard metrics like mean absolute error and comparing the output of our regression model which is a fractional number to test data from textbooks that is an integer on a discrete scale may not be an efficient approach. For instance, text i from the test set was given from the end of the A2 textbook (so the expected level is 2). The prediction for text i is 3,18 (that may be interpreted as the beginning of B1 level). In terms of linguodidactics, it is not a big mistake (the end of A2 course vs the beginning of B1 course), but it is in terms of mean absolute error.

The present paper aims to fill this gap and provide a test set with texts smoothed on the non-discrete scale. Below are our results of comparing the metrics of the same regression model with two ways of a test set annotation: standard discrete CEFR levels and Elo-based non-discrete levels. Importantly, the texts from this dataset were not used in the model training process.

To evaluate the accuracy of the regression model, which involves comparing actual and predicted values, a widely used approach is to calculate the correlation between the two sets of data. In this study, the Elo-based level scale shows a higher correlation coefficient and a lower mean absolute error compared to the CEFR level scale. Both correlations are statistically significant with p-values less than 0.05 (see Table 3).

Type of complexity scale	**Pearson's correlation coefficient with predicted level**	**p-value**	**Mean absolute error**
CEFR levels	0.81	< 0.05	0.85
Elo-based levels	0.86	< 0.05	0.77

Table 3: Pearson correlation coefficient and mean absolute error values of the predicted and observed levels depending on the chosen scale

To gain more detailed understanding of the comparison results, we analyzed the extent of the discrepancy between the expert opinions and the mathematical model predictions. The severity of the error is dependent on the magnitude of the difference between the expert opinion and the model result. For instance, an error of 0.5 signifies that the model was incorrect by half of a level, which is an acceptable margin of error, as it falls within the range of variation among expert opinions. An error of 1 level or greater suggests more significant discrepancies that require our attention. To estimate the overall magnitude of the prediction error, we used the mean absolute error metric. For the dataset analyzed in this study, the mean absolute error value was 0.77, indicating that, on average, the model's predictions are off by one level. Interestingly, the model tended to overestimate complexity levels in 30% of cases, while underestimating them in 70% of cases. Table 4 displays the distribution of absolute errors between predicted values obtained from a standard CEFR-level-based dataset and an Elo rating system dataset.

Absolute error	**Percent of cases, Elo dataset**	**Percent of cases, CEFR levels**
0-0.5 (good prediction)	38 %	41%
0.51-1 (acceptable prediction)	32%	27%
1.01 - 2 (wrong prediction)	28%	25%
> 2 (dramatically wrong prediction)	2%	7%

Table 4: The proportion of values of the average absolute error of the regression model on the resulting dataset

We consider a difference of less than 0.5 between the predicted and actual values as a correct prediction, which constitutes the majority of cases (38%). The difference greater than 0.5, but within the same level, is an acceptable quality prediction, which represents 32% of the cases. Overall, the model provides correct predictions in 70% of the cases, while being off by more than one level in the remaining 30% of the cases.

The enhanced interpretability of the error report is noteworthy. Now the absolute error distance means the real distance of the text complexity value from the level marked by the experts. This is especially important at the boundaries between levels. For instance, if a text designed for a course ending at level B1 is incorrectly predicted as a text belonging to the subsequent level B2, it will be classified as an error not in an entire level, but rather in a few tenths. The dataset is in the public domain and can be used for scientific purposes.

4 Discussion and conclusion

Construction of suitable datasets is a crucial challenge in the practical implementation of machine learning models, including the L2 linguodidactics field. Given that every language constitutes a multi-faceted living system, any categorization and partitioning into discrete levels are somewhat arbitrary. In this research, we proposed the method of creation of a dataset of texts ranked along the continuous scale of complexity for L2 learners based on CEFR levels. To accomplish this, we relied on the pairwise evaluation of the text complexity by experts and processed the resultant annotations using Elo rating system. This approach provides a non-discrete scale of text complexity, which is more in line with the view of the text complexity as a continuum of difficulty.

Among the limitations of the method, we note the small size of the dataset, which makes it possible to consider it only as a test set, but not a training data collection. Secondly, the assessment of the annotator agreement posed certain challenges. Since the main idea of the method is to compare the text with as many other texts as possible, and pairs of texts for comparison are formed randomly, there are very few identical pairs of comparisons on the basis of which the annotators' agreement can be calculated, unless it is set algorithmically, as was done in this study.

Acknowledgements

The article was prepared in full within the state assignment of Ministry of Education and Science of the Russian Federation for 2020–2024 (No. FZNM-2020-0005).

References

[1] Batinic, D., Birzer, S. Developing an English Language Placement Test for Undergraduate Students: A Crowdsourcing Approach // Educational Technology Society, 18(4), P. 259–271.

[2] Chen, X., Bennett, P. N., Collins-Thompson, K., Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 193–202). ACM.

[3] Clercq, O. D., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., Macken, L. (2014). Using the Crowd for Readability Prediction. Natural Language Engineering, 20(3), 293–325.

[4] Corlatescu, Dragos, Ștefan Ruseti Mihai Dascalu. (2022). ReaderBench: Multilevel analysis of Russian text characteristics. Russian Journal of Linguistics, 26(2), 342–370. https://doi.org/10.22363/2687-0088-30145

[5] Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors / B. North, E. Piccardo, T. Goodier. – Strasbourg: Council of Europe Publishing, 2018. – 227 p.

[6] Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.

[7] Francois, T., Fairon, C. (2012). An 'AI readability' formula for French as a foreign language. In Proceedings of the 2012 Joint Conference on Empirical methods in natural language processing and computational natural language learning (pp. 466–477).

[8] Jue Hou, Maximilian W. Koppatz, Jose Mar´ıa Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, Roman Yangarber. (2015). Modeling language learning using specialized Elo ratings. International Journal of Artificial Intelligence in Education, 25(1), 1–19.

[9] Karpov, N., Baranova, J., Vitugin, F. (2014). Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST) (pp. 91–100).

[10] Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 546–554). Association for Computational Linguistics.

[11] Laposhina, A. N. (2018). Insights from an experimental study on the text complexity for Russian as a foreign language. In Proceedings of the VI Congress of ROPRYAL (pp. 1544–1549). ROPRYAL.

[12] Laposhina, A. N., Veselovskaya, T. S., Lebedeva, M. U., Kupreshchenko, O. F. (2018). Automated Text Readability Assessment For Russian Second Language Learners. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018» (Issue 17 (24), pp. 396–406). Moscow.

[13] Mangaroska, K., Vesin, B., Giannakos, M. (2019). Elo-Rating Method: Towards Adaptive Assessment in E-Learning. In Proceedings of the 19th IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 380–382). IEEE.

[14] Ontaelio, O. (2016, may 19). Count the invisible: how to reliably test the vocabulary [Soschitat' nezrimoe: dostoverno opredelyaem slovarnyj zapac]. Habr.ru. https://habr.com/ru/companies/skyeng/articles/301214/

[15] Pelanek, R. (2016). Applications of the Elo Rating System in Adaptive Educational Systems. Computers & Education, 98, pp. 169-179.

[16] Pitler, E., Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP ’08, page 186, Honolulu, Hawaii. Association for Computational Linguistics.

[17] Reynolds, R. (2016). Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. In Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications

[18] Schwarm, S. E. Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL ’05). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 523–530.

[19] Seiffe, L., Kallel, F., Naderi, B., Moller, S. Roller, R. (2022). Subjective Text Complexity Assessment for German. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 707–714 Marseille, 20-25 June 2022.

[20] Sharoff S., Kurella S., Hartley A. (2008). Seeking needles in the web’s haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference, (TaLC-8), Lisbon, Portugal.

[21] Solovyev V., Ivanov V., Solnyshkina M. (2018). Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics. 3049–3058.