RuCoLA: Russian Corpus of Linguistic Acceptability

Vladislav Mikhailov[1]*, **Tatiana Shamardina**[2]*, **Max Ryabinin**[3,4]*
Alena Pestova[3], **Ivan Smurov**[2], **Ekaterina Artemova**[5,6]
[1]SberDevices, [2]ABBYY, [3]HSE University,
[4]Yandex, [5]Huawei Noah's Ark Lab,
[6]Center for Information and Language Processing (CIS), MaiNLP lab, LMU Munich, Germany
Correspondence: vmikhailovhse@gmail.com

Abstract

Linguistic acceptability (LA) attracts the attention of the research community due to its many uses, such as testing the grammatical knowledge of language models and filtering implausible texts with acceptability classifiers. However, the application scope of LA in languages other than English is limited due to the lack of high-quality resources. To this end, we introduce the Russian Corpus of Linguistic Acceptability (RuCoLA), built from the ground up under the well-established binary LA approach. RuCoLA consists of 9.8k in-domain sentences from linguistic publications and 3.6k out-of-domain sentences produced by generative models. The out-of-domain set is created to facilitate the practical use of acceptability for improving language generation. Our paper describes the data collection protocol and presents a fine-grained analysis of acceptability classification experiments with a range of baseline approaches. In particular, we demonstrate that the most widely used language models still fall behind humans by a large margin, especially when detecting morphological and semantic errors. We release RuCoLA, the code of experiments, and a public leaderboard[1] to assess the linguistic competence of language models for Russian.

1 Introduction

Recent NLP research has approached the linguistic competence of language models (LMs) with *acceptability judgments*, which reflect a sentence's well-formedness and naturalness from the perspective of native speakers (Chomsky, 1965). These judgments have formed an empirical foundation in generative linguistics for evaluating humans' grammatical knowledge and language acquisition (Schütze, 1996; Sprouse, 2018).

Borrowing conventions from linguistic theory, the community has put much effort into creating linguistic acceptability (LA) resources to explore whether LMs acquire grammatical concepts pivotal to human linguistic competence (Kann et al., 2019; Warstadt et al., 2019, 2020). Lately, similar non-English resources have been proposed to address this question in typologically diverse languages (Trotta et al., 2021; Volodina et al., 2021; Hartmann et al., 2021; Xiang et al., 2021). However, the ability of LMs to perform acceptability judgments in Russian remains understudied.

	Language	Size	%
CoLA	**English**	10.6k	70.5
ItaCoLA	**Italian**	9.7k	85.4
RuCoLA	**Russian**	13.4k	71.8

Table 1: Comparison of RuCoLA with related binary acceptability classification benchmarks: CoLA (Warstadt et al., 2019) and ItaCoLA (Trotta et al., 2021). **%**=Percentage of acceptable sentences.

To this end, we introduce the Russian Corpus of Linguistic Acceptability (RuCoLA), a novel benchmark of 13.4k sentences labeled as acceptable or not. In contrast to related binary acceptability classification benchmarks in Table 1, RuCoLA combines in-domain sentences manually collected from linguistic literature and out-of-domain sentences produced by nine machine translation and paraphrase generation models. The motivation behind the out-of-domain set is to facilitate the practical use of acceptability judgments for improving language generation (Kane et al., 2020; Batra et al., 2021). Furthermore, each unacceptable sentence is additionally labeled with four standard and machine-specific coarse-grained categories: morphology, syntax, semantics, and hallucinations (Raunak et al., 2021).

The main contributions of this paper are the following: (i) We create RuCoLA, the first large-scale acceptability classification resource in Rus-

*Equal contribution.
[1]Available at rucola-benchmark.com

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5207 - 5227
December 7-11, 2022

sian. (ii) We present a detailed analysis of acceptability classification experiments with a broad range of baselines, including monolingual and cross-lingual Transformer (Vaswani et al., 2017) LMs, statistical approaches, acceptability measures from pretrained LMs, and human judgements. (iii) We release RuCoLA, the code of experiments[2], and a leaderboard to test the linguistic competence of modern and upcoming LMs for the Russian language.

2 Related work

2.1 Acceptability Judgments

Acceptability Datasets The design of existing LA datasets is based on standard practices in linguistics (Myers, 2017; Scholz et al., 2021): binary acceptability classification (Warstadt et al., 2019; Kann et al., 2019), magnitude estimation (Vázquez Martínez, 2021), gradient judgments (Lau et al., 2017; Sprouse et al., 2018), Likert scale scoring (Brunato et al., 2020), and a forced choice between minimal pairs (Marvin and Linzen, 2018; Warstadt et al., 2020). Recent studies have extended the research to languages other than English: Italian (Trotta et al., 2021), Swedish (Volodina et al., 2021), French (Feldhausen and Buchczyk, 2020), Chinese (Xiang et al., 2021), Bulgarian and German (Hartmann et al., 2021). Following the motivation and methodology by Warstadt et al. (2019), this paper focuses on the binary acceptability classification approach for the Russian language.

Applications of Acceptability Acceptability judgments have been broadly applied in NLP. In particular, they are used to test LMs' robustness (Yin et al., 2020) and probe their acquisition of grammatical phenomena (Warstadt and Bowman, 2019; Choshen et al., 2022; Zhang et al., 2021). LA has also stimulated the development of acceptability measures based on pseudo-perplexity (Lau et al., 2020), which correlate well with human judgments (Lau et al., 2017) and show benefits in scoring generated hypotheses in downstream tasks (Salazar et al., 2020). Another application includes evaluating the grammatical and semantic correctness in language generation (Kane et al., 2020; Harkous et al., 2020; Bakshi et al., 2021; Batra et al., 2021).

[2]Both RuCoLA and the code of our experiments are available at github.com/RussianNLP/RuCoLA

Source	Size	%	Content
rusgram	563	49.7	Corpus grammar
Testelets (2001)	1335	73.9	General syntax
Lutikova (2010)	193	75.6	Syntactic structures
Mitrenina et al. (2017)	54	57.4	Generative grammar
Paducheva (2010)	1308	84.3	Semantics of tense
Paducheva (2004)	1374	90.8	Lexical semantics
Paducheva (2013)	1462	89.5	Aspects of negation
Seliverstova (2004)	2104	80.8	Semantics
Shavrina et al. (2020)	1444	36.6	Grammar exam tasks
In-domain	**9837**	**74.5**	
Machine Translation	1286	72.8	English translations
Paraphrase Generation	2322	59.9	Automatic paraphrases
Out-of-domain	**3608**	**64.6**	
Total	**13445**	**71.8**	

Table 2: RuCoLA statistics by source. The number of in-domain sentences is similar to that of CoLA and ItaCoLA. %=Percentage of acceptable sentences.

2.2 Evaluation of Text Generation

Machine translation (or MT) is one of the first sub-fields which has established diagnostic evaluation of neural models (Dong et al., 2021). Diagnostic datasets can be constructed by automatic generation of contrastive pairs (Burlot and Yvon, 2017), crowdsourcing annotations of generated sentences (Lau et al., 2014), and native speaker data (Anastasopoulos, 2019). Various phenomena have been analyzed, to name a few: morphology (Burlot et al., 2018), syntactic properties (Sennrich, 2017; Wei et al., 2018), commonsense (He et al., 2020), anaphoric pronouns (Guillou et al., 2018), and cohesion (Bawden et al., 2018).

Recent research has shifted towards overcoming limitations in language generation, such as copying inputs (Liu et al., 2021), distorting facts (Santhanam et al., 2021), and generating hallucinated content (Zhou et al., 2021). Maynez et al. (2020) and Liu et al. (2022) propose datasets on hallucination detection. SCARECROW (Dou et al., 2022) and TGEA (He et al., 2021) focus on taxonomies of text generation errors. Drawing inspiration from these works, we create the machine-generated out-of-domain set to foster text generation evaluation with acceptability.

3 RuCoLA

3.1 Design

RuCoLA consists of in-domain and out-of-domain subsets, as outlined in Table 2. Below, we describe the data collection procedures for each subset.

Label	Set	Category	Sentence	Source
✓	**In-domain**	✗	*Ya obnaruzhil ego lezhaschego odnogo na krovati.* I found him lying in the bed alone.	Testelets (2001)
*	**In-domain**	SYNTAX	*Ivan prileg, chtoby on otdokhnul.* Ivan laid down in order that he has a rest.	Testelets (2001)
✓	**Out-of-domain**	✗	*Ja ne chital ni odnogo iz ego romanov.* I have not read any of his novels.	Artetxe and Schwenk (2019)
*	**Out-of-domain**	HALLUCINATION	*Ljuk ostanavlivaet udachu ot etogo.* Luke stops luck from doing this.	Schwenk et al. (2021)

Table 3: A sample of RuCoLA. *=Unacceptable sentences. ✓=Acceptable sentences. The examples are translated for illustration purposes.

In-domain Set Here, the data collection method is analogous to CoLA. The in-domain sentences and the corresponding authors' acceptability judgments[3] are drawn from fundamental linguistic textbooks, academic publications, and methodological materials[4]. The works are focused on various linguistic phenomena, including but not limited to general syntax (Testelets, 2001), the syntactic structure of noun phrases (Lutikova, 2010), negation (Paducheva, 2013), predicate ellipsis, and subordinate clauses (`rusgram`[5]). Shavrina et al. (2020) introduce a dataset on the Unified State Exam in the Russian language, which serves as school finals and university entry examinations in Russia. The dataset includes standardized tests on high school curriculum topics made by methodologists. We extract sentences from the tasks on Russian grammar, which require identifying incorrect word derivation and syntactic violations.

Out-of-domain Set The out-of-domain sentences are produced by nine open-source MT and paraphrase generation models using subsets of four datasets from different domains: Tatoeba (Artetxe and Schwenk, 2019), WikiMatrix (Schwenk et al., 2021), TED (Qi et al., 2018), and Yandex Parallel Corpus (Antonova and Misyurev, 2011). We use cross-lingual MT models released as a part of the `EasyNMT` library[6]: OPUS-MT (Tiedemann and Thottingal, 2020), M-BART50 (Tang et al., 2020) and M2M-100 (Fan et al., 2021) of 418M and 1.2B parameters. Russian WikiMatrix sentences are paraphrased via the `russian-paraphrasers` library (Fenogenova, 2021) with the following models and nucleus sampling strategy: ruGPT2-Large[7] (760M), ruT5 (244M)[8], and mT5 (Xue et al., 2021) of Small (300M), Base (580M) and Large (1.2B) versions. The annotation procedure of the generated sentences is documented in §3.3.

3.2 Violation Categories

Each unacceptable sentence is additionally labeled with one of the four violation categories: morphology, syntax, semantics, and hallucinations. The annotation for the in-domain set is obtained through manual working with the sources. The categories are manually defined based on the interpretation of examples provided by the experts, topics covered by chapters, and the general content of a linguistic source. The out-of-domain sentences are annotated as described in §3.3.

Phenomena The phenomena covered by RuCoLA are well represented in Russian theoretical and corpus linguistics and peculiar to modern generative models. We briefly summarize our informal categorization and list examples of the phenomena below:

1. SYNTAX: agreement violations, corruption of word order, misconstruction of syntactic clauses and phrases, incorrect use of appositions, violations of verb transitivity or argument structure, ellipsis, missing grammatical constituencies or words.
2. MORPHOLOGY: incorrect derivation or word building, non-existent words.
3. SEMANTICS: incorrect use of negation, violation of the verbs semantic argument structure.

[3] We keep unacceptable sentences marked with the "*", "*?" and "??" labels.

[4] The choice is also based on the ease of manual example collection, e.g., high digital quality of the sources and no need for manual transcription.

[5] A collection of materials written by linguists for a corpus-based description of Russian grammar. Available at: `rusgram.ru`

[6] `github.com/UKPLab/EasyNMT`

[7] `hf.co/sberbank-ai/rugpt2large`

[8] `hf.co/cointegrated/rut5-base-paraphraser`

4. HALLUCINATION: text degeneration, nonsensical sentences, irrelevant repetitions, decoding confusions, incomplete translations, hallucinated content.

Table 3 provides a sample of several RuCoLA sentences, and examples for each violation category can be found in Appendix A.

3.3 Annotation of Machine-Generated Sentences

The machine-generated sentences undergo a two-stage annotation procedure on Toloka (Pavlichenko et al., 2021), a crowdsourcing platform for data labeling[9]. Each stage includes an unpaid training phase with explanations, control tasks for tracking annotation quality[10], and the main annotation task. Before starting, the worker is given detailed instructions describing the task, explaining the labels, and showing plenty of examples. The instruction is available at any time during both the training and main annotation phases. To get access to the main phase, the worker should first complete the training phase by labeling more than 70% of its examples correctly (Nangia and Bowman, 2019). Each trained worker receives a page with five sentences, one of which is a control one.

We collect the majority vote labels via a dynamic overlap[11] from three to five workers after filtering them by response time and performance on control tasks. Appendix B.2 contains a detailed description of the annotation protocol, including response statistics and the agreement rates.

Stage 1: Acceptability Judgments The first annotation stage defines whether a given sentence is acceptable or not. Access to the project is granted to workers certified as native speakers of Russian by Toloka and ranked top-60% workers according to the Toloka rating system. Each worker answers 30 examples in the training phase. Each training example is accompanied by an explanation that appears in an incorrect answer. The main annotation phase counts 3.6k machine-generated sentences. The pay rate is on average $2.55/hr, which is twice the amount of the hourly minimum wage in Russia. Each of 1.3k trained workers get paid, but we keep votes from only 960 workers whose annotation quality rate on the control sentences is more than 50%. We provide a shortened translated instruction and an example of the web interface in Table 6 (see Appendix B.1).

Stage 2: Violation Categories The second stage includes validation and annotation of sentences labeled unacceptable on **Stage 1** according to five answer options: "Morphology", "Syntax", "Semantics", "Hallucinations" and "Other". The task is framed as a multi-label classification, i.e., the sentence may contain more than one violation in some rare cases or be re-labeled as acceptable. We create a team of 30 annotators who are undergraduate BA and MA in philology and linguistics from several Russian universities. The students are asked to study the works on CoLA (Warstadt et al., 2019), TGEA (He et al., 2021), and hallucinations (Zhou et al., 2021). We also hold an online seminar to discuss the works and clarify the task specifics. Each student undergoes platform-based training on 15 examples before moving onto the main phase of 1.3k sentences. The students are paid on average $5.42/hr and are eligible to get credits for an academic course or an internship. Similar to one of the data collection protocols by Parrish et al. (2021), this stage provides direct interaction between authors and students in a group chat. We keep submissions with more than 30 seconds of response time per page and collect the majority vote labels for each answer independently. Sentences having more than one violation category or labeled as "Other" by the majority are filtered out. The shortened instruction is presented in Table 7 (see Appendix B.1).

3.4 General Statistics

Length and Frequency The sentences in RuCoLA are filtered by the 4–30 token range with `razdel`[12], a rule-based Russian tokenizer. There are 11 tokens in each sentence on average. We estimate the number of high-frequency tokens in each sentence according to the Russian National Corpus (RNC)[13] to control the word frequency distribution. It is computed as the number of frequently used tokens (i.e., the number of instances per million in RNC is higher than 1) divided by the number of tokens in a sentence. We use a moder-

[9] toloka.ai

[10] Control tasks are used on Toloka as common practice for discarding results from bots or workers whose quality on these tasks is unsatisfactory. In our annotation projects, the tasks are manually selected or annotated by a few authors: about 200 and 500 sentences for Stages 1 and 2, respectively.

[11] toloka.ai/docs/dynamic-overlap

[12] github.com/natasha/razdel

[13] ruscorpora.ru/new/en

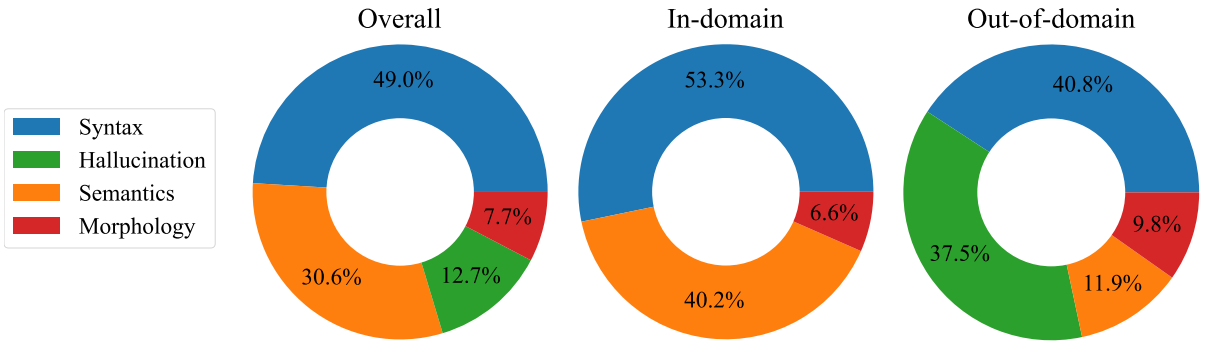

Figure 1: Distribution of violation categories in RuCoLA's unacceptable sentences.

ate frequency threshold $t \geqslant 0.6$ to keep sentences containing rare token units typical for some violations: non-existent or misderived words, incomplete translations, and others. The sentences contain on average 92% of high-frequency tokens.

Category Distribution Figure 1 shows the distribution of violation categories in RuCoLA. Syntactic violations are the most common in RuCoLA (53.3% and 40.8% in the in-domain and out-of-domain sets). The in-domain set includes 40.2% of semantic and 6.6% of morphological violations, while the out-of-domain set accounts for 11.9% and 9.8%, respectively. Model hallucinations make up a percentage of 12.7% of the total number of unacceptable sentences.

Splits The in-domain set of RuCoLA is split into train, validation and private test splits in the standard 80/10/10 ratio (7.9k/1k/1k examples). The out-of-domain set is divided into validation and private test splits in a 50/50 ratio (1.8k/1.8k examples). Each split is balanced by the number of examples per target class, the source type, and the violation category.

4 Experiments

We evaluate several methods for acceptability classification ranging from simple non-neural approaches to state-of-the-art cross-lingual models.

4.1 Performance Metrics

Following Warstadt et al. (2019), the performance is measured by the accuracy score (Acc.) and Matthews Correlation Coefficient (MCC, Matthews, 1975). MCC on the validation set is used as the target metric for hyperparameter tuning and early stopping. We report the results averaged over ten restarts from different random seeds.

4.2 Models

Non-neural Models We use two models from the `scikit-learn` library (Pedregosa et al., 2011) as simple non-neural baselines: a majority vote classifier, and a logistic regression classifier over tf-idf (Salton and Yang, 1973) features computed on word n-grams with the n-gram range $\in [1; 3]$, which results in a total of 2509 features. For the linear model, we tune the ℓ_2 regularization coefficient $C \in \{0.01, 0.1, 1.0\}$ based on the validation set performance.

Acceptability Measures Probabilistic measures allow evaluating the acceptability of a sentence while taking its length and lexical frequency into account (Lau et al., 2020). There exist several different acceptability measures, such as PenLP, MeanLP, NormLP, and SLOR (Lau et al., 2020); we use PenLP due to its results in our preliminary experiments. We obtain the PenLP measure for each sentence by computing its log-probability (computed as a sum of token log-probabilities) from the ruGPT3-medium[14] model. PenLP normalizes the log-probability of a sentence $P(s)$ by the sentence length with a scaling factor α:

$$\text{PenLP}(s) = \frac{P(s)}{((5 + |s|)(5 + 1))^{\alpha}}. \quad (1)$$

After we compute the PenLP value of the sentence, we can predict its acceptability by comparing it with a specified threshold. To find this threshold, we run 10-fold cross-validation on the train set: for each fold, we get the candidate thresholds on 90% of the data by taking 100 points that evenly split the range between the minimum and maximum PenLP values. After that, we get the best threshold per fold by evaluating each threshold on the remaining 10% of the training

[14] hf.co/sberbank-ai/rugpt3medium

Baseline	Overall		In-domain		Out-of-domain	
	Acc.	MCC	Acc.	MCC	Acc.	MCC
Non-neural models						
Majority	68.05 ± 0.0	0.0 ± 0.0	74.42 ± 0.0	0.0 ± 0.0	64.58 ± 0.0	0.0 ± 0.0
Linear	67.34 ± 0.0	0.04 ± 0.0	75.53 ± 0.0	0.17 ± 0.0	62.86 ± 0.0	-0.02 ± 0.0
Acceptability measures from LMs						
ruGPT-3	55.79 ± 0.0	0.27 ± 0.0	59.39 ± 0.0	0.19 ± 0.0	53.82 ± 0.0	0.30 ± 0.0
Russian language models						
ruBERT	75.9 ± 0.42	0.42 ± 0.01	78.82 ± 0.57	0.4 ± 0.01	74.3 ± 0.71	0.42 ± 0.01
ruRoBERTa	80.8 ± 0.47	0.54 ± 0.01	83.48 ± 0.45	0.53 ± 0.01	79.34 ± 0.57	0.53 ± 0.01
ruT5	71.26 ± 1.31	0.27 ± 0.03	76.49 ± 1.54	0.33 ± 0.03	68.41 ± 1.55	0.25 ± 0.04
Cross-lingual models						
XLM-R	65.73 ± 2.33	0.17 ± 0.04	74.17 ± 1.75	0.22 ± 0.03	61.13 ± 2.9	0.13 ± 0.05
RemBERT	76.21 ± 0.33	0.44 ± 0.01	78.32 ± 0.75	0.4 ± 0.02	75.06 ± 0.55	0.44 ± 0.01
Human	**84.08**	**0.63**	**83.55**	**0.57**	**84.59**	**0.67**

Table 4: Results for acceptability classification on the RuCoLA test set. The best score is in bold, the second best one is underlined.

data. Finally, we obtain the best threshold across folds by computing the MCC metric for each of them on the validation set. Figure 3 in Appendix D shows the distribution of scores for acceptable and unacceptable sentences, as well as the best PenLP threshold found in our experiments.

Finetuned Transformer Models We use a broad range of monolingual and cross-lingual Transformer-based language models as our baselines. The monolingual LMs are ruBERT-base[15] (178M trainable parameters), ruRoBERTa-large[16] (355M weights, available only in the large version), and ruT5-base[17] (222M parameters). The cross-lingual models are XLM-R-base (Conneau et al., 2020); 278M parameters) and RemBERT (Chung et al., 2020; 575M parameters). All model implementations, as well as the base code for finetuning and evaluation, are taken from the Transformers library (Wolf et al., 2020). Running all experiments took approximately 126 hours on a single A100 80GB GPU.

All models except ruT5 are finetuned for 5 epochs with early stopping based on the validation set performance on each epoch. We optimize the hyperparameters of these models by running the grid search over the batch sizes $\{32, 64\}$, the learning rates $\{10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$ and the weight decay values $\{10^{-4}, 10^{-2}, 0.1\}$. We fine-tune ruT5 for 20 epochs (also using early stopping) with the batch size of 128; the search space is $\{10^{-4}, 10^{-3}\}$ for the learning rate and $\{0, 10^{-4}\}$ for the weight decay respectively.

The classification task for ruT5 is framed as a sequence-to-sequence problem: we encode the "acceptable" label as "yes" and the "unacceptable" one as "no". The model takes the sentence as its input and generates the corresponding label. We interpret all strings that are not equal to "yes" or "no" as predictions of the "unacceptable" class.

4.3 Human Evaluation

We conduct a human evaluation on the entire in-domain test set and 50% of the out-of-domain test set. The pay rate is on average $6.3/hr, and the task design is similar to **Stage 1** in §3.3 (see also Table 6, Appendix B.1) with a few exceptions. In particular, (i) we remove the "Not confident" answer option, (ii) the annotators are 16 undergraduate BA and MA students in philology and linguistics from Russian universities, and (iii) the votes are aggregated using the method by Dawid and Skene (1979), which is available directly from the Toloka interface. The average quality rate on the control tasks exceeds 75%.

[15] hf.co/sberbank-ai/ruBert-base
[16] hf.co/sberbank-ai/ruRoberta-Large
[17] hf.co/sberbank-ai/ruT5-base

5 Results and Analysis

Table 4 outlines the acceptability classification results. Overall, we find that the best-performing ruRoBERTa model still falls short compared to humans and that different model classes have different cross-domain generalization abilities. Below, we discuss our findings in detail.

5.1 Acceptability Classification

ruRoBERTa achieves the best overall performance among the trained methods, which is nine points behind the human baseline in terms of overall MCC score. The second-best model is RemBERT, followed by ruBERT, with scores of 10% and 12% below ruRoBERTa, respectively. ruT5 and ruGPT-3 + PenLP perform similarly in terms of MCC, although the accuracy of ruT5 is significantly higher. XLM-R achieves the worst performance among finetuned neural models, and the majority vote and logistic regression classifiers have near-zero MCC scores.

We observe that the best models perform similarly on the in-domain and out-of-domain sets with an absolute difference of 0 to 0.04 in terms of MCC. However, the performance gap for other LMs is more prominent. RuT5, XLM-R, and the logistic regression drop by approximately 10 points, whereas the ruGPT-3 + PenLP performance increases. RuT5 and XLM-R have fewer parameters than RuRoBERTa and RemBERT, and smaller models tend to rely more on surface-level cues that poorly transfer to a different domain (Niven and Kao, 2019). The increase in quality for out-of-domain set for PenLP is due to ruGPT-3 assigning consistently lower probabilities to generated sentences. Thus, the PenLP values are skewed to the left for unacceptable sentences (see Figure 3 in Appendix D).

The human performance is higher on the out-of-domain dataset, which can be attributed to the "unnaturalness" of machine-specific features, e.g., hallucinations, nonsense, and repetitions (Holtzman et al., 2020; Meister et al., 2022). The presence of such generated text properties may directly indicate the unacceptability of a sentence.

Finally, we observe that the monolingual models tend to outperform or perform on par with the cross-lingual ones. We attribute this to the size of pre-training data in Russian, which can be five times larger (for ruRoBERTa compared to XLM-R). The size and the quality of the pre-training corpora may directly affect learning the language properties. It might be possible to test this hypothesis by comparing a series of monolingual and cross-lingual LMs pre-trained on the datasets of varying sizes and studying how scaling the data impacts the acquisition of grammatical phenomena (Zhang et al., 2021); however, such a study is beyond the scope of our work.

5.2 Error Analysis

To understand the similarities and differences between classification patterns of models and humans, we conduct an error analysis of all evaluated methods on the in-domain test set. Specifically, we study the proportion of incorrectly classified examples in each group (acceptable sentences and three violation categories).

The main quantitative results of our analysis are shown in Table 5. Our manual study of 250 ex-

Method	Acceptable	Hallucination	Morphology	Semantics	Syntax
Non-neural models					
Majority	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Linear	96.5 ± 0.0	3.3 ± 0.0	3.8 ± 0.0	3.4 ± 0.0	7.4 ± 0.0
Acceptability measures from LMs					
ruGPT-3	36.5 ± 0.0	77.4 ± 0.0	68.8 ± 0.0	63.1 ± 0.0	77.6 ± 0.0
Russian language models					
ruBERT	87.7 ± 1.9	62.6 ± 5.0	30.8 ± 3.5	33.8 ± 2.8	55.2 ± 3.6
ruRoBERTa	91.5 ± 1.2	63.4 ± 4.5	44.4 ± 4.0	37.1 ± 3.1	66.8 ± 2.9
ruT5	89.9 ± 3.6	35.4 ± 6.5	17.0 ± 4.1	20.6 ± 6.2	37.0 ± 4.6
Cross-lingual models					
XLM-R	79.9 ± 6.2	39.7 ± 9.6	29.4 ± 11.4	17.0 ± 4.6	42.9 ± 7.1
RemBERT	85.6 ± 1.7	64.6 ± 4.1	37.8 ± 3.2	35.3 ± 4.2	64.4 ± 2.8
Human	87.7 ± 0.0	84.5 ± 0.0	81.5 ± 0.0	57.1 ± 0.0	80.1 ± 0.0

Table 5: Per-category recall on the RuCoLA test set.

amples misclassified by all methods reveals that sentences with non-specific indefinite pronouns, adverbials, existential constructions, and phrases with possessive prepositions are the most challenging for models and human annotators. We also find that monolingual and cross-lingual LMs tend to judge sentences with the ungrammatical agreement and government as acceptable (e.g.,**Kakim vami viditsja buduschee strany?* "How do you see your see the future of the country?"). Humans make mistakes in long sentences with comparative and subordinate clauses and prepositional government. Another observation is that LMs are not sensitive to morphological violations, such as misderived comparative forms (**Oni v'ehali v bor, i zvuk kopyt stal zvonchee.* "They drove into the forest, and the sound of hooves became louderer."), ungrammatical word-building patterns, and declension of numerals. Finally, most acceptability classifiers achieve high recall on hallucinated sentences, which confirms a practical application potential for classifiers trained on RuCoLA.

5.3 Effect of Length

We analyze the effect of sentence length on the acceptability classification performance by dividing the test set into five length groups of equal size. The results are displayed in Figure 2. The general trend is that the behavior of performance is consistent across all methods. However, while the model performance is unstable and slightly degrades as the length increases, the human annotators outperform the language models on all example groups. Overall, our results are consistent with the findings of Warstadt and Bowman (2019).

To discover the reason behind the increase in quality of automatic methods for sentences of 13–17 tokens, we manually studied a subset of 50 sentences misclassified by ruRoBERTa for each bucket, which amounts to 250 examples in total. We observed that the domain distribution and the error type distribution vary between the length quintile groups, which could explain the differences in model performance for these groups. Specifically, the third group (10–12 tokens) contains sentences with ungrammatical agreement and government or violated argument structure, which are difficult for the models (see Section 5.2). In turn, the fourth quintile interval (13–17 tokens) has more out-of-domain examples of hallucinations, which are easier to detect both for humans and ML-based methods.

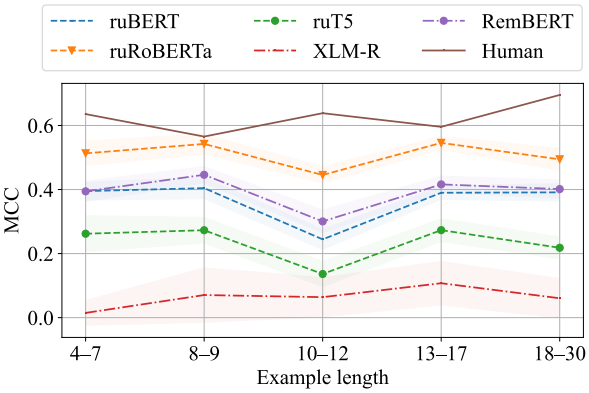

Figure 2: Results on the RuCoLA test set grouped by five quintiles of the sentence length.

6 Cross-lingual Transfer

Given the availability of acceptability classification corpora in other languages, one might be curious about the possibility of knowledge transfer between languages for this task. This is particularly important in the case of estimating sentence acceptability for low-resource languages, which are an important focus area of NLP research (Hedderich et al., 2021). However, the nature of the task makes successful transfer an open question: for instance, specific grammar violations in one language might not exist in another.

With this in mind, we explore the zero-shot cross-lingual transfer scenario, in which the training and validation datasets are provided in one language and the test data in a different one. We use four multilingual models: mBERT (Devlin et al., 2019), XLM-R_{Base}, XML-R, and RemBERT. We study the transfer between three datasets: CoLA, ItaCoLA, and RuCoLA, containing examples in English, Italian and Russian, respectively. As shown in Table 1, all datasets have similar sizes.

Due to the space constraints, we defer a detailed description of the experimental setup and results to Appendix E; here, we overview the key findings of this study. Specifically, we find that the monolingual scenarios outperform cross-lingual transfer by a large margin, which confirms and extends the results of Trotta et al. (2021). Also, we observe that RemBERT performs best in monolingual and cross-lingual setups. For the in-domain set, we observe a cross-lingual transfer gap: there is little difference in language to transfer from, and for RuCoLA, the zero-shot results are as poor as those of a linear "monolingual" classifier. However, the cross-lingual setup performs on par with the monolingual setup for out-of-domain data.

7 Conclusion and Future Work

This work introduces RuCoLA, the first large-scale acceptability classification corpus in the Russian language. The corpus consists of more than 13.4k sentences with binary acceptability judgments and provides a coarse-grained annotation of four violation categories for 3.7k unacceptable sentences. RuCoLA covers two types of data sources: linguistic literature and sentences produced by generative models. Our design encourages NLP practitioners to explore a wide range of potential applications, such as benchmarking, diagnostic interpretation of LMs, and evaluation of language generation models. We conduct extensive experimental evaluation by training baselines that cover a broad range of models. Our results show that LMs fall behind humans by a large margin. Finally, we explore the cross-lingual generalization capabilities of four cross-lingual Transformer LMs across three languages for acceptability classification. The preliminary results show that zero-shot transfer for in-domain examples is hardly possible, but the discrepancy between monolingual and cross-lingual training results for out-of-domain sentences is less evident.

In our future work, we plan to explore the benefits and limitations of RuCoLA in the context of applying acceptability classifiers to natural language generation tasks. Another direction is to augment the in-domain and out-of-domain validation sets with fine-grained linguistic annotation for nuanced and systematic model evaluation. In the long run, we hope to provide valuable insights into the process of grammar acquisition by language models and help foster the application scope of linguistic acceptability.

8 Limitations

Data Collection Acceptability judgments datasets require a source of unacceptable sentences. Collecting judgments from linguistic literature has become a standard practice replicated in multiple languages. However, this approach has several limitations. First, many studies raise concerns about the reliability and reproducibility of acceptability judgments (e.g., Gibson and Fedorenko, 2013; Culicover and Jackendoff, 2010; Sprouse and Almeida, 2013; Linzen and Oseki, 2018). Second, the linguists' judgments may limit data representativeness, as they may not reflect the errors that speakers tend to produce (Dąbrowska, 2010). Third, enriching acceptability judgments datasets is time-consuming, while creating new ones can be challenging due to limited resources, e.g., in low-resource languages.

Expert vs. Non-expert One of the open methodological questions on acceptability judgments is whether they should be collected from expert or non-expert speakers. On the one hand, prior linguistic knowledge can introduce bias in reporting judgments (Gibson and Fedorenko, 2010). On the other hand, expertise may increase the quality of the linguists' judgments over the ones of non-linguists (see a discussion by Schütze and Sprouse, 2013). At the same time, the latter tend to be influenced by an individual's exposure to ungrammatical language use (Dąbrowska, 2010). Recall that our in-domain examples and their acceptability labels are manually drawn from linguistic literature, while the out-of-domain set undergoes two stages (§3.3):

1. **Stage 1: Acceptability Judgments** — collecting acceptability labels from non-expert speakers;
2. **Stage 2: Violation Categories** — validation of the acceptability labels from **Stage 1** and fine-grained example annotation by their violation category by expert speakers.

The objective of involving students with a linguistic background is to maximize the annotation quality. We follow Warstadt et al. (2019) and report the students' evaluation results as the human baseline in this paper. Human evaluation through crowdsourcing (Nangia and Bowman, 2019) is left for future work.

Fine-grained Annotation The coarse-grained annotation scheme of the RuCoLA's unacceptable sentences relies on four major categories. While the annotation can be helpful for model error analysis, it limits the scope of LMs' diagnostic evaluation concerning linguistic and machine-specific phenomena (Warstadt and Bowman, 2019).

Distribution Shifts Many studies have discussed the role of lexical frequency in acceptability judgments (Myers, 2017). In particular, LMs can treat frequent patterns from their pretraining corpora as acceptable and perform poorly on rare or unattested sentences with low probabilities (Marvin and Linzen, 2018; Park et al., 2021;

Linzen and Baroni, 2021). Although we aim to control the number of high-frequency tokens in the RuCoLA's sentences (§3.4), we assume that potential word frequency distribution shift between LMs' pre-training corpora and our corpus can introduce bias in the evaluation. Furthermore, linguistic publications represent a specific domain as the primary source of acceptability judgments. On the one hand, it can lead to a domain shift when using RuCoLA for practical purposes. On the other hand, we observe moderate acceptability classification performance on the out-of-domain test, which spans multiple domains, ranging from subtitles to Wikipedia.

9 Ethical Considerations

Responses of human annotators are collected and stored anonymously. The average annotation pay rate exceeds the hourly minimum wage in Russia twice or four times, depending on the annotation project. The annotators are warned about potentially sensitive topics in data (e.g., politics, culture, and religion).

RuCoLA may serve as training data for acceptability classifiers, which may benefit the quality of generated texts (Batra et al., 2021). We recognize that such improvements in text generation may lead to misuse of LMs for malicious purposes (Weidinger et al., 2021). However, our corpus can be used to train adversarial defense and artificial text detection models. This paper introduces a novel dataset for **research and development needs**, and the potential negative uses are not lost on us.

10 Acknowledgements

Alena Pestova was supported by the framework of the HSE University Basic Research Program. Max Ryabinin was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139. The data annotation effort was supported by the Toloka Research Grants program.

References

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Alexandra Antonova and Alexey Misyurev. 2011. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Shreyan Bakshi, Soumya Batra, Peyman Heidari, Ankit Arun, Shashank Jain, and Michael White. 2021. Structure-to-text generation with self-training, acceptability classifiers and context-conditioning for the GEM shared task. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 136–147, Online. Association for Computational Linguistics.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, and Roberto Zamparelli. 2020. AcCompl-it@ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian. In *EVALITA*.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third*

Conference on Machine Translation: Shared Task Papers, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Peter W Culicover and Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to gibson and fedorenko. *Trends in Cognitive Sciences*, 6(14):234–235.

Ewa Dąbrowska. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2021. A Survey of Natural Language Generation. *arXiv preprint arXiv:2112.11739*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Ingo Feldhausen and Sebastian Buchczyk. 2020. Testing the Reliability of Acceptability Judgments for Subjunctive Obviation in French. In *Going romance 2020*.

Alena Fenogenova. 2021. Russian paraphrasers: Paraphrase with transformers. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 11–19, Kiyv, Ukraine. Association for Computational Linguistics.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Edward A Gibson and Evelina G Fedorenko. 2010. Weak quantitative standards in linguistics research.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6012–6025, Online. Association for Computational Linguistics.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring Gradience in Speakers Grammaticality Judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A probabilistic View of Linguistic Knowledge. *Cognitive science*, 41(5):1202–1241.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annu. Rev. Linguist*, 7:2–1.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Ekaterina Lutikova. 2010. K voprosu o kategorial'nom statuse imennykh grup v russkom yazyke. *Moscow University Philology Bulletin*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Brian W. Matthews. 1975. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et biophysica acta*, 405 2:442–51.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation.

Olga Mitrenina, Evgeniya Romanova, and Natalia Slioussar. 2017. *Vvedeniye v generativnuyu grammatiku*. Limited Liability Company "LIBROCOM".

James Myers. 2017. Acceptability Judgments. In *Oxford Research Encyclopedia of Linguistics*.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Elena Paducheva. 2004. *Dinamicheskiye modeli v semantike leksiki*. Languages of Slavonic culture.

Elena Paducheva. 2010. *Semanticheskiye issledovaniya: Semantika vremeni i vida v russkom yazyke*, second edition. Languages of Slavonic culture.

Elena Paducheva. 2013. *Russkoye otritsatel'noye predlozheniye*. Languages of Slavonic culture.

Kwonsik Park, Myung-Kwan Park, and Sanghoun Song. 2021. Deep learning can contrast the minimal pairs of syntactic data. *Linguistic Research*, 38(2):395–424.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *the Journal of machine Learning research*, 12:2825–2830.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Gerard Salton and C. S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was Built in 1776: a Case Study on Factual Correctness in Knowledge-Grounded Response Generation. *arXiv preprint arXiv:2110.05456*.

Barbara C. Scholz, Francis Jeffry Pelletier, and Geoffrey K. Pullum. 2021. Philosophy of Linguistics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Carson T Schütze and Jon Sprouse. 2013. Judgment data. *Research methods in linguistics*, pages 27–50.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Olga Seliverstova. 2004. *Trudy po semantike*. Languages of Slavonic culture.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Tatiana Shavrina, Anton Emelyanov, Alena Fenogenova, Vadim Fomin, Vladislav Mikhailov, Andrey Evlampiev, Valentin Malykh, Vladimir Larin, Alex Natekin, Aleksandr Vatulin, Peter Romov, Daniil Anastasiev, Nikolai Zinov, and Andrey Chertok. 2020. Humans keep it one hundred: an overview of AI journey. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2276–2284, Marseille, France. European Language Resources Association.

Jon Sprouse. 2018. Acceptability Judgments and Grammaticality, Prospects and Challenges. In *Syntactic Structures after 60 Years*, pages 195–224. De Gruyter Mouton.

Jon Sprouse and Diogo Almeida. 2013. The empirical status of data in syntax: A reply to gibson and fedorenko. *Language and Cognitive Processes*, 28(3):222–228.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless Green Ideas do Sleep Furiously: Gradient Acceptability and the Nature of the Grammar. *The Linguistic Review*, 35(3):575–599.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.

Yakov Testelets. 2001. *Vvedeniye v obschiy sintaksis*. Russian State University for the Humanities.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in neural information processing systems*, pages 5998–6008.

Héctor Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments. *arXiv preprint arXiv:1901.03438*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Johnny Wei, Khiem Pham, Brendan O'Connor, and Brian Dillon. 2018. Evaluating grammaticality in seq2seq models with a broad coverage HPSG grammar: A case study on machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 298–305, Brussels, Belgium. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv preprint arXiv:2112.04359*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. 2020. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and

Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Examples

This appendix provides approximately 50 examples of the RuCoLA's sentences appearing in the in-domain and out-of-domain sets and corresponding fine-grained phenomena.

A.1 In-domain Set

Morphology

(1) **Word derivation**

a. Comparatives
   (i) *Litso dlin'she, chem nado, a telo koroche. ("The face is longer than it should be, and the body is shorter.")
   (ii) *A sapogi ja vam blestjashhie prinesu, eshhjo krasivshe vashih! ("And I'll bring you shiny boots, even more beautiful than yours!")

b. Word-building patterns
   (i) *Ljudej rugajut ili hvaljat za ihnie dela, a ne za nacional'nost'. ("People are scolded or praised for their deeds, not for their nationality.")
   (ii) *Vdobavok, v koridore bylo holodno, i mal'chik sovsem ozjabnul. ("In addition, it was cold in the corridor, and the boy was completely chilled.")
   (iii) *Zarubezhnym kollegam predlozhili propoverjat' rezul'taty. ("Foreign colleagues were invited to check the results.")

c. Declension of numerals
   (i) *Delo sostoit v tom, chto "Nezhnyj vestnik" rashoditsja v vos'm'justah kopijah ezhenedel'no, po tridcati kopeek.
   (ii) *My darim podarok kazhdyj pjat'sotyj zakaz, opredeljaemyj po nomeru nakladnoj. ("We give a gift every five hundredth order, determined by the invoice number.")

Syntax

(2) **Copular constructions**

a. Vse utro on byl razdrazhitelen. ("He had been irritable all morning.")
b. *U nas sejchas est' dozhd'. ("It is raining now.")

(3) **Word order**

a. Subordinate clauses
   (i) *Den' goroda, kotorom ja zhivu v. ("The day of the city I live in.")
   (ii) Ja prines dokumenty, chtoby mne ne byt' na sude goloslovnym. ("I brought the documents so that I wouldn't be unfounded at the trial.")

b. Coordinate clauses and constructions
   (i) *I on podnjal trubku, ja pozvonil Vane. ("And he picked up the phone, I called Vanya.")
   (ii) *Ona to stihi chitaet, kartiny to pokazyvaet. ("She either reads poetry, shows or pictures.")

(4) **Agreement**

a. Number agreement
   (i) *Devochki, davaj zajdem v magazin! ("Girls, let's go to the store!")
   (ii) *Te, kto nazyvajut sebja patriotami, dolzhen horosho znat' rodnoj jazyk. ("Those who call themselves patriots should know their native language well.")

b. Case agreement
   (i) *Ego ot'ezd za granicu vsem vosprinimalsja kak pobeg. ("His departure abroad was perceived by everyone as an escape.")
   (ii) *Na melkovodnyh uchastkah rastitel'nost' obrazuet peremychki, razdeljajushhimi ozero na otdel'nye pljosy. ("In shallow areas, vegetation forms bridges dividing the lake into separate stretches.")
   (iii) *Nikogo ne bylo kholodno. ("No one was cold.")

(5) **Verb transitivity**

a. Intransitive verbs with prepositional phrases
   (i) *Na kazhdoj dorozhke bezhalo po sportsmenu ("There was an

athlete running on each track.")

(ii) *Po rebenku sdelali sebe buterbrody. ("For the child made themselves sandwiches.")

b. Transitive verbs with impersonal clauses or sentential actants

(i) U Nonny gorelo lico, ee dazhe znobilo ot volnenija. ("Nonna's face was burning, she was even shivering with excitement.")

(ii) *On znaet, chto polk perebrosjat na drugoj uchastok fronta i drugie plany komandovanija. ("He knows that the regiment will be transferred to another sector of the front and other plans of the command.")

(6) **Coordination and subordination**

a. Constructions with subordinate and infinitive clauses

(i) *Nam uzhe bylo izvestno, chto on priekhal i drugie fakty. ("We already knew that he had arrived and other facts.")

(ii) *I chto on byl razbit, bylo zamecheno vsemi. ("And that it was broken, everything was noticed.")

b. Coordinate clauses with dative constructions

(i) *Mne vystupat' sledujushhim i uzhe napomnili ob jetom. ("I will be the next to speak and have already been reminded of this.")

(ii) Mne soobshhili ob jetih planah, i oni ponravilis'. ("I was informed about these plans, and I liked them.")

Semantics

(7) **Non-specific indefinite pronouns**

a. *Khorosho, chto on kupil chto-nibud'. ("It's good that he bought something.")

b. *Kakogo-nibud' reshenija on ne prinjal. ("He didn't make any decision.")

c. *Ja ne ljublju kogo-libo. ("I don't love anyone.")

(8) **Tense**

a. *Zavtra my slyshim operu. ("Tomorrow we hear the opera")

(9) **Aspect**

a. *Zavtra budem ezdit' vo Vneshtorgbank. ("Tomorrow we will go to Vneshtorgbank.")

(10) **Negation or negative concord**

a. *Nikto ego videl? ("Has no one seen him?")

b. *On ne byl tam i razu. ("He hasn't been there once.")

(11) **Existential constructions**

a. *Sushhestvujut zajavlenija ot postradavshikh.("There are statements from victims.")

A.2 Out-of-domain Set

Morphology

(1) **Nonce words**

a. *I ja sygral pervoe dvizhenie betovennogo violetovogo koncerta. ("And I played the first movement of the beethoven violette concerto.")

b. *Rastenie harakterno dlja stepi i sil'vostepi na ravninah i na plato Moldavii na severe. ("The plant is characteristic of the steppe and the silvosteppe on the plains and on the plateau of Moldova in the north.")

c. *Aviakompanijam razreshili ispol'zovat' servis "onechuckle" dlja zakaza samyh populjarnyh aviamar. ("Airlines were allowed to use the "onechuckle" service to order the most popular aviamars.")

d. *Ona risuet horosho i mechtaet stat' hudozhn'ej. ("She draws well and dreams of becoming an artistrone.")

e. *Dlja nihsudarstvennyh organizacij razrabotali metod analiza bjudzhetov dlja razreshenija konfliktov s zhenshhinami. ("A method of budget analysis for resolving conflicts with women has been developed for themgovernmental organizations.")

Syntax

(2) **Agreement**

a. Person

(i) *On ostalsja nevredimoj i v mo-

roze. (“He remained unharmed and in the cold.”)

b. Case

(i) *Vospol’zujtes’ prokat avtomobilej i 24-chasovoj priem, chtoby vy mogli ispytat’ svoj prebyvanie, kak vy hotite. (“Take advantage the car rental and 24-hour reception so you can experience your stay the way you want.”)

(3) **Subordination and coordination**

a. *Jeto to, chto oni prishli k ponimaniju, chto samoe vazhnoe, chemu deti dolzhny nauchit’sja, jeto harakter. (“This is what they have come to understand that the most important thing children need to learn this is character.”)

(4) **Ellipsis**

a. *Bolee 30 uchenyh zashhitili kandidatskie dissertacii pod rukovodstvom. (“More than 30 scientists defended their PhD theses under the supervision of.”)

Hallucinations

(5) **Nonsensical sentences**

a. *Soobshhenie s nej tol’ko po peshke. (“Messaging her is only possible by a pawn.”)

b. *Futbolist “Liverpulja” vpervye podpisal pervyj god porazhenija kolena. (“The Liverpool footballer has signed for the first time in the first year of defeating his knee.”)

c. *I vse po vsej biblioteke raznye predmety, raznye prostranstva. (“And all throughout the library are different objects, different spaces.”)

(6) **Irrelevant repetitions**

a. *Dlja jetoj programmy byli provedeny dva programmy. (“For this program two programs were conduncted.”)

b. *Posylki pojavlenija product placement v kinematografe u brat’ev Ljum’er pojavilis’ uzhe u brat’eva Ljum’era. (“The premises of the appearance of the product placement in the cinema of the Lumiere brothers have already appeared in the Lumiere brothers.”)

Semantics

(7) **Semantics**

a. *Poberezh’ja Ivanova i Kohma proshli na severe. (“The coasts of Ivanovo and Kokhma passed in the north.”)

b. *Prezident zajavil, chto u Rossii dostatochno sil dlja provedenija profilakticheskih zabastovok. (“The President said that Russia has enough forces to carry out preventive strikes.”)

c. *Torgovlja real’nymi den’gami na virtual’nom rynke vyrosla, chtoby stat’ mnogomillionnoj industriej dollarov. (“Real money trading in the virtual market has grown to become a multimillion dollar industry.”)

d. *On vnov’ zavershil nokautom pretendenta v vos’mom raunde. (“He again finished by knocking out the challenger in the eighth round.”)

B Annotation Protocols

B.1 Instructions

Task

- Your task is to define whether a given sentence is appropriate or contains any violations (one would not say or write like this).
- Choose "Yes" if the sentence contains one or more violations.
- Choose "No" if the sentence is appropriate (you would say like this).
- Choose "Not confident" if you have doubts.
- If there are any typos, please state them in the box.

Examples of violations

- Number disagreement: *"Podrobnosti dela neizvestno."* ("No details of this case is available.")
- Semantic collisions: *"V etom godu zhenschiny vyshly zamuzh vo vtoroy raz 26 iyunya 1989 goda."* ("This year the women got married the second time on the 26th of June in 1989.")
- Nonsensical repetitions: *"Eto moya sem'ya moya sem'ya."* ("It is my family my family.")

Annotation examples

- "Yes" (the given sentence contains one or more violations): *"Ya dolzhen poiti s velichiem, chtoby prostit' eyoh.* ("I should go with greatness to forgive her.")
- "No" (the given sentence is acceptable): *"Skol'ko chasov v den' vy rabotaete?"* ("How many hours a day do you work?")

Please check the task before submission.
Thank you!

Example of web interface

Does the sentence contain **violations**?

This is a toy example.

- ○ Yes
- ○ No
- ○ Not confident

If there are any typos, please state them below:

Please check the task once again. Thank you!

Table 6: A shortened version of the instruction given to crowd-sourced annotators for judging the acceptability of machine-generated sentences (**Stage 1: Acceptability Judgments**; §3.3). The instruction is translated for illustration purposes.

Task

- Your task is to select all appropriate violation categories under which a given sentence falls: Morphology, Syntax, Semantics, Hallucinations, or Other.
- Choose "No violations" if the sentence is acceptable.
- Choose "Not confident" if you have doubts.
- If there are any typos, please state them in the box.
- If any questions or doubts, contact us in the chat.

Examples

- Morphology
  - Non-existent words: *"Eto semiduymovyi heturpin."* ("It is a seven-inch heturpin.")
  - Misderivation: *"Ona vyglyadit krasivshe."* ("She looks more beatufiuler".)
- Syntax
  - Agreement violation: *"Oni schitali ego talantlivymi."* ("They considered him to be talented.")
  - Word order: *"Plan Mashe Sashi prodat' kvartiru."* ("Plan Masha's Sasha to sell a flat.")
- Semantics
  - Semantic properties of the predicate: "Ty kogda-nibud' nakhodilsya v Moskve?" ("Have you ever been to Moscow?")
- Hallucinations
  - Incomplete translation or input copying: *"Ya rad, shto you heard o Margaret Thatcher."* ("I am glad you heard about Margaret Thatcher.")
  - Repetitive content: *"Eto moya sem'ya moya sem'ya."* ("It is my family my family.")

Example of web interface

Select all appropriate violation categories.

This is a toy example.

- ☐ Morphology
- ☐ Syntax
- ☐ Semantics
- ☐ Hallucinations
- ☐ Other
- ☐ Not confident
- ☐ No violations

If there are any typos, please state them below:

Please check the task before submission.
Thank you!

Table 7: A shortened version of the instruction given to students for validation and coarse-grained annotation of the unacceptable machine-generated sentences (**Stage 2: Violation Categories**; §3.3). The instruction is translated for illustration purposes.

	# annotators	Pay rate	Average response time, s	Average quality	# training sentences	# control sentences	# sentences
Stage 1	1300	\$2.55/hr	70	80%	28	179	5685
Stage 2	30	\$5.42/hr	143	57%	11	500	2699
Human Benchmark	16	\$6.3/hr	53	79%	10	901	2048

Table 8: Summary of the annotation design details by annotation project.

	Acceptable	Morphology	Syntax	Hallucination	Semantics	Average
Stage 1	0.80	0.83	0.88	0.88	0.78	0.83
Stage 2	0.94	0.90	0.84	0.86	0.90	0.89
Human Benchmark	0.85	0.81	0.82	0.88	0.80	0.85

Table 9: Per-category WAWA inter-annotator agreement rates by annotation project.

B.2 Design Details

This subsection summarizes the annotation design details for each annotation project: **Stage 1: Acceptability Judgments** (Section 3.3); **Stage 2: Violation Categories** (Section 3.3); and **Human Evaluation** (Section 4.3).

Annotation Project Statistics Table 8 describes the following statistics: the number of annotators who participated in the project, the pay rate (\$/hr), the average response time in seconds, the average performance on the control tasks, the number of training and control sentences, and the overall number of sentences.

Inter-annotator Agreement Rates Table 9 presents the per-category IAA rates for each annotation project. The IAA rates are computed with the Worker Agreement with Aggregate (WAWA) coefficient (Ning et al., 2018). WAWA indicates the average fraction of the annotators' responses that agree with the aggregate answer for each example. The WAWA values are above 0.8 in most cases, which implies a strong agreement between annotators. We observe that the non-expert annotators (**Stage 1**) have lower average WAVA values than the expert ones (**Stage 2**; **Human Evaluation**). Annotators in the **Human evaluation** project receive high IAA scores on acceptable sentences and sentences containing hallucinations. Although the IAA scores are lower in the other categories, they are still strong.

C Hyperparameter Values for Baseline Methods

Model	Hyperparameter	Value
Linear	ℓ_2 penalty strength	1
ruGPT-3	Threshold	-20.92

Table 10: Optimal hyperparameter values for the linear model and threshold-based baselines.

Model	Learning rate	Weight decay	Batch size
ruBERT	$3{\cdot}10^{-5}$	0.1	32
ruRoBERTa	10^{-5}	10^{-4}	32
ruT5	10^{-4}	0	128
XLM-R	10^{-5}	0.1	32
RemBERT	10^{-5}	10^{-4}	64

Table 11: Optimal hyperparameter values for finetuned language models.

D Acceptability Measure Distribution

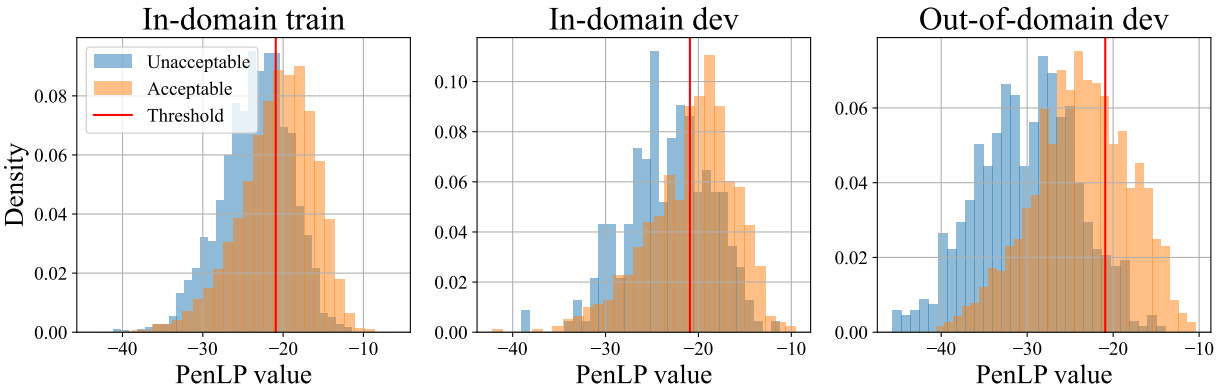

Figure 3: PenLP acceptability measure values for train and validation sets of RuCoLA.

E Cross-lingual Evaluation Details

Here, we describe the setup of experiments outlined in Section 6. We use four models: Multilingual BERT-base-cased (110M parameters), XLM-RoBERTa-base (or XLM-R_{Base}, 270M parameters), XLM-RoBERTa-large (or XLM-R, 550M parameters), and RemBERT. For each pair of **source** and **target** languages, we train and tune the hyperparameters on the train and development sets of the **source** language respectively and compute the final metrics on the **target** language. We also include the pairs consisting of the same language (i.e., the same dataset) for source and target to provide an upper bound for classification quality. We do not report accuracy both for brevity and because the test set leaderboard for CoLA reports only the MCC values.

The results are averaged over ten different random seeds: we use mean MCC on the development set for hyperparameter tuning and report the average and the standard deviation of the test set metrics. We optimize the validation score with grid search with respect to the following hyperparameters: learning rate (the search space is $\{10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$), batch size $\{32, 64\}$, weight decay $\{10^{-4}, 10^{-2}, 0.1\}$. Each model is trained for 5 epochs with early stopping based on the validation MCC.

Table 12 shows the results of our study. The RemBERT model outperforms other cross-lingual encoders, which aligns with the results of Chung et al. (2020). A key observation is that although the quality on the out-of-domain data is indeed similar for all source languages, for the in-domain test set of RuCoLA, the cross-lingual generalization gap remains quite large, similarly to other studied datasets. This indicates that source on other datasets does not induce the language understanding capabilities that are necessary for estimating acceptability in the Russian language, which is expected given its typological differences from the English and Italian languages.

Model	Training data	CoLA		ItaCoLA	RuCoLA	
		In-domain	OOD		In-domain	OOD
mBERT	CoLA	**0.41 ± 0.04**	**0.35 ± 0.03**	0.03 ± 0.03	-0.04 ± 0.02	0.07 ± 0.02
	ItaCoLA	0.07 ± 0.05	0.09 ± 0.05	**0.36 ± 0.04**	0.0 ± 0.03	0.0 ± 0.02
	RuCoLA	-0.02 ± 0.04	0.01 ± 0.05	0.0 ± 0.02	**0.18 ± 0.02**	**0.15 ± 0.03**
XLM-R_{Base}	CoLA	**0.55 ± 0.03**	**0.51 ± 0.02**	0.2 ± 0.02	0.1 ± 0.01	0.3 ± 0.02
	ItaCoLA	0.05 ± 0.07	0.03 ± 0.06	**0.25 ± 0.08**	0.01 ± 0.02	0.01 ± 0.02
	RuCoLA	0.05 ± 0.08	0.06 ± 0.06	0.01 ± 0.05	**0.23 ± 0.05**	**0.2 ± 0.06**
XLM-R	CoLA	**0.61 ± 0.02**	**0.57 ± 0.03**	0.3 ± 0.02	0.2 ± 0.03	**0.42 ± 0.02**
	ItaCoLA	0.3 ± 0.03	0.32 ± 0.03	**0.52 ± 0.03**	0.13 ± 0.03	0.24 ± 0.02
	RuCoLA	0.24 ± 0.12	0.26 ± 0.13	0.17 ± 0.08	**0.36 ± 0.05**	0.4 ± 0.06
RemBERT	CoLA	**0.65 ± 0.02**	**0.6 ± 0.02**	0.29 ± 0.03	0.17 ± 0.02	**0.44 ± 0.03**
	ItaCoLA	0.48 ± 0.04	0.44 ± 0.04	**0.52 ± 0.02**	0.15 ± 0.03	0.39 ± 0.04
	RuCoLA	0.46 ± 0.03	0.44 ± 0.01	0.29 ± 0.02	**0.41 ± 0.02**	**0.44 ± 0.02**

Table 12: MCC for cross-lingual acceptabiliy classification. The best score is in bold.