# CEFR Level Prediction for Short Russian L2 Texts: Evaluating Classifiers and Instruction-Based LLMs

**Anna Glazkova**[1,2], **Antonina Laposhina**[3], **Dmitry Morozov**[2,4]
[1] University of Tyumen, Tyumen, Russia
[2] Russian National Corpus, Moscow, Russia
[3] Pushkin State Russian Language Institute, Moscow, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
a.v.glazkova@utmn.ru, antonina.laposhina@gmail.com, morozowdm@gmail.com

**Abstract**

This study explores the automated prediction of text complexity levels for short Russian texts on the Common European Framework of Reference for Languages (CEFR) scale. The dataset consists of 7,322 nonfictional fragments (15–30 words) extracted from textbooks for learners of Russian as a second language and filtered according to linguistic feature distributions typical of each CEFR level, with additional validation conducted by 3 human experts. Each text fragment was annotated with 127 linguistic features, including lexical, morphological, syntactic, and length-based characteristics. We evaluate several approaches to text complexity assessment: traditional machine learning classifiers, fine-tuned transformer models, and instruction-based large language models (LLMs). Among all models, RuBERT achieved the best strict F1-score (47.8%) and the lowest mean absolute error (0.56), while instruction-based LLMs such as YandexGPT captured overall complexity trends but underperformed in exact classification. Feature ablation experiments demonstrated that lexical features are the most informative for CEFR prediction. Our findings confirm that fine-tuned language models currently offer the most reliable results for short-text CEFR assessment in Russian, whereas instruction-based LLMs show potential for qualitative analysis of text difficulty patterns.

**Keywords:** text complexity, CEFR levels, large language models, Russian as a second language

## 1. Introduction

Although text complexity assessment for L2 teaching and learning has been extensively studied at the level of full texts, considerably less attention has been given to short text passages in the context of the Russian language. Yet, assessing proficiency levels for such short text fragments is highly relevant for language teaching practice, particularly when adapting materials from large national corpora—such as concordance lines—for second language learners.

Within the framework of the Common European Framework of Reference for Languages (CEFR), automatic assessment of text complexity is typically approached as a multi-class classification task across six proficiency levels (A1–C2).

In this study, we investigate whether CEFR levels can be reliably predicted for short Russian nonfictional fragments (15–30 words). We construct **RU-CEFR-Short**, a dataset derived from L2 textbooks and refined using a linguistic feature–based representativeness criterion with partial expert validation. We then compare three modeling paradigms under a unified evaluation protocol: traditional classifiers, fine-tuned transformer models, and instruction-based large language models (LLMs).

The contributions of this paper are:

- We introduce **RU-CEFR-Short**, a dataset of 7,322 short CEFR-graded Russian fragments selected using a feature-based filtering procedure. The dataset is publicly available[1].
- We provide a systematic comparison of traditional machine learning models, fine-tuned masked language models, and instruction-based LLMs for fragment-level CEFR prediction.
- We analyze feature informativeness and the impact of class imbalance, showing that fine-tuned masked language models outperform instruction-based LLMs in strict classification, while lexical features remain the strongest predictors of proficiency level.

## 2. Related Work

Research on automatic text complexity assessment has developed along three main directions: feature-based modeling, neural contextual representations, and, more recently, instruction-based LLMs. Most studies operationalize complexity either as readability level (e.g., school grade) or as CEFR proficiency level.

Early research on L2 text complexity prediction for Russian primarily relied on statistical and linguistic features, such as length-based metrics, lex-

[1] https://github.com/Digital-Pushkin-Lab/RU-CEFR-Short



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1081–1091
11–16 May 2026. 

ical frequency, morphological distributions, and measures of syntactic complexity. For instance, Reynolds (2016) and Karpov et al. (2014) focused on classification models, whereas Laposhina et al. (2018) explored regression models using a variety of linguistic features. Their work demonstrated that feature-based approaches remain competitive, especially when domain-specific linguistic indicators are carefully selected.

With the emergence of contextualized embeddings, transformer models have become standard baselines for Russian NLP tasks. Studies such as (Sharoff, 2022; Corlatescu et al., 2022) demonstrated that pretrained language models capture substantial information about linguistic complexity, often outperforming traditional feature-based systems. Nevertheless, most current work operates at the document or paragraph level. The applicability of such models to short fragments remains underexplored.

More recently, Lavrovskiy et al. (2025) compared several modern Russian-language embedding models for CEFR classification, including contextual representations derived from BERT- and GPT-based models. Using these embeddings as input features, they trained traditional classifiers and multilayer perceptrons, showing that BERT-based representations outperform handcrafted linguistic features. However, their study does not evaluate instruction-based large language models in a prompt-based setting, nor does it address fragment-level CEFR prediction from minimal context.

LLM prompt-based approaches to CEFR text complexity assessment have used strategies such as specifying the CEFR scale and providing example passages along with their levels (Kogan et al., 2025), or more detailed lexical and grammatical features from CEFR descriptors or experts in teaching L2 (Katinskaia et al., 2025; Imperial et al., 2024). Although findings show that carefully engineered prompts tend to improve performance (Uchida, 2025), studies indicate that prompt-based LLMs still underperform compared to other models and cannot yet be reliably used for CEFR-related tasks (Benedetto et al., 2025).

A persistent challenge in CEFR-related research is the lack of reliable training data. L2 textbook passages appear to be a natural solution, yet they present clear limitations. Texts from different textbooks often diverge in difficulty due to learners' L1 backgrounds, and even within a single textbook, variation across text genres and pedagogical aims can be substantial (Sharoff, 2022). For short-fragment studies, assigning the proficiency level of a whole text to its excerpts introduces an additional assumption that researchers must accept. Expert ratings may also be constrained by limited inter-rater agreement, especially across adjacent levels (Kogan et al., 2025). To mitigate these issues, this study adopts a combined approach, integrating textbook-based annotations with linguistic features and multiple expert judgments to strengthen the reliability of the dataset.

## 3. Dataset

### 3.1. Corpus of CEFR-graded Texts

The dataset for this study is based on texts from the *RuFoLa corpus* of Russian L2 textbooks and resources, which includes metadata on genre, style, and the CEFR proficiency level specified by the textbook authors (Laposhina et al., 2018). For this study, we selected only nonfictional monologic texts — such as media articles, reports, academic and popular science texts, essays, and everyday narratives — and segmented them into fragments of 15–30 words without breaking sentence boundaries. Each fragment was assigned the CEFR level of its source text. This involves the assumption that a fragment reflects the same level of linguistic complexity as the entire text and that the CEFR label provided in the textbook reliably captures this complexity. To address these limitations, we applied additional data filtering and expert validation procedures, as detailed in Section 2.3.

### 3.2. Linguistic Features

The fragments were annotated with length-based, lexical, morphological, and syntactic features (see Table 1). The **length-based** group of metrics contains simple features commonly used in various readability formulas: average and median word length in characters and syllables, average sentence length in words, and the percentage of words with more than four syllables. In addition, we employ two formulas adapted for Russian that combine these parameters: the Flesch formula in Oborneva's adaptation (Oborneva, 2006), and the Flesch–Kincaid formula in the adaptation by Solovyev et al. (2024).

| Category | Number of features |
|---|---|
| Morphological | 44 |
| Lexical | 27 |
| Syntactic | 48 |
| Length-based | 8 |
| **Total** | **127** |

Table 1: Linguistic features subgroups.

**Morphological Features** are defined as the proportion of each POS and morphological tag relative to the total number of words in the text. All morphological features as well as text tokenization and lemmatization were carried out using the **pymys-**

**tem3**[2], a Python wrapper for Yandex's morphological analyzer Mystem 3.0 (Zobnin and Nosyrev, 2015).

**Lexical Feature set** encompasses several subgroups of different types. The CEFR-graded features capture the proportion of words in the text that occur in 6 CEFR-graded wordlists (Andryushina, 2009) and a list of multiword expressions (Laposhina et al., 2024) compiled for Russian L2 learners. The frequency-based features reflect the proportion of words belonging to different frequency bands, calculated using the New Frequency Dictionary of Russian Vocabulary (Lyashevskaya and Sharov, 2009) and a list of the 5,000 most frequent words from a corpus of texts addressed to children (Maslinskij et al., 2021). Both sets were computed on lemmatized texts, with proper names and toponyms excluded. The lexico-grammatical subgroup covers the proportion of certain indefinite pronouns, as well as the number of parenthetical constructions and causal discourse markers. Finally, the general lexical features comprise the type–token ratio (TTR), lexical density (ratio of content words), the proportion of proper and geographical names, and the proportion of words absent from the Mystem dictionary.

**Syntactic Features** cover the distribution of specific dependency roles extracted using **SpaCy**[3] library (Honnibal and Montani, 2017), the proportion of sentences with a verbal root, and the average maximum depth of the syntactic tree. They also include measures derived from syntactic role analysis, such as the average number of clauses per sentence and the average number of markers characteristic of subordinate constructions per sentence.

### 3.3. Data Filtering and Evaluation

To select the most representative fragments for each proficiency level, we first calculated, for each fragment, the proportion of linguistic features whose values fell within the second and third quartiles for that level. This proportion reflects how typical a fragment is in terms of its linguistic profile. We then ranked all fragments according to this proportion and retained the top 75% with the highest concentration of mid-range features.

To evaluate the proposed filtering procedure and to assess the overall adequacy of the training data, we conducted a selective expert review involving three independent raters, Russian L2 teachers each with more than 10 years of experience in Russian language teaching. A total of 400 fragments were sampled, representing four groups of 100 texts each, balanced by the number of examples per CEFR level:

- **Median** — fragments retained in the final dataset, with the majority of feature values close to the median for their level.
- **T_conflict** — fragments that showed the largest absolute discrepancy between the CEFR level indicated in the textbook and the level predicted by Textometr, an existing features-based ridge regression model for L2 longer text complexity assessment (Laposhina et al., 2018).
- **Q1** — fragments, characterized by an excess of feature values in the first quartile over those in the fourth quartile. These texts are presumed to be more *difficult* than the CEFR level indicated in the textbook.
- **Q4** — fragments, characterized by an excess of feature values in the fourth quartile over those in the first quartile. These texts are presumed to be *easier* than their assigned CEFR level.

Fragments were presented to experts in tabular form. The texts were randomized with respect to textbook-assigned CEFR levels, but all experts received them in the same order. Experts were asked to evaluate the difficulty of each text fragment on a six-point scale, where 1 corresponded to A1 (Beginner) and 6 to C2 (Mastery). The results of the expert evaluation are presented in Table 2.

The inter-rater reliability, measured by Krippendorff's alpha ($\alpha = 0.79$), indicates substantial agreement among the three experts. While strict percent agreement was moderate (39%), allowing $\pm 1$ level difference increased agreement to 87%. The rating results between pairs of raters ranged from $\rho = 0.75$ to $\rho = 0.88$. This suggests that experts broadly share similar judgments of text level, but often differed in the granularity of their ratings. When comparing across text subgroups, the highest relaxed inter-rater agreement (91%) and the strongest correlation between expert ratings and textbook-assigned levels ($\rho = 0.85$) were observed for texts from the **Median** group. This supports the validity of our data filtering methodology based on this criterion. Notably, the highest strict percent agreement (44%) occurred in group **Q4**, which contains texts that are, based on their linguistic features, presumably simpler than their textbooks-assigned CEFR levels. This suggests that indicators of text simplicity are more salient and easily identifiable to experts than those of text complexity. The **T_conflict** group is also noteworthy: the highest CEFR deviation values (1.61) and the lowest correlations between the experts' mean ratings and the textbook-assigned levels indicate that models based on linguistic features can effectively detect

[2] https://github.com/nlpub/pymystem3

[3] https://spacy.io/models/ru#ru_core_news_lg

[4] CEFR level deviation = Absolute value of (mean expert score — textbook's CEFR level) / number of examples.

| Metric | Median | T_conflict | Q1 | Q4 | Total |
|---|---|---|---|---|---|
| Percent agreement strict | 34% | 38% | 40% | 44% | 39% |
| Percent agreement relaxed (± level) | 91% | 83% | 89% | 86% | 87% |
| Krippendorff alpha | 0.82 | 0.65 | 0.62 | 0.85 | 0.79 |
| CEFR level deviation[4] | 0.72 | 1.61 | 0.70 | 0.70 | 0.94 |
| Pearson (experts vs. textbook) | 0.85 | 0.04 | 0.61 | 0.78 | 0.60 |
| Pearson (experts vs. Textometr) | 0.84 | 0.57 | 0.73 | 0.77 | 0.78 |

Table 2: Expert agreement and correspondence with CEFR levels across fragment groups.

cases of inconsistency between the actual level of a text fragment and its textbook labeling. The difficulty of determining the actual level of fragments in this group is also reflected in the comparatively low inter-rater agreement values. As a result of the described data filtering based on the proportion of features falling within the median, we obtained a dataset **RU-CEFR-Short**. Its size and composition are presented in Table 3.

Some examples of sample fragments from the dataset are given in Appendix A.

## 4. Evaluation

### 4.1. Models

We formulated text complexity detection as a multi-class text classification task. Evaluation was performed using three groups of models: traditional machine learning classifiers, fine-tuned language models, and instruction-based LLMs. The machine learning classifiers included a feedforward neural network (**MLP**), a linear support vector classifier (**SVM**), extreme gradient boosting (**XGBoost**), and ridge regression (**Ridge**). Sentence embeddings were obtained using **Sbert**[5] and **E5**[6] (Wang et al., 2024). As model input, we used the concatenation of the embedding and the feature vector. The features were pre-normalized using MinMax scaling.

Fine-tuned language models comprised **RuBERT** (180M parameters)[7] (Kuratov and Arkhipov, 2019) and **RuRoBERTa** (355M)[8] (Zmitrovich et al., 2024). Instruction-based models included Russian-language LLMs **YandexGPT** (8B)[9] and **T-lite** (7.6B)[10], as well as the multilingual **Llama** (8B)[11]. These models were applied using a prompt-based approach. Training parameters are summarized in Appendix B. For both traditional machine learning classifiers and fine-tuned language models, we utilized class weighting, where the weights are inversely proportional to the class frequencies in the training set.

Instruction-based LLMs were applied with a system prompt designed to classify Russian texts according to the CEFR scale. The prompt specifies linguistic criteria for each level, including vocabulary frequency, grammatical constructions, verb forms, sentence complexity, and stylistic markers from the detailed Russian language curriculum (Glazunova and Kolesova, 2022). Illustrative examples are provided for each level, and the model is instructed to output a single CEFR label (A1, A2, B1, B2, C1, or C2) per text without additional commentary.

### 4.2. Metrics

We calculated **the strict F1-score** based on the exact match between the predicted and gold labels. **The relaxed F1-score** was calculated similarly. However, the answer was considered correct if the difference between the predicted and gold CEFR levels is no more than one in absolute value. **MAE** was calculated as the Mean Absolute Error, that is, the average absolute deviation of each prediction from the corresponding actual level.

## 5. Results

### 5.1. General Results

Table 4 presents the performance of the models on the dataset. For traditional machine learning classifiers and fine-tuned language models, we report the average results obtained from five-fold cross-validation and indicate standard deviation values using the symbol ±. The folds were constructed to ensure that fragments originating from the same source text did not appear simultaneously in both the training and test sets. In the case of instruction-based models, the predictions were generated by running the model sequentially on each individual

[5] https://huggingface.co/ai-forever/sbert_large_nlu_ru

[6] https://huggingface.co/intfloat/multilingual-e5-large-instruct

[7] https://huggingface.co/DeepPavlov/rubert-base-cased

[8] https://huggingface.co/ai-forever/ruRoberta-large

[9] https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

[10] https://huggingface.co/t-tech/T-lite-it-1.0

[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

| CEFR level | Number of fragments | Number of words | Mean fragment length (words) ± SD | Mean fragment length (sentences) ± SD |
|---|---|---|---|---|
| A1 | 763 | 18622 | 24.4 ± 4 | 3.4 ± 1.8 |
| A2 | 1136 | 27111 | 23.9 ± 4 | 2.5 ± 1.1 |
| B1 | 2375 | 55879 | 23.5 ± 4 | 2.0 ± 0.8 |
| B2 | 1716 | 39783 | 23.3 ± 4 | 1.9 ± 0.8 |
| C1 | 1182 | 27074 | 22.9 ± 4 | 1.7 ± 0.7 |
| C2 | 150 | 3574 | 23.8 ± 4 | 1.8 ± 0.8 |
| Total | 7322 | 172043 | 23.5 ± 4 | 2.2 ± 1.1 |

Table 3: Data distribution across CEFR levels in the RU-CEFR-Short dataset.

fragment in the dataset. The best result for each metric is highlighted in **bold**, while the second-best result is indicated in ***bold italic***.

Among traditional machine learning classifiers, MLP-based models generally perform better than SVM, XGBoost, and Ridge. In particular, MLP E5 achieves high relaxed F1 scores (86.17% macro, 90.63% weighted). Fine-tuned language models, such as RuBERT, show the best results across most metrics, achieving the highest strict F1 macro (47.80%) and weighted (52.99%) scores, the best relaxed F1 weighted (91.41%), and the lowest MAE (0.5595). RuRoBERTa has slightly lower strict F1 scores but is still competitive, especially on weighted metrics. Instruction-based LLMs (YandexGPT, T-lite, Llama) perform poorly on strict F1 metrics. However, for relaxed metrics, YandexGPT is better than Ridge, showing that these models can capture general patterns even if they fail on exact predictions. Overall, fine-tuned models perform best on most metrics, while MLP models remain a strong alternative for relaxed evaluation.

## 5.2. Feature Informativeness

Figure 1 shows the results of evaluating the informativeness of different feature types in terms of the strict F1 macro. In this experiment, we sequentially removed different types of features from the model input and measured its performance. The results indicate that length-based features have weak impact on the overall performance. Depending on the model and the type of embeddings used, the best results were achieved either by the model trained with all features and embeddings or by the model trained with all features except length-based and the embeddings. The difference in performance between these two input variants was small. This may also be related to the small amount of length-based features.

The removal of lexical features had the greatest impact on classifier performance, highlighting the importance of CEFR-graded wordlists and frequency lists for level prediction. The models trained only on features or only on embeddings generally achieved the lowest results in our evaluation.

## 5.3. Error Analysis

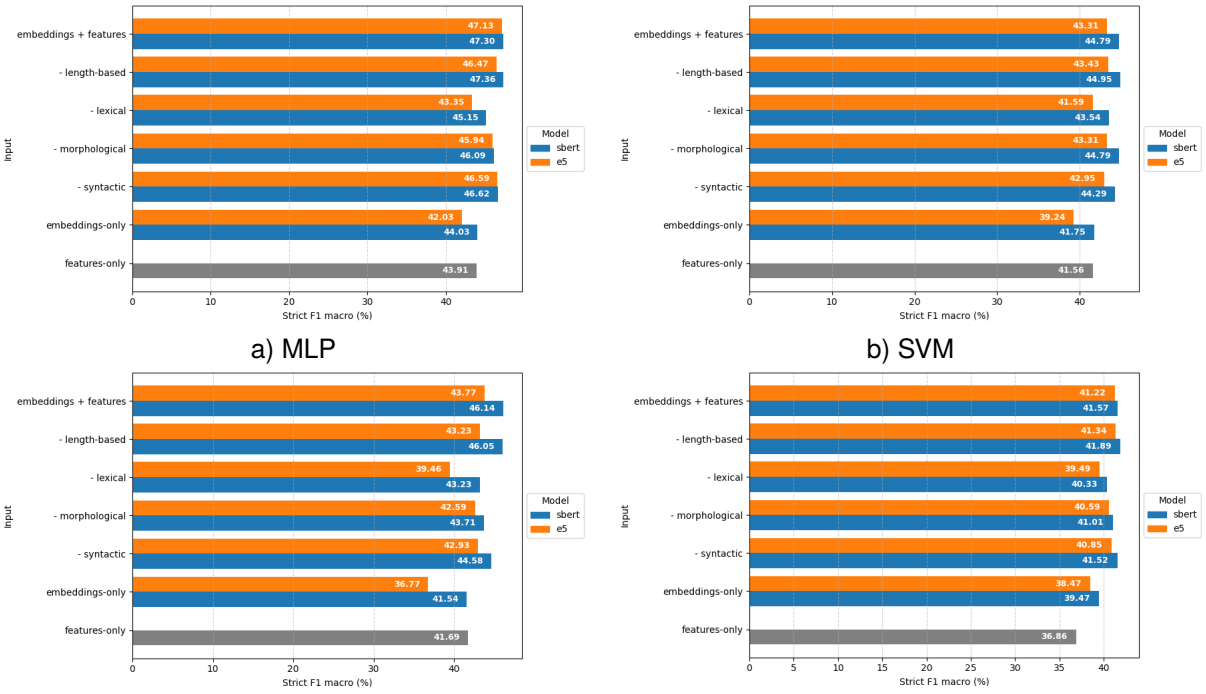
Figure 2 presents the confusion matrix for RuBERT. Out of 7,322 classified instances, 3,903 were predicted correctly, resulting in 3,419 misclassifications. A significant portion of errors (81.9%) occurs between adjacent levels, which explains the large difference between the strict and relaxed metrics. Large deviations ($\geq$2 levels) account for 618 cases (18.1% of all errors and 8.4% of all predictions); severe misclassifications ($\geq$3 levels) are extremely rare: 59 cases were observed (1.7% of all errors and 0.81% of all instances), 50 of which involve texts labeled C1 and C2 that the model predicted as simpler.

Class-wise evaluation further clarifies the model's behaviour across CEFR levels. The best performance is observed for A1 (F1 = 0.67), indicating that elementary texts are relatively well distinguished from higher levels. Intermediate categories show moderate and fairly balanced scores: A2 (F1 = 0.52), B1 (F1 = 0.56), and B2 (F1 = 0.47), which aligns with the substantial confusion observed at the A2–B1 and B1–B2 boundaries. C1 fragments achieve an F1-score of 0.52, suggesting a tendency toward underestimation, while C2 exhibits the weakest results (F1 = 0.17). Among 150 examples labeled as C2, the model correctly classifies 15, giving a recall of 0.10. This low performance is most likely due to the extremely small number of available data sources at this level, as well as the nature of complexity at advanced levels, where difficulty tends to stem less from linguistic features and more from discourse and stylistic factors.

A qualitative analysis of significant errors ($\geq$2 levels) reveals several contributing factors. Some errors reflect genuine model misclassifications, e.g. Example 1 illustrates a fragment that the model predicted as A2, whereas the presence of figurative language (e.g., *people of the same blood; I absorbed it*), interrupted or embedded clauses all suggest that the fragment corresponds to a higher proficiency level. Other errors arise from a mismatch between the proficiency level of an individual fragment and that of the full text (see Example 2), for instance when a very simple excerpt is taken from a complex historical text about women's rights

| Model | Strict F1 macro | Strict F1 weighted | Relaxed F1 macro | Relaxed F1 weighted | MAE |
|---|---|---|---|---|---|
| MLP Sbert | 47.30±2.8 | 51.13±2.9 | 85.48±1.5 | 89.87±0.9 | 0.5986±0.03 |
| MLP E5 | 47.13±2.9 | 51.27±3.3 | **86.17±3.0** | ***90.63±1.3*** | 0.6030±0.01 |
| SVM Sbert | 44.79±1.8 | 49.93±2.9 | 84.40±2.0 | 88.20±1.1 | 0.6708±0.02 |
| SVM E5 | 43.31±1.9 | 48.89±2.5 | 83.83±4.4 | 87.97±1.7 | 0.6743±0.03 |
| XGBoost Sbert | 46.14±2.8 | 52.36±1.9 | 83.39±6.3 | 89.37±1.4 | 0.5877±0.04 |
| XGBoost E5 | 43.77±2.0 | 50.83±2.1 | 83.51±6.4 | 89.20±1.4 | 0.6040±0.04 |
| Ridge Sbert | 41.57±1.3 | 45.18±0.9 | 75.97±4.2 | 83.86±1.1 | 0.7752±0.04 |
| Ridge E5 | 41.22±1.1 | 44.88±1.0 | 75.48±4.4 | 82.92±0.8 | 0.7881±0.02 |
| RuBERT | **47.80±1.9** | **52.99±0.5** | ***85.92±5.7*** | **91.41±0.8** | **0.5595±0.01** |
| RuRoBERTa | ***47.33±3.2*** | ***52.85±2.3*** | 83.62±5.5 | 90.41±0.9 | ***0.5724±0.04*** |
| YandexGPT | 28.62 | 35.69 | 76.04 | 85.84 | 0.7665 |
| T-lite | 20.39 | 26.29 | 66.00 | 78.08 | 1.5153 |
| Llama | 17.85 | 25.34 | 62.65 | 76.95 | 1.3646 |

Table 4: Results. The scores for all F1-score variants are given in percentages.




a) MLP b) SVM

c) XGBoost d) Ridge

Figure 1: Feature type informativeness.

struggles. Another data-related issue concerns ambiguous cases in which multiple CEFR levels could plausibly apply. For instance, in a Russian course labeled A1, a text with simple grammatical structures formally contains very difficult words (26% not included in A1), but it may be perceived as easier by learners familiar with European languages due to international words (e.g., *optimistic, impulsive, aggressive*). In contrast, the model provides a more accurate CEFR level for learners without such language background (B1 level).

(1) Мы были как люди одной крови, и это очень важно: страсть уходит, а вот это, базовое, остаётся. А как я была счастлива на его концертах, по-настоящему счастлива, я впитывала это! *[We were like people of the same blood, and that's very important: passion fades, but this fundamental bond remains. And how happy I was at his concerts — truly happy — I absorbed every moment of it!]* (BERT prediction: A2, dataset level: C1)

(2) Впервые этот праздник отметили в 1911 году, но только 19 марта, в Австрии, Дании, Германии и Швейцарии. *[This holiday was celebrated for the first time in 1911, but only on March 19, in Austria, Denmark, Germany, and Switzerland.]* (BERT prediction: B1, dataset level: C1)

This analysis shows that the model performs best at the lower boundary of the CEFR scale

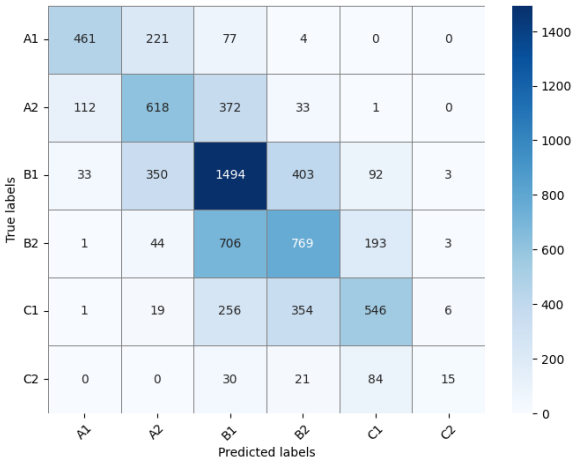



Figure 2: Confusion matrix (RuBERT).

and struggles particularly with distinguishing upper-intermediate and advanced levels. It also highlights the challenge of obtaining high-quality training data for short text fragments with reliable complexity labels derived from textbooks, even when applying a filtering procedure based on linguistic features.

### 5.4. Experiments on Balanced Data

The original dataset is unbalanced across CEFR classes. To examine how this imbalance influences classification performance, we conducted an additional experiment on a more balanced dataset. First, we merged the C1 and C2 categories into a single class. This decision was motivated by their relatively small sizes and the high degree of linguistic similarity between these two advanced proficiency levels, which often leads to systematic confusion in automatic classification. Second, we reduced the size of the B1 class, which was by far the largest in the dataset. To achieve this, we randomly undersampled B1 until its size matched that of the second-largest class. A new class distribution is shown in Table 5. Similar to previous experiments, we applied class weighting to mitigate remaining imbalance effects. Then, we calculated evaluation metrics on this balanced set using five-fold cross-validation.

The experiment produced results similar to those obtained earlier (see Table 6). The best overall performance was achieved by RuBERT, while MLP E5 and RuRoBERTa showed comparable outcomes. The average increase in strict macro F1 on the balanced dataset was 8–10 percentage points, which can be explained by the reduced number of classes and the decreased class imbalance.

| CEFR level | Number of fragments |
|---|---|
| A1 | 763 |
| A2 | 1136 |
| B1 | 1716 |
| B2 | 1716 |
| C1-C2 | 1332 |
| Total | 6663 |

Table 5: Data distribution across CEFR levels after balancing classes.

## 6. Conclusion

In this study, we first present the **RU-CEFR-Short** dataset of 7,322 short CEFR-graded Russian non-fictional texts. These fragments were extracted from longer texts in L2 textbooks and resources, and only those with a high proportion of linguistic features near the median values for each level were retained for the final dataset. The filtering procedure was selectively validated through expert annotation.

Examination of the expert annotations showed only moderate strict inter-rater agreement, highlighting the challenges of formalizing text difficulty without taking into account learner-specific factors (e.g., native language) and instructional objectives. At the same time, the high agreement among experts under a ±1 level tolerance supports the overall validity of the training data and underscores the importance of employing relaxed evaluation metrics for this task. Consequently, we employed a set of metrics capturing different aspects of classification quality, including relaxed F1 and mean absolute error (MAE). Future research directions here include a detailed analysis and additional expert commentary on cases of discrepancies between the model's predictions and textbook-assigned levels, as well as the potential development of more fine-grained annotations for the text fragments, taking into account the learners' native language and the type of instructional task for which each text is intended.

Second, we evaluated various text complexity assessment methods for Russian on this dataset, comparing traditional classifiers, fine-tuned transformer models, and instruction-based LLMs. The results demonstrate that fine-tuned transformer models, particularly RuBERT, achieved the best overall performance, showing the highest strict and relaxed F1-scores and the lowest MAE. While instruction-based LLMs underperformed in exact classification, they were able to capture general difficulty trends, indicating potential for improvement through prompt engineering. Feature ablation confirmed the central role of lexical and frequency-based features, whereas length-based metrics contributed minimally. Overall, the study contributes to the development of reliable approaches to auto-

| Model | Strict F1 macro | Strict F1 weighted | Relaxed F1 macro | Relaxed F1 weighted | MAE |
|---|---|---|---|---|---|
| MLP Sbert | 55.67±1.0 | 54.48±1.1 | 90.62±1.8 | 90.58±2.2 | 0.5547±0.03 |
| MLP E5 | **56.17±2.4** | ***54.82±2.7*** | 90.62±1.7 | 90.50±2.2 | 0.5508±0.03 |
| SVM Sbert | 52.14±1.7 | 50.83±1.5 | 88.23±1.3 | 88.51±1.4 | 0.6216±0.03 |
| SVM E5 | 52.53±1.0 | 50.89±1.2 | 87.33±1.6 | 87.51±1.9 | 0.6311±0.03 |
| XGBoost Sbert | 54.33±2.8 | 53.05±2.5 | 89.65±0.9 | 89.75±1.2 | 0.5848±0.04 |
| XGBoost E5 | 54.65±1.5 | 53.39±1.4 | 89.69±1.4 | 89.79±1.6 | 0.5825±0.03 |
| Ridge Sbert | 50.42±1.7 | 49.04±1.4 | 86.28±1.1 | 86.78±1.4 | 0.6582±0.03 |
| Ridge E5 | 50.86±0.6 | 49.43±0.7 | 85.56±1.1 | 86.03±1.5 | 0.6634±0.02 |
| RuBERT | ***56.15±2.1*** | **54.88±2.3** | **90.92±1.5** | ***91.03±1.6*** | ***0.5491±0.03*** |
| RuRoBERTa | 55.96±3.3 | 54.79±3.7 | ***90.84±1.5*** | **91.05±1.4** | **0.5444±0.04** |

Table 6: Results on the balanced dataset. The scores for all F1-score variants are given in percentages.

matic CEFR prediction for short texts and outlines directions for enhancing both data quality and model robustness in future work. The outcomes of this study may support the development of educational applications and research on tailoring corpus interfaces and content to users' Russian language proficiency levels.

## 7. Ethics Statement

Short texts for the dataset were extracted from the RuFoLa corpus, which is derived from published Russian L2 textbooks and online resources and is shared under fair use for research and educational purposes only. To ensure ethical transparency, all fragments were segmented, randomized, and anonymized so that the reconstruction or reuse of the original full texts is impossible. All original sources are appropriately acknowledged and cited in the accompanying bibliography file. Authors may request the removal of any fragment at [email]. Users of the dataset are expected to cite it appropriately and to use it solely for non-commercial research and educational purposes.

## 8. Limitations

We identified the following limitations of our study.

**Limited Text Genre and Source.** The dataset is derived exclusively from L2 textbook texts, which may not fully represent the linguistic diversity of authentic Russian language use. The range of texts for the research is limited to non-fiction texts. This choice was motivated by two primary factors. First, we did not include fiction because the learning goals for reading stories differ from those for reading factual texts. With fiction, the focus is more on understanding feelings and style, not just the facts. Second, fiction texts often use special grammar, creative language, and storytelling techniques that differ from non-fiction. If we included them, it would make our data less clear and harder to find patterns that match CEFR levels. This decision is supported by earlier research (Sharoff, 2022).

**Fragment-Level CEFR Prediction.** CEFR levels are traditionally defined at the level of extended discourse rather than short excerpts. In this study, proficiency is inferred from fragments of 15–30 words, which provide only local lexical and morphosyntactic evidence and do not capture discourse-level coherence or pragmatic complexity. Therefore, the results should be interpreted as assessing the feasibility of CEFR prediction from minimal linguistic context rather than as a substitute for full-text proficiency evaluation.

**Data Imbalance.** The dataset exhibits class imbalance, particularly for the C2 level, which contains only 150 examples. Although we conducted additional experiments on a balanced subset, the generalizability of the models across all CEFR levels remains constrained by data availability. Additionally, fragments were selected using a feature-based filtering procedure that retains the top 75% most representative examples per level. This threshold reflects a compromise between dataset purity and size. While the procedure removes linguistically atypical fragments, it does not modify textbook-assigned CEFR labels and is independent of model predictions.

**Underperformance of Instruction-Based LLMs.** The performance of instruction-based LLMs was notably weaker than that of fine-tuned models, suggesting that current prompt-based methods are not yet sufficient for precise CEFR classification. This result may be partially explained by the fact that the evaluated instruction-based models have up to 8B parameters. Larger or proprietary LLMs were not included due to computational constraints. Consequently, the observed performance gap should be interpreted within this parameter regime rather than as a general limitation of instruction-based modeling approaches. Future work may explore larger-scale LLMs and more advanced prompting strategies, including structured reasoning prompts and few-shot demonstrations.

## References


N. P. Andryushina. 2009. Leksicheskie minimumy v Rossijskoj sisteme testirovaniya po russkomu yazyku kak inostrannomu [Lexical Minima in the Russian system of testing in Russian as a Foreign Language]. *Problemy istorii, filologii, kul'tury [Problems of History, Philology, Culture]*, (2 (24)):767–771. In Russ.

L. Benedetto, G. Gaudeau, A. Caines, and P. Buttery. 2025. Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

D. Corlatescu, Ș. Ruseti, and M. Dascalu. 2022. ReaderBench: Multilevel analysis of Russian text characteristics. *Russian Journal of Linguistics*, 26(2):342–370.

O.I. Glazunova and D. V. Kolesova. 2022. *Russian as a Foreign Language: Curriculum*. Russkiy Yazyk. Kursy, Moscow, Russia.

M. Honnibal and I. Montani. 2017. spaCy: Industrial-strength natural language processing in Python. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics - System Demonstrations*, pages 85–90.

J.M. Imperial, G. Forey, and H.T. Madabushi. 2024. Standardize: Aligning language models with expert-defined standards for content generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA. Association for Computational Linguistics.

N. Karpov, J. Baranova, and F. Vitugin. 2014. Single-sentence readability prediction in Russian. In *Proceedings of Analysis of Images, Social Networks, and Texts Conference (AIST)*, pages 91–100, Nizhny Novgorod, Russia. National Research University Higher School of Economics.

A. Katinskaia, D. Vu Anh, J. Hou, U. Vanhatalo, Y. Wu, and R. Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 594–611, Vienna, Austria. Association for Computational Linguistics.

D. Kogan, M. Schumacher, S. Nguyen, M. Suzuki, M. Smith, C. S. Bellows, and J. Bernstein. 2025. Ace-CEFR: A dataset for automated evaluation of the linguistic difficulty of conversational texts for LLM applications. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 594–611, Vienna, Austria. Association for Computational Linguistics.

Y. Kuratov and M. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 333–339.

A. N. Laposhina, T. A. Khramchanka, and M. Yu. Lebedeva. 2024. Multi-word expressions for Russian L2 learners: Corpora-based selection with expert verification. *Research Result. Theoretical and Applied Linguistics*, 10(2):117–137. In Russ.

A. N. Laposhina, T. S. Veselovskaya, M.U. Lebedeva, and O. F. Kupreshchenko. 2018. Automated text readability assessment for Russian second language learners. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*, volume 17, pages 396–406, Moscow, Russia. Dialog.

V. A. Lavrovskiy, N. S. Lagutina, and O. B. Lavrovskaya. 2025. Modern Russian-language texts models comparison for the task of CEFR levels classification. *Modeling and Analysis of Information Systems*, 32(3):298–310.

O. N. Lyashevskaya and S. A. Sharov. 2009. *Chastotnyj slovar' sovremennogo russkogo yazyka (na materialakh Nacional'nogo korpusa russkogo yazyka) [Frequency Dictionary of Contemporary Russian Language (Based on the Russian National Corpus)]*. Azbukovnik, Moscow. In Russ.

K. Maslinskij, E. Lekarevich, and L. Aleinik. 2021. Korpus russkoj prozy dlya detej i junoshestva [a corpus of Russian prose for children and adolescents]. Repository of Open Data on Russian Literature and Folklore, V2. In Russ.

I. V. Oborneva. 2006. *Avtomatizirovannaja ocenka slozhnosti uchebnyx tekstov na osnove statisticheskix parametrov [Automatic evaluation of the complexity of educational texts on the basis of statistical parameters]*. Ph.D. thesis, Ph.D. thesis. In Russ.

R. Reynolds. 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 Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA. Association for Computational Linguistics.

S. A. Sharoff. 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics*, 26(2):371–390.

V. Solovyev, V. Ivanov, and M. Solnyshkina. 2024. Readability formulas for three levels of Russian school textbooks. *Journal of Mathematical Sciences*, 285:100–111.

S. Uchida. 2025. Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction*, 14(1):2078.

L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

D. Zmitrovich, A. Abramov, A. Kalmykov, V. Kadulin, M. Tikhonova, E. Taktasheva, D. Astafurov, M. Baushenko, A. Snegirev, T. Shavrina, S. Markov, V. Mikhailov, and A. Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

A. I. Zobnin and G. V. Nosyrev. 2015. Mystem 3.0 Morphological Analyzer. *Proceedings of the V.V. Vinogradov Russian Language Institute*, 3(6):300–307.

## A. Examples of CEFR Levels

Examples 4-6 illustrate sample fragments from different CEFR levels included in the final dataset:

(3) **A1 level.** Весной, летом и осенью почти каждую субботу он играет в футбол. У них в школе есть футбольная команда. А зимой он играет в хоккей.
*[In spring, summer, and autumn he plays football almost every Saturday. His school has a football team. In winter he plays hockey.]*

(4) **B1 level.** В вузе легче встретить людей с похожими интересами и целями, что очень объединяет студентов. Поэтому многие из них становятся хорошими друзьями на всю жизнь.
*[At university, it is easier to meet people with similar interests and goals, which brings students together. As a result, many of them become lifelong friends.]*

(5) **C1 level.** Природа вложила в человека некоторые врождённые инстинкты, как то: чувство голода, половое чувство и т.п., и одно из самых сильных чувств этого порядка — чувство собственности. *[Nature has instilled in humans certain innate instincts, such as hunger, sexual desire, and so on, and among the strongest of these is the sense of ownership.]*

## B. Model Architectures and Training Parameters

The following models and configurations were used in the experiments:

- **MLP.** Two hidden layers (256 and 128 units, ReLU activation), Softmax output layer, Dropout = 0.5. Optimizer: Adam (lr = 1e-3, $\beta_1$ = 0.9, $\beta_2$ = 0.999, $\epsilon$ = 1e-8). Trained for 100 epochs with early stopping (patience = 20).
- **SVM.** Linear Support Vector Classifier (C = 1.0, kernel = linear, tol = 1e-4, max_iter = 1000).
- **XGBoost.** Objective = multi:softmax, n_estimators = 100, max_depth = 3, learning_rate = 1e-1, gamma = 0, min_child_weight = 1, subsample = 1, colsample_bytree = 1.
- **Ridge Classifier.** Alpha = 1.0, solver = auto, tolerance = 1e-5, max_iterations = None.
- **RuBERT.** Fine-tuned for 5 epochs; maximum sequence length = 128; learning rate = 4e-5; optimizer: AdamW (lr = 4e-5, $\beta_1$ = 0.9, $\beta_2$ = 0.999, eps = 1e-8); batch size = 4.
- **RuRoBERTa.** Fine-tuned for 5 epochs; maximum sequence length = 128; learning rate = 5e-6; optimizer: AdamW (lr = 5e-6, $\beta_1$ = 0.9, $\beta_2$ = 0.999, eps = 1e-8); batch size = 4.
- **Instruction-based LLMs.** Decoding parameters: temperature = 0.1; maximum number of generated tokens = 3.

## C. Prompting Details

For prompting instruction-based LLMs, we compiled structured lists of linguistic criteria intended to characterize each CEFR proficiency level from the detailed Russian language curriculum (Glazunova and Kolesova, 2022). These criteria operationalize level descriptions in terms of lexical frequency bands, grammatical constructions, and morphosyntactic patterns.

For example, the A1 level was specified as follows (translated from Russian):

*A1: highly frequent everyday vocabulary (approximately the first 700 most frequent lexical items); basic motion verbs (*идти-ходить, ехать-ездить*) without prefixes in various inflectional forms; prepositional case with o; dative case expressing recipient (*позвонить маме *'to call one's mother'); genitive case expressing absence (*нет урока *'there is no lesson') and possession (*тетрадь студента *'the student's notebook'); constructions with the prepositions* из, для, без, около, после *followed by genitive; simple sentences or short complex sentences with subordinating conjunctions (*что, потому что*);*

*basic impersonal constructions with the dative (*мне холодно *'I am cold').*

The model was instructed to analyze a Russian-language text fragment and assign a CEFR level (A1–C2). The prompt included level descriptions and illustrative examples.

Two prompt configurations were evaluated. In the first configuration, the original expert-formulated level descriptions were reformatted as structured feature lists to emphasize discrete linguistic criteria. In the second configuration, each level description was compressed into a concise summary (up to 200 characters) generated with ChatGPT-4o, aiming to provide a more compact and abstract representation of proficiency characteristics.

For YandexGPT, the feature-list prompt yielded a strict F1 macro of 27.51% and a relaxed F1 macro of 74.04%. The compressed-description prompt resulted in a strict F1 macro of 25.69% and a relaxed F1 macro of 69.%. Thus, prompt reformulation did not lead to performance improvements for YandexGPT.

In addition, two multilingual instruction-tuned models, such as Phi-4-mini-instruct[12] and Aya Expanse 8B[13], were evaluated under the same protocol. Their performance was substantially lower than that of the models reported in the main experimental section: Phi-4-mini-instruct achieved a strict F1 macro of 14.69% and a relaxed F1 macro of 57.53%, while Aya Expanse 8B obtained 11.45% and 55.11%, respectively.

Overall, the results indicate that prompt-level modifications and alternative multilingual instruction-based models did not close the performance gap with fine-tuned transformer models in fragment-level CEFR classification.

[12] https://huggingface.co/microsoft/Phi-4-mini-instruct

[13] https://huggingface.co/CohereLabs/aya-expanse-8b