RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs

Ekaterina Taktasheva[1*], **Maxim Bazhukov**[2*], **Kirill Koncha**[3,4*†], **Alena Fenogenova**[5], **Ekaterina Artemova**[6], **and Vladislav Mikhailov**[7]
[1]University of Edinburgh, [2]HSE University, [3]University of Groningen, [4]Ghent University, [5]SaluteDevices, [6]Toloka AI, [7]University of Oslo

Abstract

Minimal pairs are a well-established approach to evaluating the grammatical knowledge of language models. However, existing resources for minimal pairs address a limited number of languages and lack diversity of language-specific grammatical phenomena. This paper introduces the **Ru**ssian **B**enchmark of **Li**nguistic **M**inimal **P**airs (RuBLiMP), which includes 45k pairs of sentences that differ in grammaticality and isolate a morphological, syntactic, or semantic phenomenon. In contrast to existing benchmarks of linguistic minimal pairs, RuBLiMP is created by applying linguistic perturbations to automatically annotated sentences from open text corpora and decontaminating test data. We describe the data collection protocol and present the results of evaluating 25 language models in various scenarios. We find that the widely used LMs for Russian are sensitive to morphological and agreement-oriented contrasts, but fall behind humans on phenomena requiring the understanding of structural relations, negation, transitivity, and tense. RuBLiMP, the codebase, and other materials are publicly available.

1 Introduction

Acceptability judgments are the main empirical test in generative linguistics for assessing humans' linguistic competence and language acquisition (Chomsky, 1965; Schütze, 1996). One of the well-established approaches to judging a sentence's acceptability is a forced choice between *minimal pairs* of sentences, where a native speaker is expected to prefer a grammatical sentence to an ungrammatical one, as in Example 1.

(1) a. The cat is on the mat. (grammatical)
b. *The cat are on the mat. (ungrammatical)

	Language	Size	# Paradigm	Method
BLiMP	English	67k	67	Dictionary & templates
CLiMP	Chinese	16k	16	Translation & templates
JBLiMP	Japanese	331	39	Extract from articles
SLING	Chinese	38k	38	UD Treebank & templates
NoCoLA$_{zero}$	Norwegian	99.1k	11	Extract from an L2 corpus
DaLAJ	Swedish	4.8k	4	Extract from an L2 corpus
LINDSEA	Indonesian	380	38	Expert-written min. pairs
	Tamil	200	20	
CLAMS	English	153.5k	13	Translation & templates
	French	49.3k	7	
	German	47.8k	7	
	Hebrew	40.8k	7	
	Russian	40.1k	7	
RuBLiMP	Russian	45k	45	Open text corpora, rules, automatic UD annotation, pretraining data detection

Table 1: Comparison of benchmarks of linguistic minimal pairs for different languages: BLiMP (Warstadt et al., 2020), CLiMP (Xiang et al., 2021), JBLiMP (Someya and Oseki, 2023), SLING (Song et al., 2022), NoCoLA$_{zero}$ (Jentoft and Samuel, 2023), DaLAJ (Volodina et al., 2021), LINDSEA (Leong et al., 2023), CLAMS (Mueller et al., 2020), and RuBLiMP (ours).

The paradigm of minimal pairs has been widely adopted for evaluating the grammatical knowledge of language models (LMs) across various linguistic phenomena (Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Warstadt et al., 2019, 2020). The evaluation design implies that an LM assigns a higher probability to the grammatical sentence than the ungrammatical one if it is sensitive to the isolated phenomenon. Over the last few years, a broad range of LMs has been analyzed via this paradigm in typologically diverse languages, except for Russian (e.g., Hartmann et al., 2021; Pérez-Mayos et al., 2021; Leong et al., 2023).

This paper introduces the **Ru**ssian **B**enchmark of **Li**nguistic **M**inimal **P**airs (RuBLiMP), which consists of 45 datasets, each including 1k minimal pairs. Our benchmark covers morphological, syntactic, and semantic phenomena well-represented in Russian theoretical linguistics. In contrast to existing benchmarks of linguistic minimal pairs (see Table 1), RuBLiMP is created by (i) extracting

[*] Equal contribution.
[†] Work is partially done while at HSE University.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9268–9299
November 12-16, 2024 ©2024 Association for Computational Linguistics

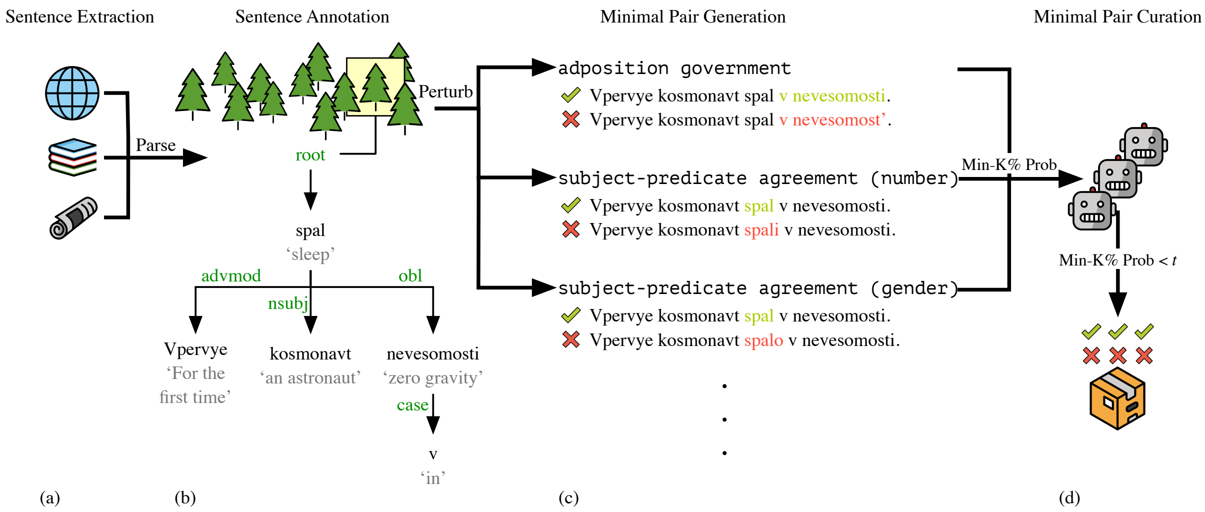


Figure 1: Overview of the RuBLiMP's minimal pair generation approach. Example: *Vpervye kosmonavt spal v nevesomosti* "For the first time an astronaut slept in zero gravity". (a) Extract sentences from publicly available corpora of Wikipedia texts, news articles, and books. (b) Annotate each extracted sentence in the Universal Dependencies scheme (Nivre et al., 2017) with a multidomain morphosyntactic parser for Russian (Anastasyev, 2020). (c) Search the dependency trees for specific lexical units and linguistic structures and apply expert-written perturbation rules to create a pool of minimal pairs for a target paradigm. (d) Compute MIN-K% PROB (Shi et al., 2023) for each grammatical sentence in the pool using a set of LMs. Select t (the threshold for the maximum MIN-K% PROB value), which allows to find an intersection of 1k minimal pairs between the LMs. The minimal pairs in the intersection contain grammatical sentences that are not detected as the LMs' pretraining examples.

sentences from open text corpora across multiple domains, (ii) annotating the sentences with one of the state-of-the-art multidomain morphosyntactic parsers, (iii) creating minimal pairs by perturbing the annotated sentences with expert-written rules, and (iv) discarding the pairs if the grammatical sentence is detected as a pretraining corpus example for at least one of 25 widely used LMs for Russian. Our method allows for generating minimal pairs at scale and ensures high customizability w.r.t. domain, dataset size, and LMs. Validating RuBLiMP by 20 native speakers with a background in linguistics confirms that the generated minimal pairs unambiguously isolate the target phenomenon and contrast in grammaticality.

Our main *contributions* are: (i) we create RuBLiMP, the first diverse and large-scale benchmark of minimal pairs in Russian, (ii) we conduct ablation studies to analyze the effect of pretraining data decontamination on the model performance, (iii) we evaluate 25 monolingual and cross-lingual Transformer LMs (Vaswani et al., 2017) and crowdsourcing workers, (iv) we release RuBLiMP[1], our codebase[2], and all data collection, data annotation, and other materials.

[1] hf.co/datasets/RussianNLP/rublimp
[2] github.com/RussianNLP/RuBLiMP

2 RuBLiMP

Figure 1 outlines our approach to generating minimal pairs for RuBLiMP, which includes the following stages: sentence extraction and annotation (§2.1), minimal pair generation (§2.2) and curation (§2.4). Our framework allows the user to customize each component and provides the foundation to mitigate the limitations of static benchmarks (Bowman and Dahl, 2021) through continuous generation of minimal pairs for a domain of interest and decontaminating the data for specific Russian LMs.

2.1 Corpora Annotation

Sentence Extraction Three open text corpora are used as the source of grammatical sentences: Wikipedia[3], Wikinews[4], and Librusec, a collection of digitalized Russian books (Panchenko et al., 2017). We extract articles from Wikipedia and Wikinews using WikiExtractor (Attardi, 2015) and literary texts from Librusec using `corus`[5]. Next, we segment the documents into sentences and tokenize the sentences with the help of `natasha`[6]. We filter out the sentences based on the number of to-

[3] dumps.wikimedia.org/ruwiki/latest
[4] dumps.wikimedia.org/ruwikinews/latest
[5] github.com/natasha/corus
[6] github.com/natasha/natasha

kens (6-to-50) and shallow heuristics to avoid the sentence segmentation errors.

Sentence Annotation Each extracted sentence is annotated in the Universal Dependencies scheme (Nivre et al., 2017) with a multidomain morphosyntactic parser for Russian (Anastasyev, 2020).

2.2 Minimal Pair Generation

We search the dependency trees for specific lexical units and linguistic structures and edit them using expert-written perturbation rules to create a pool of minimal pairs for a target paradigm (§2.3). Our rules are written by three authors of this paper (native Russian computational linguistics) based on theoretical works on Russian morphology, syntax, and semantics. Each set of rules undergoes a peer-review stage by one of the authors. Below, we provide a general description of the minimal pair generation procedure, which involves four main edit operations: addition, replacement, swapping, and movement. These operations ensure the equal length of the grammatical and ungrammatical sentences. The implementation details and a complete list of the literature are documented in Appendix B.

Morphology Our morphological perturbations violate the principles of the affix order (Greenberg, 1963; Reynolds, 2013) and properties of inflectional classes. We introduce derivational and inflectional errors using `pymorphy2`[7] (Korobov, 2015), morphological dictionaries (Bocharov et al., 2013) available in `pymorphy2`, and word formation dictionaries (Bolshakova and Sapin, 2021).

Syntax Here, we corrupt adpositional and verbal government, negative concord rules, and agreement in number, gender, person, or case (Testelets, 2001). We search for a word from curated lists or with specific morphosyntactic features in relevant syntactic constructions and move it to a different constituent or change its form using `pymorphy2`. We consider various types of the subject (a noun phrase, genitive, and clause) and additional contexts with attractors, which introduce contextual ambiguity in the ungrammatical sentence.

Semantics Our semantic perturbations alter the verb's argument structure and introduce temporal and aspectual violations across the entire sentence.

[7]A rule-based morphological analyzer, which allows for inflecting a word w.r.t. a given set of grammatical features and searching a word and its grammatical properties in the supported dictionaries.

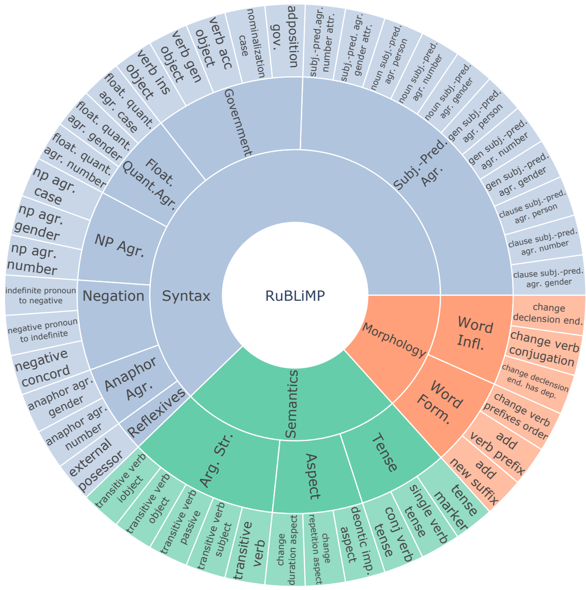


Figure 2: Distribution of phenomena in RuBLiMP.

(Hopper and Thompson, 1980; Paducheva, 2010). We search for a word or phrases with certain morphosyntactic features (e.g., a transitive verb) and semantic properties using a manually curated list of temporal markers and word co-occurrence and semantic dictionaries from the Russian National Corpus (Savchuk et al., 2024).

2.3 Phenomena

RuBLiMP includes 45 minimal pair types or *paradigms*, each containing 1k minimal pairs. All paradigms are grouped into 12 *phenomena* (see Figure 2), which are well represented in Russian theoretical and corpus linguistics. We provide a minimal pair example for each paradigm in Appendix A and describe each phenomenon below.

- WORD FORMATION: uninterpretable combinations of derivational affixes and violation of verb prefix stacking rules.
- WORD INFLECTION: incorrect use of declension affixes or verb conjugation endings.
- GOVERNMENT: incorrect use of a word governed by a nominalization, preposition, or verb.
- SUBJECT-PREDICATE AGREEMENT: violations of the subject-predicate agreement w.r.t. number, gender, person, or case. We include more complex agreement violation contexts with attractors.
- ANAPHOR AGREEMENT: incorrect agreement between an anaphoric relative pronoun and its antecedent in number or gender.

- NOUN PHRASE AGREEMENT: agreement violation between the head of a noun phrase and its modifiers, such as adjectives and determiners, w.r.t. number, gender, or case.
- FLOATING QUANTIFIER AGREEMENT: lack of number, gender, or case agreement between a floating quantifier and a noun.
- REFLEXIVES: incorrect use of a reflexive pronoun in constructions with an external possessor.
- NEGATION: negative particle movement and inappropriate use of negative and indefinite pronouns.
- ARGUMENT STRUCTURE: violations of the animacy requirement for a transitive verb's arguments via the replacement of a subject, direct or indirect object, and predicate in the active or passive voice.
- ASPECT: incorrect use of perfective verbs in contexts with semantics of duration and repetition and in negative constructions with deontic verbs.
- TENSE: incorrect choice of (i) a single or conjoined verb form in a sentence with temporal adverbial (an adverb or a noun phrase) and (ii) a temporal adverbial in a sentence with a tense-marked verb.

2.4 Minimal Pair Curation

Detecting pretraining data helps measure test data contamination and becomes a necessary component of the evaluation design (Brown et al., 2020; Gao et al., 2023). In our work, we employ a pretraining data detection method as a filtering stage when creating RuBLiMP. In particular, we use MIN-K% PROB (Shi et al., 2023), which relies on the hypothesis that a pretraining example is less likely to include outlier tokens with low probability compared to a non-pretraining example. The main idea is to compute the average log-likelihood of K% tokens with minimum probability and determine a threshold t used to classify an example as pretraining or non-pretraining. MIN-K% PROB does not require an access to an LM's pretraining corpus and is highly efficient for scoring-based evaluation, since both MIN-K% PROB and a sentence's probability are computed via a single forward pass.

We compute MIN-K% PROB for each grammatical sentence in a pool of generated minimal pairs using 25 LMs described in §3. For each paradigm, we then run a grid search for t, which allows to find an intersection of 1k minimal pairs between *all* LMs. The minimal pairs in the intersection contain unique grammatical sentences, which are not detected as pretraining examples for any LM[8]. We conduct ablation studies on choosing K% in §4.

2.5 Human Validation

Annotation Design We conduct an in-house human validation to verify that the generated minimal pairs unambiguously isolate a target phenomenon and illustrate a grammaticality contrast. We create a team of 20 undergraduate BA and MA students in fundamental and computational linguistics from several Russian universities. We collaborate closely with the students over the course of the annotation project and maintain communication in a group chat. Our project includes a training phase and a main annotation phase. Each student is given detailed annotation guidelines available at any time during both annotation phases. We train the students to perform the task on 10 examples with explanations and ensure that their training performance is above 70% (Nangia and Bowman, 2019). The main annotation phase counts 2,350 examples (50 minimal pairs per paradigm). Each student receives a page with 5 minimal pairs, one of which is a honeypot example[9]. The pay rate is on average $20/hr, the minimum response time per page is 25 seconds, and the average honeypot performance exceeds 75%. A shortened version of the guidelines and an example of the web interface are in Appendix C.1.

Vote Aggregation The students' votes are aggregated with the Dawid-Skene method (Dawid and Skene, 1979) using `Crowd-Kit` (Ustalov et al., 2024). We compute the inter-annotator agreement using the Worker Agreement with Aggregate (WAWA) coefficient (Ning et al., 2018), which indicates the average fraction of the annotators' votes that agree with the aggregated vote for each pair.

[8] We limit the maximum number of the generated minimal pairs for each paradigm to 350k. If the threshold search allows us to find more than 1k pairs in the LMs' intersection, we downsample the decontaminated pairs to 1k in a stratified fashion w.r.t. domain, length, and paradigm-specific features.

[9] Honeypot examples are a standard practice to estimate the annotation quality (Ustalov et al., 2024). Three authors of this paper prepare 250 honeypot minimal pairs by manually labelling the generated pairs as "positive" and "negative". Various inconsistencies are manually introduced to balance the number of "negative" examples, such as violation of several phenomena, perturbing multiple sentence units, or usage of ambiguous word forms. An annotator labels a honeypot example without knowing the ground truth label, and then the annotator's labels are compared against the authors' labels in order to measure the annotator's performance.

Paradigm	%	WAWA
Word Formation	95.77	92.83
Word Inflection	95.33	93.90
Government	91.83	91.84
Subject-Predicate Agreement	95.87	92.46
Anaphor Agreement	94.06	93.00
Noun Phrase Agreement	96.50	94.33
Floating Quantifier Agreement	97.28	92.37
Reflexives	100.0	96.50
Negation	93.33	92.60
Argument Structure	93.51	89.94
Aspect	95.28	92.97
Tense	93.79	92.10
Average	94.35	92.51

Table 2: The ratio of plausible minimal pairs (%) by phenomenon and per-phenomenon WAWA inter-annotator agreement rates.

Results We report the per-phenomenon results in Table 2 and per-paradigm results in Table 7 (see Appendix C.2). Overall, we observe a high ratio of plausible minimal pairs (94.35%), with more than 85% of correctly generated pairs for most of the paradigms. The average IAA as measured by WAWA is 92.5, indicating a strong agreement.

2.6 General Statistics

The RuBLiMP's general statistics are summarized in Table 3 and compared with the Russian subset of CLAMS (Mueller et al., 2020), a pattern-generated benchmark for subject-predicate agreement.

Length and Frequency We compute the ratio of high-frequency tokens in the grammatical sentences as follows. We divide the number of tokens whose number of instances per million in our corpus (§2.1) is ≥ 1 by the sentence length in tokens. The sentences contain on average 11.3 tokens and 87.4% of high-frequency tokens. In CLAMS, the sentences are shorter on average (7.55 tokens) and similar in terms of the high-frequency tokens ratio (86.3%). We also observe that the overall number of unique tokens in CLAMS's 40.1k grammatical sentences is 126, which indicates its low lexical diversity. In contrast, RuBLiMP's subset for the syntactic phenomena counts 57.9k unique tokens.

Syntactic Diversity We compute the dependency tree depth and the number of unique POS 5-grams and syntactic patterns at the benchmark- and sentence-level. The sentences vary in terms of the word order, with the number of unique POS 5-grams ranging between 18.9k (morphology) and 50k (syntax). The average tree depth in RuBLiMP is 4.18, and there are 24.6k unique syntactic patterns, with the average pattern frequency of 1.82 (see Appendix D). Comparing RuBLiMP's minimal pairs for the syntactic phenomena to CLAMS, we find that CLAMS has significantly less variety, with 70 unique syntactic patterns, and their average frequency of 573.1. The number of unique POS 5-grams and average tree depth are smaller: 99 and 2.94, respectively. This confirms that utilizing open text corpora promotes high linguistic diversity. We report the CLAMS's manual analysis results in §6.

	RuBLiMP				CLAMS
	Morphology	Semantics	Syntax	Overall	
Benchmark-level					
# Pairs	6k	11k	28k	45k	40k
# Patterns	3,9k	7,4k	15,9k	24,6k	70
Pattern Frequency	1.52	1.48	1.76	1.82	573.1
# Unique Tokens	20.7k	33.8k	57.9k	86.5k	126
# POS 5-Grams	18,9k	30,9k	50k	64,9k	99
Sentence-level					
Frequency (%)	86.6	88.9	87.0	87.4	86.3
Depth	4.02	4.41	4.12	4.18	2.94
# Tokens	10.46	12.23	11.14	11.31	7.55
# POS 5-Grams	6.46	8.23	7.14	7.31	3.56

Table 3: Benchmark- and sentence-level general statistics in comparison with CLAMS.

3 Experimental Setup

Model	Source	Size	Corpus
Encoder-only LMs			
ruBERT-base ruBERT-large	Zmitrovich et al. (2024)	178M 427M	Wikipedia, news
ruRoBERTa	Zmitrovich et al. (2024)	355M	Wikipedia, news, books
distil-MBERT MBERT	Sanh et al. (2019) Devlin et al. (2019)	134M 177M	Wikipedia
XLM-R_{base} XLM-R_{large}	Conneau et al. (2020)	279M 560M	C4
RemBERT MDeBERTa	Chung et al. (2021) He et al. (2022)	575M 276M	Wikipedia C4
Decoder-only LMs			
ruGPT-small ruGPT-medium ruGPT-large	Zmitrovich et al. (2024)	125M 355M 760M	Wikipedia, C4, news, books
ruGPT-3.5-13B	N/A	13B	Wikipedia, news, books, other
SambaLingo	Csaki et al. (2023)	7B	CulturaX
mGPT-1.3B mGPT-13B	Shliazhko et al. (2024)	1.3B 13B	Wikipedia, C4
bloom-1b7 bloom-3b bloom-7b1	Scao et al. (2023)	1.7B 3B 7.1B	ROOTS
xglm-1.7B xglm-4.5B xglm-7.5B	Lin et al. (2022)	1.7B 4.5B 7.5B	C4
Llama-7b Llama-13b	Touvron et al. (2023)	7B 13B	Web corpora
Mistral	Jiang et al. (2023)	7B	Web corpora

Table 4: The LMs used in our work. Corpora references: C4 (Raffel et al., 2020), CulturaX (Nguyen et al., 2024), and ROOTS (Laurençon et al., 2022).

Language Models Table 4 summarizes a broad range of 25 pretrained decoder- and encoder-

	ruBERT-base	ruBERT-large	ruRoBERTa	ruGPT-small	ruGPT-medium	ruGPT-large	ru-GPT-3.5-13B	SambaLingo	distil-MBERT	MBERT	XLM-R-base	XLM-R-large	RemBERT	MDeBERTa	mGPT-1.3B	mGPT-13B	bloom-1b7	bloom-3b	bloom-7b1	xglm-1.7b	xglm-4.5b	xglm-7.5b	Llama-7b	Llama-13b	Mistral	Average
k=30	-4.2	-4.2	-4.6	-5.5	-4.1	-4.9	-2.9	-7	-8.1	-9.6	-9.3	-8.6	1.3	2.7	-7.2	-6.7	-5.4	-5.8	-6.3	-9.4	-7.9	-7.6	-7.6	-6.7	-6.1	-5.8
k=40	-4.1	-4.3	-4.7	-6.1	-4.3	-5.5	-3.1	-8	-8.2	-10	-9.8	-9.2	1.3	2.9	-7.7	-7.3	-5.5	-5.9	-6.4	-11	-8.7	-8.3	-8.2	-7.3	-6.6	-6.2
k=50	-4.4	-4.4	-5	-6.5	-4.7	-5.9	-3.4	-8.9	-8.4	-11	-10	-9.9	1.2	2.9	-8.2	-7.8	-5.7	-6.3	-6.7	-11	-9.4	-9	-8.8	-7.9	-7.1	-6.7
k=60	-4.3	-4.4	-5	-6.8	-4.9	-6.2	-3.6	-9.6	-8.5	-11	-11	-10	1.3	3.1	-8.6	-8.1	-5.7	-6.3	-6.8	-12	-9.9	-9.4	-9.2	-8.3	-7.5	-6.9

Figure 3: Δ-scores (↓) for each LM and K% $\in \{30, 40, 50, 60\}$. All values are in %.

only LMs used in our work and accessed via `Transformers` (Wolf et al., 2020). Each LM is used in our MIN-K% PROB ablation studies (§4) and empirical evaluation experiments in monolingual (§5) and cross-lingual scenarios (§6).

Method The sentences in a minimal pair are ranked based on their perplexity (PPL) or pseudo-perplexity (PPPL). The PPL of a sentence s is inferred with a decoder-only LM as Equation 1, where $|s|$ is the sentence length in tokens and Θ denotes the LM's parameters.

$$PPL(s) = exp(-\frac{1}{|s|}\sum_{i=0}^{|s|}\log P_{\Theta}(x_i|x_{<i})) \quad (1)$$

The PPPL (Salazar et al., 2020) is computed with an encoder-only LM as in Equation 2. Each token x_j in s is masked out and predicted based on the past and future tokens $x_{\backslash i} = (x_1, \ldots, x_{i-1}, \ldots, x_{i+1}, \ldots, x_{|s|})$.

$$PPPL(s) = exp(-\frac{1}{|s|}\sum_{i=0}^{|s|}\log P_{\Theta}(x_i|x_{\backslash i})) \quad (2)$$

Human Baseline We establish the human baseline on 5% of RuBLiMP (2,350 pairs; 50 pairs per paradigm) using ABC[10], a crowdsourcing platform. Each of the 144 hired workers is certified as a native Russian speaker and paid \$15/hr on average. The annotation task is to select a grammatical sentence in a given pair (see Appendix E). The sentences in a pair are randomly shuffled. We use 10 training and 100 honeypot examples and aggregate the votes using the Dawid-Skene method. The average response time per one pair is 10 seconds, and the average honeypot performance exceeds 90%.

4 MIN-K%: Ablation Studies

We begin with ablation studies on the effect of the minimal pair curation stage and the hyperparameter K% $\in \{30, 40, 50, 60\}$. For each paradigm in RuBLiMP, (i) we randomly sample 1k generated minimal pairs and evaluate the LMs to get the reference scores (the accuracy scores are averaged over 100 runs), and (ii) decontaminate the generated minimal pairs through a greed search for t and select 1k pairs with the maximum MIN-K% PROB as described in §2.4 and evaluate the LMs' performance. We then compute the Δ-score between (i) and (ii) for each LM, which measures the performance drop when using MIN-K% PROB with certain K%.

Higher K% is More Effective Figure 3 shows that MIN-K% PROB ensures adversarial filtering of the pool of generated minimal pairs. In general, the higher K% value, the lower the Δ-score for most LMs. We find that the overall performance can drop from 2.9% to 12% and the Δ-score can depend on the model size (e.g., `ruGPT`, `bloom`, and `Llama-2`). However, the Δ-scores for `RemBERT` and `MDeBERTa` are positive; we relate it to the fact that these LMs perform close to random guessing on RuBLiMP (§5) and other related benchmarks (§6). We select K% of 60 to create RuBLiMP.

5 Results on RuBLiMP

This section describes the empirical evaluation results on RuBLiMP. We report the results by phenomenon in Table 5 and by paradigm in Appendix F. Overall, we find that the best performing and the largest monolingual LM (`ruGPT-3.5-13B`) still falls short compared to humans, whose performance exceeds 95% on all RuBLiMP's paradigms. Analyzing the results for the monolingual and multilingual LMs, we observe that the former generally perform better, and the latter can achieve the random baseline performance (e.g., `RemBERT`, `MDeBERTa`, `xglm-1.7B`). We evaluate the multilingual LMs on five related BLiMP-style benchmarks to explore this behavior in more detail (§6). Below,

[10]Available only in Russian: `elementary.center`

Model	Word Form.	Word Infl.	Government	Subj.-Pred. Agr.	Anaphor Agr.	NP Agr.	Float. Quant. Agr.	Reflexives	Negation	Arg. Structure	Aspect	Tense	Average
`ruBERT-base`	81.90	84.87	90.72	91.16	85.90	83.57	91.40	78.70	77.77	88.52	96.07	87.17	86.48
`ruBERT-large`	82.83	86.03	90.66	91.43	86.35	84.70	91.23	81.00	82.13	89.20	96.47	88.17	87.52
`ruRoBERTa`	89.67	91.63	96.68	93.62	95.60	88.83	96.17	91.10	89.83	91.64	97.20	92.40	92.86
`ruGPT-small`	88.53	91.57	92.94	90.33	94.30	95.33	87.63	83.20	73.17	88.82	93.93	84.30	88.67
`ruGPT-medium`	91.77	86.37	94.88	91.57	95.90	97.37	96.17	79.90	80.53	91.98	95.60	88.70	90.89
`ruGPT-large`	89.23	91.37	94.58	91.51	95.80	96.77	90.83	87.80	78.60	92.24	95.53	87.03	90.94
`ruGPT-3.5-13B`	94.33	95.20	97.10	96.12	97.05	98.47	98.17	94.70	87.53	96.34	97.77	95.37	95.68
`SambaLingo`	79.87	85.73	89.20	80.85	92.95	89.83	90.43	96.20	77.63	82.74	87.40	80.47	86.11
`distil-MBERT`	83.97	79.63	70.84	75.90	52.35	79.43	83.13	56.00	75.27	55.88	59.47	55.83	68.98
`MBERT`	88.83	84.63	78.88	80.37	86.35	87.07	82.77	52.90	66.30	61.22	59.77	52.70	73.48
`XLM-R`$_{\texttt{base}}$	88.57	90.57	88.42	87.55	91.85	92.67	92.97	69.90	72.50	75.48	81.57	74.67	83.89
`XLM-R`$_{\texttt{large}}$	88.80	91.03	90.52	87.73	93.15	94.37	93.07	79.10	80.67	81.30	87.70	79.77	87.27
`RemBERT`	51.40	54.70	48.90	49.77	32.05	51.17	62.63	45.30	51.17	49.28	52.40	52.20	50.08
`MDeBERTa`	52.57	43.63	47.50	36.77	75.35	41.03	37.43	40.20	43.57	41.90	44.10	53.53	46.47
`mGPT-1.3B`	94.37	95.97	89.64	87.69	92.15	94.20	85.13	82.50	67.80	79.94	85.43	79.53	86.20
`mGPT-13B`	94.53	95.53	92.08	88.75	94.35	95.33	88.50	85.60	68.27	84.46	87.70	83.03	88.18
`bloom-1b7`	86.10	89.70	67.86	85.55	69.75	79.10	66.87	13.10	65.20	53.78	54.67	68.87	66.71
`bloom-3b`	89.53	90.23	71.44	86.66	65.85	81.10	68.33	19.30	67.83	54.22	51.13	72.30	68.16
`bloom-7b1`	88.90	91.87	73.62	88.91	73.10	84.63	75.37	23.30	68.40	57.64	55.07	77.43	71.52
`xglm-1.7B`	37.70	45.03	51.72	44.26	65.70	61.57	47.23	64.10	54.63	38.20	75.27	51.93	53.11
`xglm-4.5B`	92.40	92.17	87.96	82.70	92.80	92.70	91.30	82.90	73.97	82.62	90.47	80.57	86.88
`xglm-7.5B`	92.80	93.43	88.46	83.75	93.45	93.70	91.03	90.80	74.43	83.12	90.27	82.37	88.13
`Llama-7b`	94.70	90.83	85.20	89.45	48.35	89.23	72.10	84.80	72.40	79.96	81.20	81.93	80.85
`Llama-13b`	95.83	93.50	88.50	91.23	56.00	91.53	76.97	89.00	74.00	83.08	85.50	85.63	84.23
`Mistral`	96.87	95.00	88.16	92.99	72.10	93.20	87.83	32.40	72.40	83.28	86.60	88.13	82.41
Human	100.0	99.33	98.80	98.53	98.0	98.67	99.33	98.00	100.0	100.0	100.0	99.33	98.62

Table 5: The average accuracy scores (%) of the 25 LMs and human baseline by phenomenon. Random baseline is 50%. The monolingual and multilingual LMs are separated by a line.

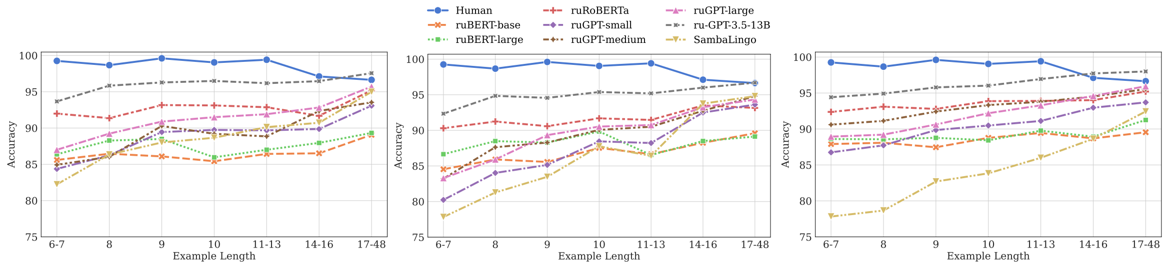
we discuss our findings from the perspective of the LM size, phenomenon, domain, and length.

Larger $\neq$ Better We find that smaller LMs can outperform or perform on par with larger LMs. In particular, `ruGPT-medium` performs close to `ruGPT-large` on average, while `ruBERT-base` & `ruBERT-large`, `xglm-4.5B` & `xglm-7.5B`, and `mGPT-1.3B` & `mGPT-13B` perform on par on certain phenomena (e.g., Word Formation, Anaphor Agreement, and Tense). This finding aligns with Warstadt et al. (2020); Song et al. (2022).

Higher Sensitivity to Local Edits The LMs are robust to local perturbations for Word Inflection and Word Formation. We observe that the LMs can perform on par with humans in identifying an incorrect order of the verb prefixes. The presence of a modifier helps the LMs resolve an incorrect word's declension, improving the accuracy by up to 5% (e.g., `ruBERT`, `ruGPT`, and `mGPT`).

Lower Sensitivity to Structural Relations The LMs achieve lower performance on the structural phenomena (Reinhart, 2016). The behavior is more pronounced for the multilingual LMs, which fall behind humans by up to 40% on Anaphor Agreement and 45% on Reflexives.

LMs Struggle with Negative Pronouns Negation is one of the challenging phenomena in RuBLiMP. In particular, most LMs are least sensitive to the replacement of a negative pronoun with an indefinite one (see Appendix F), which requires understanding of the pronoun licensing conditions (e.g., *On nikogda/*kogda-nibud' ne hodit v teatr*



(a) Wikipedia (b) Wikinews (c) Librusec

Figure 4: Results on RuBLiMP for the monolingual LMs per domain grouped by seven quintiles of the length.

"He never/*ever goes to the theatre"). However, the LMs distinguish well between the sentences without a negative particle *ne* "not" where an indefinite pronoun is replaced with a negative one (e.g., *Petr kogda-to/*nikogda byl v Moskve* "Petter was once/*never in Moscow").

Attractors Confuse LMs Analyzing the effect of the attractor presence (see Appendix B.2 for details), we observe that the LMs' performance can drop by up to 10% on SUBJECT-PREDICATE AGREEMENT if an attractor is added (see Appendix F; e.g., `bloom`, `SambaLingo`, and `mGPT`).

LMs are Less Sensitive to Tense Another finding is that the LMs struggle to identify a violated tense form of a single verb, with the accuracy ranging between the random guessing (`xglm-1.7B`) to 90.7% (`ruGPT-3.5-13B`). However, having a conjoined verb increases the performance by up to 17.3% (`mGPT-1.3B`), which indicates that the LMs utilize the context reliably.

Effect of Length & Domain We estimate the effect of length per domain by dividing RuBLiMP into 7 length groups of equal size. The results for the monolingual and multilingual LMs are in Figure 4 and Figures 5-6 (Appendix F), respectively. While human performance is consistent, the LMs' performance improves as the length increases. The first length groups (6-10 tokens) contain pairs related to the most challenging phenomena for the LMs (syntax: NEGATION, REFLEXIVES; semantics: ARGUMENT STRUCTURE, ASPECT). We find that some LMs are more domain-sensitive (e.g., `SambaLingo`, `ruGPT`), while others receive similar scores (e.g., `ruGPT-3.5`, `XLMR`).

6 Multilingual Analysis

To analyze the multilingual LMs in more detail, we evaluate their sensitivity to linguistic phenomena in six benchmarks: BLiMP, CLiMP, SLING, JBLiMP, CLAMS, and RuBLiMP (see Table 1 for statistics). We detail the experimental setup and empirical evaluation results in Appendix G and outline our key findings here. (i) no single LM performs consistently well in all languages, (ii) the LMs' performance for AGREEMENT in a given language depends on the benchmark, and the Δ-scores between the benchmarks can be up to 15% for English, 20% for Chinese, and 35% for Russian, and (iii) the manual analysis of CLAMS reveals its concerning quality: 20% of Russian minimal pairs are semantically implausible, 15% do not isolate a phenomenon, and 5% contain repetitive text (a native Russian speaker will unlikely say or write this way). Besides, there are only 126 unique tokens in the 40.1k grammatical sentences, which limits the sentence diversity. These findings raise the need for a more detailed comparison of LMs on peer-reviewed evaluation resources and their additional validation, which aligns with Song et al. (2022).

7 Related Work

Evaluating Russian LMs' Grammatical Knowledge Earlier studies introduce mono- and multilngual probing suites to explore how the LMs' representations encode Russian grammatical phenomena, ranging from a word's part of speech to gapping (e.g., Ravishankar et al., 2019; Şahin et al., 2020; Mikhailov et al., 2021; Choenni and Shutova, 2022; Serikov et al., 2022). RuCoLA (Mikhailov et al., 2022) includes expert-written and machine-generated (un-)acceptable sentences and aims to test the LM's linguistic competence via supervised acceptability classification. Our work extends the direction of evaluating Russian LM's grammatical knowledge and focuses on unsupervised acceptability judgments over linguistic minimal pairs.

Benchmarks of Linguistic Minimal Pairs The idea of discriminating between linguistic minimal pairs has gained visibility in NLP due to its several advantages, such as controlling the sentences' length and lexical units and providing a local view of an LM's decision boundary (Lau et al., 2017; Warstadt and Bowman, 2022). With the creation of BLiMP (Warstadt et al., 2020), similar resources have been proposed to evaluate LMs' acquisition of grammatical phenomena in languages other than English (see Table 1). In contrast to these benchmarks, RuBLiMP cover diverse phenomena in Russian morphology, syntax, and semantics beyond subject-verb agreement in CLAMS and includes pairs generated from naturally occurring and decontaminated sentences across three domains.

8 Conclusion and Future Work

This work introduces RuBLiMP, the first large-scale multidomain benchmark of 45k minimal pairs for the Russian language. RuBLiMP covers 45 minimal pair types grouped into 12 linguistic phenomena in morphology, syntax, and semantics. The RuBLiMP creation approach ensures the linguistic diversity and high quality of the minimal pairs and minimizes the data contamination risk. We conduct an extensive empirical evaluation of 25 widely used monolingual and multilingual LMs for Russian and analyze their performance w.r.t. various criteria. Our results show that the LMs are better at identifying morphological and agreement-oriented contrasts than violations of structural relations, negation, transitivity, and tense. Furthermore, we analyse the 17 multilingual LMs in seven languages and find that no single LM performs well in all languages. Our *future work* includes (i) comparison of pretraining data detection methods (ii) implementation of new phenomena (e.g., islands), and (iii) a more detailed multilngual study of the LMs' linguistic abilities. By releasing RuBLiMP, we hope to foster further research on how the Russian language is acquired by LMs.

Limitations

This section describes the limitations of our work associated with our multi-stage minimal pair generation approach and computational costs. Noise in the publicly available data and automatic data extraction and annotation errors can generate implausible pairs. However, each stage is highly customizable based on the user needs, and expert validation of our approach shows that roughly 2,235 out of 2,350 generated minimal pairs unambiguously isolate a target phenomenon and display the required grammaticality contrast (§2.5).

Corpus Annotation In our work, we utilize data sources that undergo human review and editing (e.g., Wikipedia and Wikinews articles). However, there is still a high chance of noise in the data, such as web page artifacts or errors of optical character recognition systems. Another disadvantage of this stage is errors in the text segmentation tools and morphosyntactic parsers. We use the current state-of-the-art Russian NLP libraries and models and create a set of shallow heuristics to filter out irrelevant sentences through a series of manual data analysis iterations.

Minimal Pair Generation On the one hand, our multidomain corpus represents a large-scale source of sentences with a high degree of diversity in terms of lexis, length, frequency, and linguistic structures. On the other hand, there are a few challenges due to the rich Russian morphology, a high degree of ambiguity, and a flexible word order. In particular, not all grammatical sentences with relevant linguistic constructions can be perturbed into ungrammatical ones, e.g., many word perturbations still result in plausible sentences and require additional heuristics to prevent semantic and syntactic felicity, which is not always possible. This is the main reason for narrowing down a set of linguistic structures and contexts to ensure control over the perturbations. We limit the number of the (i) phenomena criteria (e.g., considering nominalizations only with a specific set of endings), (ii) perturbation options (e.g., discarding ambiguous case forms during the government violations), or both (i) and (ii) (e.g., selecting verbs with only two prefixes during the word formation violations and only changing their order instead of adding more prefixes). Last but not least, the search for relevant lexical units and linguistic structures depends on the domain, which limits the scope of the domain-specific performance analysis (e.g., the temporal markers describing the duration or repetition of an event are primarily found in the news domain).

Minimal Pair Curation Recent research has proposed a broad range of pretraining data detection methods. Our work does not aim to compare different solutions to this problem; we recognize that more advanced methods can be applied (e.g., MIN-

K%++; Zhang et al., 2024). We also acknowledge that the MIN-K% PROB method may still identify sentences that *do appear* in the LMs' pretraining corpora as non-pretraining examples and select sentences with rare vocabulary items, which may lead to the performance decrease. Naturally, the effectiveness of the curation stage and the resulting LM's performance depends on the quality of the pretraining data detection method, which is an open question in the LM evaluation & benchmarking research direction (Oren et al., 2023). However, our approach allows one to continuously update RuBLiMP and create multiple versions of the benchmark, which can be decontaminated w.r.t. a set of LMs and another test data decontamination methods (or their ensemble).

Domain Shifts Many studies report that LMs can judge frequent linguistic patterns in their pretraining corpora as grammatical and perform worse on rare sentences with low probabilities (Marvin and Linzen, 2018; Linzen and Baroni, 2021). Our benchmark design implies potential word frequency and domain distribution shifts between an LM's pretraining corpus and RuBLiMP, which can introduce bias in the evaluation. Nevertheless, we demonstrate a high diversity of syntactic patterns and a moderate word frequency in RuBLiMP's sentences (§2.6), and show that the LMs can generalize well to out-of-domain examples (§5).

Computational Costs Each stage in our minimal pair generation approach requires efficient computational resources. However, the morphosyntactic parser in §2.1 can be replaced with a more lightweight one with possible changes in the annotation quality (e.g., `slovnet`[11]). Note that the minimal pair curation stage costs are reduced as follows. First, we filter out pairs based on MIN-K% PROB for decoder-only LMs due to their optimal inference speed. Next, we filter out the remaining pairs based on MIN-K% PROB for encoder-only LMs. Recall that both MIN-K% PROB and a sentence's probability are computed via a single forward pass.

Ethics Statement

Human Annotation The annotators' votes in our annotation projects (see §2.5; §3) are collected anonymously. The average pay rate significantly exceeds the hourly minimum wage in Russia. The annotators are warned about potentially sensitive topics in the examples, such as politics, culture, and religion.

Inference Costs Evaluating an LM on RuBLiMP depends on the LM architecture and size and can be optimized with distributed inference libraries (e.g., `accelerate`[12]). Running the complete evaluation experiment on a single V100 GPU takes approx. 1.5h and 11h for a decoder-only and encoder-only LM, respectively.

Potential Misuse RuBLiMP can be used as training data for acceptability classifiers, potentially enhancing the quality of generated texts (Batra et al., 2021). We acknowledge that these improvements in text generation might lead to the misuse of LMs for harmful purposes (Lucas et al., 2023). RuBLiMP's intended use is for **research and development purposes**, and the potential negative uses are not lost on us.

Transparency We release RuBLiMP, our minimal pair generation framework, and all annotation materials under the permissive license following the standard open research practices. Our GitHub repository and HuggingFace dataset card (Lhoest et al., 2021) provide detailed documentation on the codebase, benchmark creation methodology, and human annotation.

Use of AI-assistants We improve and proofread the text of this paper using Grammarly[13] to correct grammatical, spelling, and style errors and paraphrasing sentences. Therefore, specific segments of our publication can be detected as AI-generated, AI-edited, or human-AI-generated.

[11] `github.com/natasha/slovnet`

[12] `github.com/huggingface/accelerate`

[13] `grammarly.com`

References

D. G. Anastasyev. 2020. Exploring pretrained models for joint morpho-syntactic parsing of Russian. In *Computational Linguistics and Intellectual Technologies*, pages 1–12.

Aysa Arylova. 2013. *Possession in the Russian clause: towards dynamicity in syntax*. Ph.D. thesis, University of Groningen. Relation: http://www.rug.nl/ Rights: University of Groningen.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

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.

Victor Bocharov, Svetlana Alexeeva, Dmitry Granovsky, Ekaterina Protopopova, Maria Stepanova, and Aleksei Surikov. 2013. Crowdsourcing morphological annotation. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue*, volume 1, pages 109–114.

Elena Bolshakova and Andrey Sapin. 2021. Building dataset and morpheme segmentation model for russian word forms. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 154–161.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

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

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently Adapting Pre-trained Language Models to New Languages. *arXiv preprint arXiv:2311.05741*.

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.

Ferdinand de Haan. 2002. Strong modality and negation in russian. In Randi Reppen, Susan Fitzmaurice, and Douglas Biber, editors, *Using Corpora to Explore Linguistic Variation*, pages 91–110. John Benjamins, Amsterdam/Philadelphia.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, chapter 5, pages 58–90. The MIT Press.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *Language*, pages 251–299.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

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.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models. *Preprint*, arXiv:2309.06085.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

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.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel, and Ekaterina Artemova. 2021. RuSentEval: Linguistic source, encoder force! In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 43–65, Kiyv, Ukraine. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational 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.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

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.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving test set contamination for black-box language models. In *The Twelfth International Conference on Learning Representations*.

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

Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia Loukachevitch, and Chris Biemann. 2017. *Human and Machine Judgements for Russian Semantic Relatedness*, pages 221–235. Springer International Publishing, Cham.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. Association for Computational Linguistics.

Tanya Reinhart. 2016. *Anaphora and semantic interpretation*. Routledge.

Robert Reynolds. 2013. Out of order?: Russian prefixes, complexity-based ordering and acyclicity. *University of Pennsylvania Working Papers in Linguistics*.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. Linspector: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems*.

Svetlana Savchuk, Timofey Arkhangelsky, Anastasia Bonch-Osmolovskaya, Olga Donina, Yulia Kuznetsova, Olga Lyashevskaya, Boris Orekhov, and Maria Podryadchikova. 2024. Natsionalny Korpus Russkogo Yazyka 2.0: Novye Vozmozhnosti i Perspektivy Razvitiya (in Russian). *Voprosy Yazykoznaniya*, 2:7–34.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo

Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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

Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova, and Tatiana Shavrina. 2022. Universal and independent: Multilingual probing framework for exhaustive model interpretation and evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 441–456, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

D. Sichinava. 2018. Preposition. Materials for the project of corpus description of Russian grammar (http://rusgram.ru).

Natalia Slioussar and Anton Malko. 2016. Gender agreement attraction in russian: production and comprehension evidence. *Frontiers in psychology*, 7:166019.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leon Stassen. 2013. Predicative possession. In Matthew S. Dryer and Martin Haspelmath, editors, *WALS Online*, v2020.3 edition. Zenodo. Available online at http://wals.info/chapter/117, Accessed on 2023-11-16.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96):6227.

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.

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. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

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.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

Andrey Zaliznyak. 1987. *Grammatical Dictionary of Russian Language: Word Inflection*. Moscow.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2024. Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. *arXiv preprint arXiv:2404.02936*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena 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 Examples of Minimal Pairs

Phenomenon	PID	Acceptable Example	Unacceptable Example
Word Formation	add_new_suffix	Priekhala **staren'kaya**, malen'kaya, khuden'kaya zhenshchina.	Priekhala **starsken'kaya**, malen'kaya, khuden'kaya zhenshchina.
	add_verb_prefix	Vesnoy lichinki **vyedayut** pochki iznutri i v mae okuklivayutsya.	Vesnoy lichinki **vyv"edayut** pochki iznutri i v mae okuklivayutsya.
	change_verb_prefixes_order	Khochu dolozhit', chto plany po etoy rabote my **perevypolnili**.	Khochu dolozhit', chto plany po etoy rabote my **vyperepolnili**.
Word Inflection	change_declension_ending	Fil'm byl dostatochno podrobno rassmotren v **zhurnale** "Iskusstvo kino".	Fil'm byl dostatochno podrobno rassmotren v **zhurnali** "Iskusstvo kino".
	change_declension_ending_has_dep	Znachitel'nye ploshchadi pashni podverzheny vodnoy **erozii**.	Znachitel'nye ploshchadi pashni podverzheny vodnoy **eroziu**.
	change_verb_conjugation	I nikomu uzhe ne **dokazhesh'**, chto ty – eto ty, – tak on dumal.	I nikomu uzhe ne **dokazhish'**, chto ty – eto ty, – tak on dumal.
Government	adposition_government	Vpervye kosmonavt spal v **nevesomosti**.	Vpervye kosmonavt spal v **nevesomost'yu**.
	verb_acc_object	My opishem nashu **aksiomatizatsiyu** putem opisaniya struktury formul.	My opishem nashu **aksiomatizatsiya** putem opisaniya struktury formul.
	verb_gen_object	Summy ne ochen' bol'shie, no **azarta** oni dobavlyayut.	Summy ne ochen' bol'shie, no **azartom** oni dobavlyayut.
	verb_ins_object	Dioksid kremniya obladaet **polimorfizmom**.	Dioksid kremniya obladaet **polimorfizma**.
	nominalization_case	Pakhotnye ploshchadi podverzheny smyvu i vyduvaniyu **vetrom**.	Pakhotnye ploshchadi podverzheny smyvu i vyduvaniyu **veter**.
Subject-predicate Agreement	noun_subj_predicate_agreement_number	Pua-Katiki — potukhshij shchitovidnyj **vulkan** na ostrove Paskhi.	Pua-Katiki — potukhshij shchitovidnyj **vulkany** na ostrove Paskhi.
	genitive_subj_predicate_agreement_number	Predposylok dlya muzykal'noy kar'jery v ee sem'je ne **bylo**.	Predposylok dlya muzykal'noy kar'jery v ee sem'je ne **byli**.
	clause_subj_predicate_agreement_number	Takim obrazom, dlya bol'shikh programm **prikhodilos'** ispol'zovat' overlei.	Takim obrazom, dlya bol'shikh programm **prikhodilis'** ispol'zovat' overlei.
	subj_predicate_agreement_number_attractor	Rasprostranennost' drugikh yazykov **nevelika**.	Rasprostranennost' drugikh yazykov **neveliki**.
	noun_subj_predicate_agreement_gender	Na territorii kompleksa **postroen** Kongress-tsentr.	Na territorii kompleksa **postroena** Kongress-tsentr.
	genitive_subj_predicate_agreement_gender	Ushedshikh iz kluba v dannyj transfernyj period ne **bylo**.	Ushedshikh iz kluba v dannyj transfernyj period ne **byla**.
	clause_subj_predicate_agreement_gender	Dalee **neobkhodimo** sdelat' obratnuyu zamenu.	Dalee **neobkhodima** sdelat' obratnuyu zamenu.
	subj_predicate_agreement_gender_attractor	Mestnost' vokrug sela sil'no **zabolochena**.	Mestnost' vokrug sela sil'no **zabolocheno**.
	noun_subj_predicate_agreement_person	Liturgicheskaya komissiya **rabotaet** v Monreale.	Liturgicheskaya komissiya **rabotayu** v Monreale.
	genitive_subj_predicate_agreement_person	Detey u Magnusa i Elizavety ne **bylo**.	Detey u Magnusa i Elizavety ne **budu**.
	clause_subj_predicate_agreement_person	Po otsenkam, **ostaetsya** raskopat' okolo 350 m.	Po otsenkam, **ostaesh'sya** raskopat' okolo 350 m.
Anaphor Agreement	anaphor_agreement_number	Est' neskol'ko rastenij, **kotorye** mozhno nayti tol'ko v Velikobritanii.	Est' neskol'ko rastenij, **kotoroe** mozhno nayti tol'ko v Velikobritanii.
	anaphor_agreement_gender	Tekhnika, **kotoruyu** on izobrel, poluchila nazvanie «skul'ptura sveta».	Tekhnika, **kotoryj** on izobrel, poluchila nazvanie «skul'ptura sveta».
Noun Phrase Agreement	np_agreement_number	No **malen'kaya** geroinya vashego naroda ostalas' tverda.	No **malen'kie** geroinya vashego naroda ostalas' tverda.
	np_agreement_gender	Titul luchshej komandy Anglii **togo** sezona takzhe otoshel «osam».	Titul luchshej komandy Anglii **toj** sezona takzhe otoshel «osam».
	np_agreement_case	Zoloto bylo obnaruzheno v **etom** rajone v 1923 godu.	Zoloto bylo obnaruzheno v **etogo** rajone v 1923 godu .
Floating Quant. Agreement	floating_quantifier_agreement_number	Informatsiyu podtverdili i v **samoj** shkole.	Informatsiyu podtverdili i v samoj shkolakh.
	floating_quantifier_agreement_gender	Pri etom **samo** povestvovanie nikuda vas ne gonit.	Pri etom **sama** povestvovanie nikuda vas ne gonit.
	floating_quantifier_agreement_case	Ego **samogo** uzhe malo kto priznaet avtoritetom.	Ego **samomu** uzhe malo kto priznaet avtoritetom.
Reflexives	external_possessor	Potomki metsenata perebralis' v Moskvu, gde u **Zhivago** byl biznes.	Potomki metsenata perebralis' v Moskvu, gde u **sebya** byl biznes.
Negation	negative_concord	I konechno, nikto **ne** toropilsya vzyat' vinu na sebya.	I konechno, nikto toropilsya vzyat' vinu **ne** na sebya.
	negative_pronoun_to_indefinite	Poetomu **nikogda** ne ostanavlivaytes', vsegda idite vpered!	Poetomu **kogda-libo** ne ostanavlivaytes', vsegda idite vpered!
	indefinite_pronoun_to_negative	**Chto-to** podskazyvaet, chto nechto pokhozhee my uvidim i v Parizhe.	**Nichto** podskazyvaet, chto nechto pokhozhee my uvidim i v Parizhe.
Argument Structure	transitive_verb	Ya rasshchityval, chto budet mnogo smaylov i vse **zatsenyat** sarkazm.	Ya rasshchityval, chto budet mnogo smaylov i vse **voskhodyat** sarkazm.
	transitive_verb_subject	**Shante** teryaet soznanie i snova prosypaetsya v svoej krovati.	**Khimiya** teryaet soznanie i snova prosypaetsya v svoej krovati.
	transitive_verb_passive	Al'tron byl unichtozhen **Vizhenom**, kotoryj prines sebya v zhertvu.	Al'tron byl unichtozhen **navykom**, kotoryj prines sebya v zhertvu.
	transitive_verb_object	Professor Farnsvort naznachaet **Lilu** kapitanom kosmicheskogo korablya.	Professor Farnsvort naznachaet **krug** kapitanom kosmicheskogo korablya.
	transitive_verb_iobject	Nasledniki posle ego smerti prodali dvorets **Oginskim**.	Nasledniki posle ego smerti prodali dvorets **fragmentam**.
Aspect	change_duration_aspect	Pri etom vopros avtorstva dolgo **ostavalsya** otkrytym.	Pri etom vopros avtorstva dolgo **ostalsya** otkrytym.
	change_repetition_aspect	Boll kazhdyj god **posylala** tsvety na den' rozhdeniya svoej podruge.	Boll kazhdyj god **poslala** tsvety na den' rozhdeniya svoej podruge.
	deontic_imperative_aspect	Vse serii kogda-to zakanchivayutsya, ne stoit etomu **udelyat'** vnimanie.	Vse serii kogda-to zakanchivayutsya, ne stoit etomu **udelit'** vnimanie.
Tense	single_verb_tense	A vchera on **dopustil** ochen' grubuyu oshibku.	A vchera on **dopustit** ochen' grubuyu oshibku.
	conj_verb_tense	Poslezavtra utrom on uzhe **pokinet** MKS i budet na Zemle.	Poslezavtra utrom on uzhe **pokinul** MKS i budet na Zemle.
	tense_marker	Tonnel' na Sinopskoy naberezhnoy otkroyut na **budushchey** nedele.	Tonnel' na Sinopskoy naberezhnoy otkroyut na **minuvshey** nedele.

Table 6: Examples of all 45 paradigms in RuBLiMP.

B Minimal Pair Generation

In this section, we provide a detailed description of the minimal pair generation procedure for each phenomenon in RuBLiMP.

B.1 Morphology

B.1.1 Word Formation

The minimal pairs in this phenomenon are created to violate the principles of affix ordering, namely (i-ii) prefix stacking rules (Reynolds, 2013), and (iii) suffixation universals (Greenberg, 1963).

Contexts We create a list of affixes (including all their possible allomorphs) that we can add or swap and manually annotate, dividing them into two subtypes: derivation and inflection. Thus, we limit the contexts to sentences where at least one word has one or several affixes from the list. This gives us more control when generating minimal pairs since not every random affix change leads to ungrammaticality. Additionally, we limit the number of prefixes a target verb can have to two since having more prefixes is less common in Russian.

Implementation Details We search the sentences for a possible target word (e.g., a verb with a prefix) and segment it into morphological elements using `pymorphy2` and dictionaries from Bolshakova and Sapin (2021). To generate the minimal pairs we then (i) add a new prefix to a verb (e.g., *za-pisat'* 'to write down' → **pro-za-pisat'*); swap verb prefixes to change their order (*pri-**u**-krasit'* 'to embellish' → ****u**-pri-krasit'*); or add a derivational suffix between the root and existing suffixes (*vodoprovod-n-aya* 'tap [water]' → **vodoprovod-**ist**-n-aya*). We check that the added affixes co-occur with the root to make the examples more probable, w.r.t. co-occurrence frequency. Finally, we check that the target word does not exist in the `pymoprhy2` dictionaries to ensure that the obtained word is ungrammatical.

B.1.2 Word Inflection

Word Inflection phenomenon includes errors in (i) verb conjugation and declension of (ii) a single noun or (iii) a noun with modifiers.

Contexts Since the list of inflectional affixes is not unique to every declension and conjugation (i.e., there are intersections between classes), we curate a dictionary of possible suffix perturbations. We create the dictionary so that the new suffix will not be interpreted as a different form of the same conjugation/declension. This way, each suffix replacement will lead to ungrammatical forms.

Implementation Details We use the manually crafted dictionaries to violate declension or conjugation of target words. In the verb conjugation violations (i) we replace the verb's inflection with inflection of the opposite conjugation (`I` ↔ `II`) with the same tense, number and person values. For example, the affix ***-et*** (*fut.3sg*) of the `I` conjugation verb *chita-**et*** 'is reading' is replaced with ***-it***, the `II` conjugation affix for the *fut.3sg* verb form.

For the declension violations (ii-iii), we change the inflectional suffixes of a noun to the suffixes of another declension. Similarly to (i), we ensure that the new inflection suffixes preserve the gender and case values of the word. For example, *stol-**a*** 'table' (`m.sg.gen`, `II` declension) is changed to *stol-**i***, where ***-i*** is the `m.sg.gen` affix of the `III` declension).

We then check that the resulting word does not contain any combinations of letters that do not exist in Russian. We created a list of non-occurring letter sequences based on RNC data. Finally, we check that the new word form is ungrammatical, using the `pymoprhy2` dictionaries.

B.2 Syntax

B.2.1 Government

Government refers to the government of the grammatical case of a noun, wherein a verb or a preposition determines the grammatical case of its noun phrase complement. We violate the government rules by changing the case of the objects of verbs governing (i) Accusative, ii) Instrumental, (iii) Genitive, (iv) a (pro)noun in a prepositional phrase, or (v) a dependent of the nominalization.

Contexts Since several adpositions allow different cases (e.g., *v* 'in' allows both, *v dome* 'in the house (Locative)' and *v litso* 'in the face (Accusative)'), we create a list of adpositions and their allowed cases based on Sichinava (2018). To find nominalizations, we check for words ending with *-nie* as in *odobrenie* 'blessing'. Since many modifiers in Russian agree with their heads in number, case, and gender, a change in any of those categories will lead to agreement violations. To ensure that the phenomenon is isolated, we only include the sentences where the target word (i.g. a verb's object, a dependent of a nominalization, or an adposition) has no modifiers.

Implementation Details We search the sentences for required constructions (e.g., a noun with a preposition in its dependents) and use pymorphy2 to change the form of the target word. Notably, to isolate the phenomenon, we ensure that the resulting word form is not ambiguous, i.e., it cannot be interpreted as two different forms (e.g., `acc.sg` is often the same as `nom.pl`).

B.2.2 Subject-Predicate Agreement

Subject-Predicate Agreement phenomenon includes agreement errors in the domain of the clause, where the subject controls agreement on the predicate. The predicate is often a verb, but sometimes it is an adjective, a participle, or, rarely, a noun. Subject kinds are described below. Our paradigms include violations of agreement in one of the three features: number, gender, and person, which happen in one of the four contexts: with the nominal subject, with the genitive subject, with the clausal subject, and with any subject, but in the presence of an attractor.

Contexts In an ungrammatical sentence, a single feature of the predicate or the subject is altered in the following contexts:

- Nominal subject: The subject is a noun phrase (including pronouns) in the nominative case. The predicate agrees with it for number and gender (past tense verbs and adjectives) or number and person (present tense verbs).
- Genitive subject: The subject is nominal in the genitive case with the predicate negated. The predicate must have default features (*3sg.N*). Only the predicate is altered here (to features other than *3sg.N*).
- Clausal subject: The subject is a clause. The predicate must have default features (*3sg.N*). Only the predicate is altered here (to features other than *3sg.N*).
- Subject with attractor: The subject is nominal or a clause, and there exists an attractor in terms of Slioussar and Malko (2016) – a nominal in the subject tree with features distinct from those of the actual subject. Only the predicate is altered here (to features that match those of the attractor) to produce an attraction error.

Implementation Details We determine the subject and the predicate via a syntactic parser from (Anastasyev, 2020). We ensure the subject and the predicate can inflect for at least some features like number, gender, or person. Using pymorphy2, we match syntactic and morphological analyses. We ensure the subject and the predicate agree according to pymorphy2 feature analysis. We then perform minimal alternation, one feature at a time, for number, gender, and person. In principle, the subject and the predicate can be alternated. However, we never alternate the subject if it controls the agreement of any other word, except for the predicate, so the change is minimal, and the phenomena are kept distinct. We ensure the changed form is not ambiguous: a changed nominal form should have no homonyms in its declension paradigm. Also, we do not alternate the predicate for gender in sentences where the subject is a proper noun or a word denoting jobs, as these can plausibly agree for either feminine or masculine. A curated list of job words from RNC is used.

B.2.3 Anaphor Agreement

Anaphor Agreement phenomenon includes errors in agreement of a relative pronoun (anaphor) with its head noun. These pronouns do not inflect for person, so there are two paradigms: incorrect (i) number or (ii) gender

Contexts For this phenomenon, we search for relative clauses with a pronoun *kotoryj* 'which,' that is not a subject of this relative clause (such nominative relative pronoun fits neither this phenomenon nor Subject-Predicate Agreeement because it always enters two agreement relations simultaneously).

Implementation Details We determine the head noun and relative pronoun via a morphosyntactic parser by Anastasyev (2020). Using pymorphy2, we match syntactic and morphological analyses. We ensure the head noun and relative pronoun agree according to pymorphy2 feature analysis. We then perform minimal alternation, one feature at a time, for number and gender. In principle, the head noun and the relative pronoun can be alternated. However, we never alternate the head noun or the pronoun if they enter several agreement relations simultaneously, so the change is minimal, and the phenomena are kept distinct.

B.2.4 Noun Phrase Agreement

Noun Phrase Agreement phenomenon includes agreement errors in the domain of the noun phrase. Adjectives and adjectival pronouns agree with their head noun; as such, the violations include errors in agreement for (i) number, (ii) gender, and (iii) case.

Contexts Here, we search for clauses with noun phrases with a single modifier, as we only alter a word in our phenomena. These modifiers could be adjectives, adjective-like pronouns, numerals, and participles that agree with their head nouns.

Implementation Details We determine the head noun and modifier via a morphosyntactic parser by Anastasyev (2020). Using `pymorphy2`, we match syntactic and morphological analyses. We ensure the head noun and modifier agree according to `pymorphy2` feature analysis. We then perform minimal alternation, one feature at a time, for number, gender, and case. In principle, the head noun and the modifier can be alternated (the head noun can be alternated for number, but not for the case as that would be the GOVERNMENT phenomenon). We never alternate the head noun if it enters several agreement relations simultaneously, so the change is minimal, and the phenomena are kept distinct.

B.2.5 FLOATING QUANTIFIER AGREEMENT

FLOATING QUANTIFIER AGREEMENT phenomenon includes errors in agreement of a floating quantifier (or "intensifier") *sam* 'self' with its antecedent head noun. The violations include incorrect (i) number, (ii) gender, and (iii) case.

Contexts For this phenomenon, we search for sentences with a floating quantifier *sam* 'self.' We determine its antecedent head noun heuristically (see below). The floating quantifier has some freedom to appear in different spots in the sentence for which we account.

Implementation Details The syntactic analysis does not connect the floating modifier to its antecedent noun. In each sentence, we heuristically search the whole clause for a single verbal argument (a subject or an object: direct, indirect, or oblique) that has all the same features as a floating quantifier: the number, the gender, and the case – this will be its antecedent head noun (highlighted brown in Example 2). (If such a noun is not found, or more than one is found, we discard the sentence). Then, using `pymorphy2`, we match syntactic and morphological analyses. We ensure the antecedent noun and modifier agree according to `pymorphy2` feature analysis. We then perform minimal alternation, one feature at a time, for number, gender, and case. In principle, the head noun and the relative pronoun can be alternated (the head noun can be alternated for number, but not for case as that would be the GOVERNMENT phenomenon). We never alternate the head noun if it enters several agreement relations simultaneously, so the change is minimal, and the phenomena are kept distinct.

(2) a. *Zhdali samogo bossa kompanii.*
'They waited for the company boss[m.sg] himself.'

b. **Zhdali samo bossa kompanii.*
'They waited for the company boss[m.sg] itself.'

B.2.6 REFLEXIVES

We only consider the case of an external possessor, a so-called *u*-phrase inside the existential *be*-possessive construction that allows a noun phrase or a personal pronoun but cannot bind a reflexive; see Example (3) (Arylova, 2013; Stassen, 2013).

Contexts We define the appropriate external possessor contexts as sentences with a *be*-verb (*byt'*, *est'*), where a noun phrase or a personal pronoun has the preposition *u* in its dependents. Additionally, we limit the contexts to those sentences where the *u*-phrase preceded the verb. This is required because noun phrases following can be used with a preposition *u* in other contexts, namely locative (e.g., *On byl* ***u*** <u>*doma*</u> 'He was **by** <u>the house</u>'). However, this interpretation is less common for cases when the *u*-phrase precedes the verb.

Implementation Details To create violations, we change the noun phrase or a pronoun to a reflexive pronoun *sebya* 'self'. Since the reflexive has no gender, number, or case features, we do not need to inflect it.

(3) a. *U nego byli druz'ya.*
'He had friends.'

b. *U sebya byli druz'ya.*
'Himself had friends.'

B.2.7 NEGATION

We implement several ways to violate the rules of *negative concord*, namely (i) shifting the negative particle *ne* from a negated verb to another word in the sentence; replacement of (ii) a negative pronoun with an indefinite one, and (iii) an indefinite pronoun with a negative one.

Contexts For this paradigm, we search sentences containing a verb under negation used with a negative pronoun (i-ii) or an indefinite pronoun used with a non-negated verb (iii). We do not consider interrogative and conditional sentences and sentences containing an imperative, as their syntactic structures differ from affirmative sentences.

Implementation Details To create violations for paradigm (i), we move the negative particle *ne* 'not' from a verb to the head of another noun, adjective, or another phrase. We ensure that the particle is moved not randomly but to specific syntactic constructions to avoid non-logical combinations of words. Such constructions can be negated in other contexts. Thus, the resulting combinations are more plausible and natural. Our systematic approach to replacing a negative pronoun with an indefinite one (and vice versa) ensures that only some replacements lead to ungrammatical sentences. We curate a list of possible replacements, which consistently lead to the violation of negative concord. This list is then systematically applied to paradigms (ii-iii), resulting in the necessary changes to the pronouns.

B.3 Semantics

B.3.1 Argument Structure

Argument Structure phenomenon includes errors in the verb's argument structure. Similarly to BLiMP, we focus on cases where the animacy requirement for the arguments of a transitive verb (from now on in this section – TV) is violated due to the verb, subject, and object replacement. Additionally, we include a more straightforward case, employing the differences between the argument structure of a transitive and an intransitive verb. Thus, the paradigms include swapping: (i) a TV with an intransitive one; an animate subject of a TV in (ii) active or (iii) passive voice with its inanimate object or replacing it with a random inanimate word; (iv) animate direct object of a TV with a random inanimate word; (v) animate indirect object of a TV with an inanimate subject, or replacing it with a random inanimate word.

Contexts We consider sentences with a transitive verb in finite form, active or passive, with an inanimate object, both direct and indirect. TVs sometimes allow inanimate subjects, typically metaphorically, so we limit allowed contexts using the RNC semantic annotation. We avoid subjects with semantics of heterogeneous groups of people (e.g., *crowd*); organizations (*bank*); events (*elections*); instruments, weapons, and their parts (*gun, bullet*); means of transport (*bus*); space, place, and time (*planet, spring*); and proper nouns (*Moscow*). For paradigm (v), we search for sentences with an open clausal complement (`xcomp`) dependent on an animate object and following the said object.

Implementation Details To generate minimal pairs for this paradigm, we filter the sentences with a transitive verb and check their dependents for the required arguments. In cases where several arguments are swapped places (paradigms ii-v), to isolate the phenomenon, we ensure that the words to be swapped do not have any modifiers, ensuring that no agreement errors appear after the perturbation. We also make sure to inflect the swapped words to preserve sentence structure. For transitivity (i), that includes replacing a verb with a verb of the same aspect, tense, number, person, and gender values. Subject and object swaps include sampling the nouns with the same number and gender features as the original. See Example (4), the TV is underlined, the original subject and object are highlighted in gray and brown, respectively. Both subject and object have the same gender category (feminine) and number (singular), so we can swap them. In the generated sentence (b), the original object *sumku* 'the bag' takes the Agent argument of the TV, which requires it to be in Nominative, so we change its case from Accusative to Nominative and do the opposite for the object *ona* 'she' (Nominative), which becomes *ee* 'her' (Accusative).

(4) a. *Ona ostavila sumku na stole.*
'She left the bag on the table.'

b. * *Sumka ostavila ee na stole.*
'The bag left her on the table.'

B.3.2 Aspect

Aspect is the grammatical category of verbs that indicates whether an action is complete (perfective) or incomplete (imperfective) at a particular time. Such semantic difference limits the contexts where each category of verbs can be used, so we employ this to generate minimal pairs for this phenomenon. We replace an imperfective verb with a perfective one in the following contexts, which do not allow a perfective verb: (i) duration; (ii) repetition; contexts with a negated deontic verb, which only allows a (iii) single or (iv) conjoined imperfective (de Haan, 2002; Paducheva, 2010).

Contexts We curate a list of words and constructions that indicate the required semantics and use them to filter the contexts. The following lexical cues are used:

- Duration (i): *dolgo*, *dlitel'no*, *prodolzhitel'no*, all with the semantics of 'continuously, for a long time'.
- Repetition (ii): *kazhdyj* 'every' + X construction, where X is a noun denoting a time period, such as *kazhdyj den'/god* 'every day/year', etc.; and adverbs like *ezhechasno* /*ezheminutno* 'ocurring every hour/minute'.
- Deontic modality (iii-iv): *stoit* and *sleduet* 'should', *nado* and *nuzhno* 'need'.

Implementation Details To generate minimal pairs, we find sentences with an imperfective verb and check its dependents for one of the lexical cues from the list. We then use a dictionary of aspect pairs (Zaliznyak, 1987) to change the verb with its perfective counterpart. Note that for some verbs, the dictionary presents several possible versions of pairs (e.g., *sbrasyvat'* 'to throw' has two perfective forms: *sbrosit'* and *sbrosat'*)). We filter the dictionary by IPM and only leave the pairs with the higher frequency.

B.3.3 Tense

The phenomenon focuses on the semantics of tense, expressed in sentences with a tense-marked verb in the presence of a temporal adverbial. We include three paradigms: incorrect choice of a (i) single or (ii) conjoined verb form in a sentence with temporal adverbial, and (iii) wrong temporal adverbial in a sentence with a tense-marked verb.

Contexts We only consider sentences with a perfective verb in future or past tense. This way, we ensure that the pairs are minimal and that the perturbations would lead to ungrammaticality. Additionally, we filter out clausal complements that are verbs to account for constructions like *sobirayus' sdelat'* "am going to do", which can be used with markers of both past and future tenses when changed.

To find sentences with the required semantics, we look for a temporal adverbial – a word or an expression that specifies the time of the event. We include several types of such expressions:

- Adverbs: simple one word expressions like *vchera* 'yesterday', *zavtra* 'tomorrow', etc. We curate a list of adverbs using RNC.
- Adpositional Phrases: `PREP + ADJ + NOUN` constructions, such as *v sleduyushchij raz* 'next time', *na proshloj nedele* 'last week', etc.
- Numerical Phrases: constructions of the type `NUM + NOUN(pl) + ADP`, e.g., *neskol'ko dnej nazad* 'a few days ago', *paru nedel' nazad* 'a couple of weeks ago', etc.

Implementation Details To introduce ungrammaticality, we find sentences that include a verb in past or future tense and check its dependents for one of the temporal adverbials from the list. We change the verb form or the temporal adverbial to the one of the 'opposite' tense (future ↔ past). Example (5) illustrates the two possible perturbations. We can either change the verb form *poletit* 'will fly' to *poletel* 'flew', or *zavtra* 'tomorrow' to *vhera* 'yesterday'. Both alterations result in ungrammatical sentences.

(5) a. *Zavtra on poletit v Italiyu.*
'Tomorrow he will fly to Italy.'

b. **Zavtra on poletel v Italiyu.*

c. * *Vchera on poletit v Italiyu.*

C Human Validation

C.1 Annotation Guidelines

Annotation Task: Verify the quality of a linguistic minimal pair

Overview Judge the correctness of a given minimal pair in which the grammatical sentence is taken from the corpus of natural texts, and the ungrammatical sentence is automatically generated using expert-written rules and natural language processing tools.

What is a minimal pair? A minimal pair consists of two sentences that differ in grammatical acceptability due to a single morphological, syntactic, or semantic feature. Please note that the minimal pair should isolate only one linguistic feature, such as number, gender, case, and more. The ungrammatical sentence is obtained by perturbing the grammatical one using one of the following operations.:

- Changing a feature, e.g., changing of one inflectional category: number, case, gender, tense, etc.
- Replacing a word, e.g., replacing a lexeme while maintaining the original grammatical form;
- Swapping two words in a sentence;
- Moving a word to another position.

Your task

1. Carefully read the grammatical and ungrammatical sentences and the linguistic feature that should be isolated.
2. Decide whether the minimal pair is designed correctly. Does it isolate the specified linguistic feature?
3. If everything is correct, select "Yes".
4. If the minimal pair is implausible, does not isolate the mentioned feature, contains two grammatical sentences, perturbs multiple sentence units or linguistic features, select "No".
5. If the original sentence is ungrammatical, select "N/A".
6. If there are any typos, please state them in the box.

Do you have any questions or difficulties with completing your task? Reach out in our group chat.

The guidelines further provide an extensive list of minimal pair examples for each paradigm and annotation examples for each answer option. You can access the complete guidelines in our GitHub repository.

Example of web interface

Minimal pair

This is a toy grammatical sentence.

*This **are** a toy ungrammatical sentence.

Phenomenon

This is the linguistic feature.

Is the minimal pair designed correctly?

○ Yes ○ No ○ N/A

Comment

Enter your comment

C.2 Data validation results

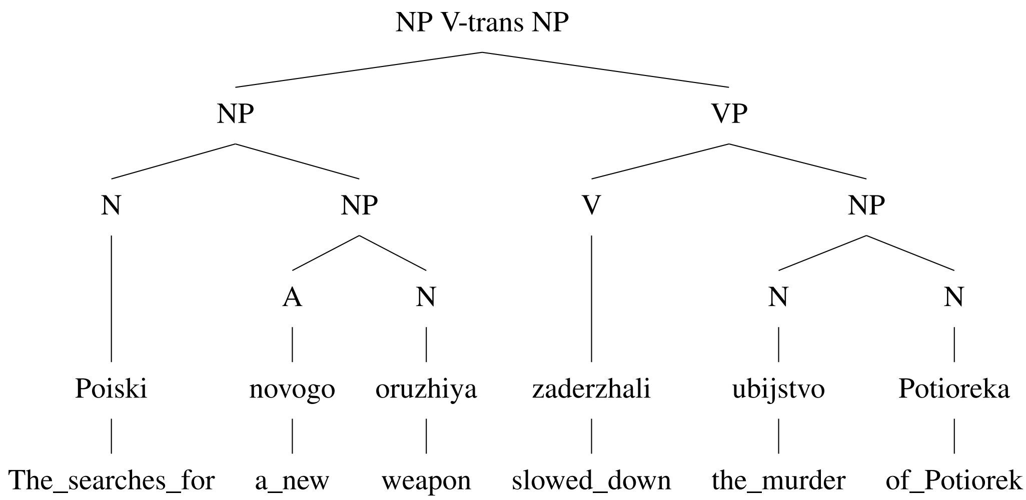
Phenomenon	Paradigm	%	WAWA
Word Formation	Addition of Extra Morphemes: Uninterpretable Suffix Combinations	93.48	91.1
	Addition of Extra Morphemes: Verb Prefixes	97.83	93.6
	Morpheme Permutation: Verb Prefixes	96.00	93.8
Word Inflection	Replacement of Inflectional Affixes: Noun Declensions (Simple)	98.00	96.0
	Replacement of Inflectional Affixes: Declensions of Nouns With Agreeing Dependents	94.00	89.7
	Inflectional Affixes: Verbal Conjugation Swap	94.00	96.0
Government	Prepositional Government	100.00	94.0
	Verbal Government: Direct Object	87.50	89.7
	Verbal Government: Genitive Object	93.62	88.8
	Verbal Government: Object in Instrumental Case	100.00	100.0
	Verbal Government: Nominalizations	78.05	86.7
Subject-predicate Agreement	Subject-Predicate Agreement (Number)	96.00	96.0
	Genitive Subject-Predicate Agreement (Number)	85.71	89.2
	Clausal Subject-Predicate Agreement (Number)	97.83	88.0
	Subject-Predicate Agreement in Presence of an Attractor (Number)	100.00	93.6
	Subject-Predicate Agreement (Gender)	97.96	94.5
	Genitive Subject-Predicate Agreement (Gender)	91.84	93.6
	Clausal Subject-Predicate Agreement (Gender)	100.00	91.5
	Subject-Predicate Agreement in Presence of an Attractor (Gender)	97.96	92.4
	Subject-Predicate Agreement (Person)	100.00	99.3
	Genitive Subject-Predicate Agreement (Person)	89.36	85.8
	Clausal Subject-Predicate Agreement (Person)	97.96	93.2
Anaphor Agreement	Anaphor Agreement (Number)	92.68	93.2
	Anaphor Agreement (Gender)	95.45	92.8
Noun Phrase Agreement	Noun Phrase Agreement (Number)	91.49	92.3
	Noun Phrase Agreement (Gender)	98.00	95.5
	Noun Phrase Agreement (Case)	100.00	95.2
Floating Quant. Agreement	Floating Quantifier Agreement (Number)	95.92	87.8
	Floating Quantifier Agreement (Gender)	97.92	96.0
	Floating Quantifier Agreement (Case)	98.00	93.3
Reflexives	External Possessor	100.00	96.5
Negation	Negative Concord	100.00	95.6
	Replacement of a Negative Pronoun with an Indefinite One	80.00	87.8
	Replacement of an Indefinite Pronoun with a Negative One	100.00	94.4
Argument Structure	Transitivity	97.67	91.4
	Animate Subject of a Transitive Verb	94.00	86.4
	Animate Subject of a Passive Verb	93.88	92.7
	Animate Direct Object of a Transitive Verb	82.00	82.4
	Animate Indirect Object of a Transitive Verb	100.00	96.8
Aspect	Incompatibility of the Perfective with the Semantics of Duration	92.00	92.7
	Impossibility of the Perfective in Repetitive Situations	97.83	91.2
	Impossibility of the Perfective Under Negated Strong Deontic Verbs	96.00	95.0
Tense	Tense	95.92	92.6
	Tense (Coordination)	87.50	89.3
	Tense Markers	97.96	94.4

Table 7: The per-paradigm ratios of plausible minimal pairs (%) and WAWA inter-annotator agreement rates.

D Statistics for Syntactic Patterns

We extract syntactic structures from a grammatical sentence's dependency tree to compute a high-level diversity w.r.t. syntactic patterns in RuBLiMP. Using expert-written rules, we linearize the dependency tree by merging its subtrees into a single constituent. We never merge the verb arguments with it and parse the main and dependent clauses similarly. We then compute the total number of unique patterns and the pattern frequency at the benchmark level. Consider Example 6 for the sentence *Poiski novogo oruzhiya zaderzhali ubijstvo Potioreka* "The searches for a new weapon slowed down the murder of Potiorek", where we extract the sentence's syntactic structure as NP V-TRANS NP (transitive verb). We provide the word translations with the articles and prepositions in the same nodes for illustration purposes.

(6)



E Human Baseline

E.1 Annotation Guidelines

Select a Grammatical Sentence

Your task

1. Carefully read two sentences.
2. Determine which of the two sentences is grammatical (a Russian native speaker would say or write like this).
3. Choose "Sentence #1" if the first sentence is grammatical, or choose "Sentence #2" otherwise.
4. If there are any typos, please state them in the box.

Below, you can find annotation examples and examples of possible grammatical errors. For clarity, we mark the sentences with a grammatical error with the "*" symbol and highlighted the word in bold. Choose the sentence that has *no* grammatical errors. If you find a given pair of sentences difficult, choose the sentence that seems *more* natural and *more* grammatically correct from your perspective.

The guidelines further provide an extensive list of minimal pair examples for each paradigm and annotation examples for each answer option. You can access the complete guidelines in our GitHub repository.

Example of web interface

Which of the two sentences has no errors?

1. This is a toy sentence #1.
2. This is a toy sentence #2.

○ Sentence #1
○ Sentence #2

Comment

Enter your comment

F Fine-grained Results

Phenomenon	PID	ruBERT-base	ruBERT-large	ruRoBERTa	ruGPT-small	ruGPT-medium	ruGPT-large	ruGPT-3.5-13B	SambaLingo	**Human**
Word Formation	add_new_suffix	78.30	79.50	95.30	94.80	95.80	94.30	97.70	80.60	100.0
	add_verb_prefix	74.40	74.80	77.70	72.90	81.30	74.40	86.90	61.50	100.0
	change_verb_prefixes_order	93.00	94.20	96.00	97.90	98.20	99.00	98.40	97.50	100.0
Word Inflection	change_declension_ending	83.50	84.40	92.00	90.40	85.60	90.80	95.30	86.80	100.0
	change_declension_ending_has_dep	86.50	88.10	94.90	94.30	86.80	95.30	97.40	92.80	100.0
	change_verb_conjugation	84.60	85.60	88.00	90.00	86.70	88.00	92.90	77.60	98.00
Government	adposition_government	95.80	96.40	97.80	94.60	95.80	95.40	97.30	90.30	100.0
	verb_acc_object	93.80	93.50	96.10	92.20	93.60	93.90	97.10	89.70	100.0
	verb_gen_object	93.20	94.10	94.30	88.30	91.90	90.90	95.70	77.10	98.00
	verb_ins_object	78.70	77.30	99.20	94.50	96.60	96.60	98.00	96.50	100.0
	nominalization_case	92.10	92.00	96.00	95.10	96.50	96.10	97.40	92.40	96.00
Subject-predicate Agreement	noun_subj_predicate_agreement_number	91.70	92.70	95.40	90.30	92.20	92.20	96.00	86.60	98.00
	genitive_subj_predicate_agreement_number	95.30	95.60	96.00	95.60	96.20	96.70	97.90	82.90	97.96
	clause_subj_predicate_agreement_number	90.60	89.60	89.80	91.40	93.80	91.90	96.00	73.50	97.96
	subj_predicate_agreement_number_attractor	92.70	92.30	96.50	90.50	91.40	92.10	96.20	87.60	100.0
	noun_subj_predicate_agreement_gender	83.00	83.60	88.40	81.50	83.70	84.00	90.90	80.20	98.00
	genitive_subj_predicate_agreement_gender	97.10	97.40	97.00	96.40	96.00	96.80	98.60	89.50	98.00
	clause_subj_predicate_agreement_gender	97.00	96.10	94.00	94.30	95.30	94.60	97.90	79.90	100.0
	subj_predicate_agreement_gender_attractor	88.40	89.50	92.60	84.90	86.80	87.10	94.90	83.90	98.00
	noun_subj_predicate_agreement_person	86.40	86.90	93.40	86.10	86.40	87.00	94.60	79.40	100.0
	genitive_subj_predicate_agreement_person	87.80	89.60	92.60	92.50	92.50	93.10	97.90	78.90	98.00
	clause_subj_predicate_agreement_person	92.80	92.40	94.10	90.10	93.00	91.10	96.40	66.90	97.96
Anaphor Agreement	anaphor_agreement_number	84.10	83.70	93.20	92.70	93.80	93.90	95.30	87.50	98.00
	anaphor_agreement_gender	87.70	89.00	98.00	95.90	98.00	97.70	98.80	98.40	98.00
Noun Phrase Agreement	np_agreement_number	82.80	84.90	84.20	94.70	97.20	96.70	98.60	90.60	100.0
	np_agreement_gender	79.50	80.90	97.00	93.40	96.20	95.10	97.40	83.20	98.00
	np_agreement_case	88.40	88.30	85.30	97.90	98.70	98.50	99.40	95.70	98.00
Floating Quant. Agreement	floating_quantifier_agreement_number	83.30	85.20	96.60	89.50	93.60	93.00	98.10	83.20	100.0
	floating_quantifier_agreement_gender	95.40	94.30	93.30	79.80	96.50	83.30	97.70	97.60	98.00
	floating_quantifier_agreement_case	95.50	94.20	98.60	93.60	98.40	96.20	98.70	90.50	100.0
Reflexives	external_posessor	78.70	81.00	91.10	83.20	79.90	87.80	94.70	96.20	98.00
Negation	negative_concord	99.50	99.20	99.70	99.90	99.90	99.90	100.0	99.80	100.0
	negative_pronoun_to_indefinite	33.90	47.40	71.10	19.90	41.80	36.40	62.60	33.40	100.0
	indefinite_pronoun_to_negative	99.90	99.80	98.70	99.70	99.90	99.50	100.0	99.70	100.0
Argument Structure	transitive_verb	96.50	96.40	98.60	93.00	95.40	95.40	98.60	77.90	100.0
	transitive_verb_subject	83.60	85.40	81.10	79.00	83.50	84.40	90.30	74.60	100.0
	transitive_verb_passive	90.00	90.50	93.10	89.90	94.10	93.80	98.20	91.40	100.0
	transitive_verb_object	88.00	87.20	93.60	94.30	96.50	97.10	98.40	86.60	100.0
	transitive_verb_iobject	84.50	86.50	91.80	87.90	90.40	90.50	96.20	83.20	100.0
Aspect	change_duration_aspect	96.20	96.50	97.10	92.60	94.60	94.60	97.00	85.40	100.0
	change_repetition_aspect	95.20	95.50	97.00	94.30	95.60	95.30	97.80	91.00	100.0
	deontic_imperative_aspect	96.80	97.40	97.50	94.90	96.60	96.70	98.50	85.80	100.0
Tense	single_verb_tense	85.00	86.80	87.80	76.30	82.60	81.70	90.70	69.30	100.0
	conj_verb_tense	93.00	93.00	96.90	92.60	94.60	94.60	98.50	85.40	100.0
	tense_marker	83.50	84.70	92.50	84.00	88.90	84.80	96.90	86.70	98.00
	Average	87.95	88.74	93.13	89.28	91.62	91.29	95.86	84.56	99.15

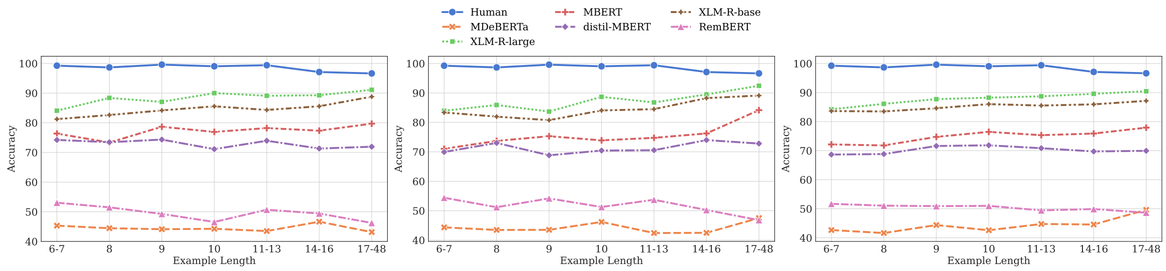
Table 8: Accuracy scores (%) for the monolingual LMs by paradigm. Random baseline is 50%.

Phenomenon	PID	distil-MBERT	MBERT	XLM-R$_{base}$	XLM-R$_{large}$	RemBERT	MDeBERTa	mGPT-1.3B	mGPT-13B	**Human**
Word Formation	add_new_suffix	86.70	91.20	93.20	94.10	48.30	49.40	97.20	97.20	100.0
	add_verb_prefix	81.50	86.70	82.20	81.30	53.20	62.30	87.30	87.30	100.0
	change_verb_prefixes_order	83.70	88.60	90.30	91.00	52.70	46.00	98.60	99.10	100.0
Word Inflection	change_declension_ending	79.80	81.30	89.40	90.20	56.30	43.80	92.70	92.30	100.0
	change_declension_ending_has_dep	86.10	89.50	93.90	94.80	58.80	40.00	97.40	97.30	100.0
	change_verb_conjugation	73.00	83.10	88.40	88.10	49.00	47.10	97.80	97.00	98.00
Government	adposition_government	80.00	85.90	91.10	92.00	55.10	47.60	92.70	93.20	100.0
	verb_acc_object	63.20	67.30	82.50	85.90	43.40	48.60	84.90	89.90	100.0
	verb_gen_object	57.00	64.90	82.20	84.70	47.70	31.70	83.60	87.40	98.00
	verb_ins_object	72.20	89.70	94.30	96.80	44.20	49.80	95.20	95.70	100.0
	nominalization_case	81.80	86.60	92.00	93.20	54.10	59.80	91.80	94.20	96.00
Subject-predicate Agreement	noun_subj_predicate_agreement_number	74.50	79.20	87.70	89.30	51.70	34.80	88.60	89.70	98.00
	genitive_subj_predicate_agreement_number	85.70	86.90	91.10	92.00	45.70	43.00	95.50	96.20	97.96
	clause_subj_predicate_agreement_number	72.50	82.80	87.90	80.70	38.40	26.00	94.60	95.60	97.96
	subj_predicate_agreement_number_attractor	69.10	76.90	88.90	89.90	49.10	40.10	84.50	87.20	100.0
	noun_subj_predicate_agreement_gender	72.10	72.70	82.00	84.20	52.50	36.50	82.00	83.80	98.00
	genitive_subj_predicate_agreement_gender	86.80	92.10	94.50	94.80	42.60	51.40	98.10	98.20	98.00
	clause_subj_predicate_agreement_gender	74.50	82.90	89.90	86.50	43.80	27.10	87.60	88.00	100.0
	subj_predicate_agreement_gender_attractor	70.30	78.50	85.10	86.50	54.70	34.40	84.20	86.40	98.00
	noun_subj_predicate_agreement_person	67.30	67.50	84.40	85.10	52.70	39.60	74.20	76.90	100.0
	genitive_subj_predicate_agreement_person	80.50	76.00	86.20	88.40	62.20	39.20	82.70	82.70	98.00
	clause_subj_predicate_agreement_person	81.60	88.60	85.30	87.60	54.10	32.40	92.60	91.50	97.96
Anaphor Agreement	anaphor_agreement_number	74.70	82.30	89.60	89.80	45.30	65.40	90.80	92.40	98.00
	anaphor_agreement_gender	30.00	90.40	94.10	96.50	18.80	85.30	93.50	96.30	98.00
Noun Phrase Agreement	np_agreement_number	78.80	88.40	93.20	94.70	48.20	51.60	95.30	96.20	100.0
	np_agreement_gender	69.10	80.10	88.90	91.60	55.70	26.50	89.50	91.10	98.00
	np_agreement_case	90.40	92.70	95.90	96.80	49.60	45.00	97.80	98.70	98.00
Floating Quant. Agreement	floating_quantifier_agreement_number	71.00	77.20	88.40	90.40	57.40	40.20	88.20	90.70	100.0
	floating_quantifier_agreement_gender	88.40	83.30	94.90	95.70	79.10	33.40	74.80	83.00	98.00
	floating_quantifier_agreement_case	90.00	87.80	95.60	93.10	51.40	38.70	92.40	91.80	100.0
Reflexives	external_posessor	56.00	52.90	69.90	79.10	45.30	40.20	82.50	85.60	98.00
Negation	negative_concord	96.30	97.90	99.20	98.80	54.60	55.30	99.80	99.70	100.0
	negative_pronoun_to_indefinite	50.80	17.30	19.10	43.90	93.00	54.90	3.90	5.40	100.0
	indefinite_pronoun_to_negative	78.70	83.70	99.20	99.30	5.90	20.50	99.70	99.70	100.0
Argument Structure	transitive_verb	62.40	71.70	84.80	89.40	46.50	38.40	82.20	86.00	100.0
	transitive_verb_subject	53.80	56.00	64.80	69.60	49.60	42.70	70.40	73.70	100.0
	transitive_verb_passive	63.60	69.40	79.70	85.50	44.80	39.40	87.00	91.20	100.0
	transitive_verb_object	48.40	53.60	73.80	80.70	52.40	45.40	80.20	86.00	100.0
	transitive_verb_iobject	51.20	55.40	74.30	81.30	53.10	43.60	79.90	85.40	100.0
Aspect	change_duration_aspect	58.80	62.20	80.70	87.50	54.90	43.80	83.00	85.00	100.0
	change_repetition_aspect	58.70	62.90	80.40	87.70	48.10	47.20	86.90	89.00	100.0
	deontic_imperative_aspect	60.90	54.20	83.60	87.90	54.20	41.30	86.40	89.10	100.0
Tense	single_verb_tense	63.10	58.00	66.40	72.40	54.00	48.90	65.70	71.50	100.0
	conj_verb_tense	71.60	70.20	79.30	86.60	47.80	55.50	83.00	87.10	100.0
	tense_marker	32.80	29.90	78.30	80.30	54.80	56.20	89.90	90.50	98.00
	Average	70.65	75.03	84.81	87.46	50.55	44.22	86.37	88.26	99.15

Table 9: Accuracy scores (%) for the multilingual LMs by paradigm (part 1). Random baseline is 50%.

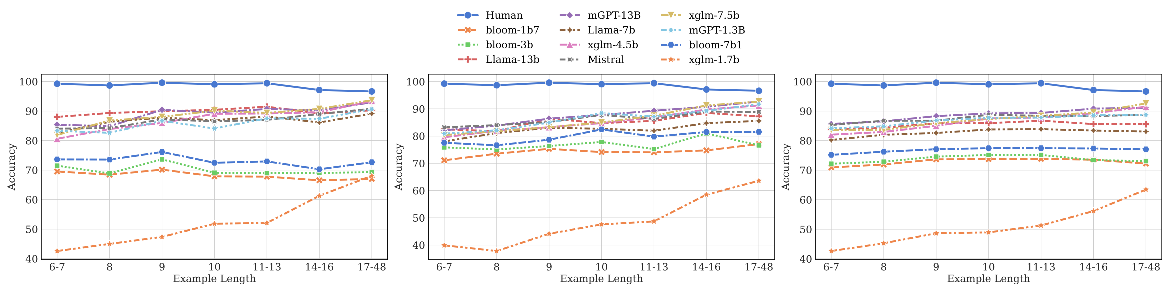
Phenomenon	PID	bloom-1b7	bloom-3b	bloom-7b1	xglm-1.7B	xglm-4.5B	xglm-7.5B	Llama-7b	Llama-13b	Mistral	**Human**
Word Formation	add_new_suffix	90.30	90.90	91.40	23.80	96.20	96.60	93.30	94.90	96.50	100.00
	add_verb_prefix	90.50	92.10	92.30	17.20	82.00	83.10	92.10	93.40	95.40	100.00
	change_verb_prefixes_order	77.50	85.60	83.00	72.10	99.00	98.70	98.70	99.20	98.70	100.00
Word Inflection	change_declension_ending	86.00	85.40	87.80	43.10	91.30	92.20	90.50	92.10	93.50	100.0
	change_declension_ending_has_dep	90.20	91.40	93.90	47.90	95.30	96.20	93.30	96.20	97.00	100.0
	change_verb_conjugation	92.90	93.90	93.90	44.10	89.90	91.90	88.70	92.20	94.50	98.00
Government	adposition_government	70.00	73.50	77.00	56.90	90.40	91.80	88.40	90.90	92.50	100.0
	verb_acc_object	65.80	69.30	71.50	43.90	81.80	82.30	85.70	88.70	87.50	100.0
	verb_gen_object	60.80	64.30	63.90	35.40	81.60	79.30	69.70	73.80	77.40	98.00
	verb_ins_object	65.30	69.70	71.40	68.90	93.60	95.10	90.30	93.50	87.50	100.0
	nominalization_case	77.40	80.40	84.30	53.50	92.40	93.80	91.90	95.60	95.90	96.00
Subject-Predicate Agreement	noun_subj_predicate_agreement_number	80.90	79.50	85.00	45.20	87.20	87.80	85.40	86.00	90.40	98.00
	genitive_subj_predicate_agreement_number	89.10	89.90	91.70	48.40	89.10	90.90	90.00	91.40	95.80	97.96
	clause_subj_predicate_agreement_number	93.50	95.00	95.00	31.90	79.40	83.80	95.60	96.50	96.90	97.96
	subj_predicate_agreement_number_attractor	75.10	76.20	83.20	56.90	84.50	85.90	84.20	86.50	87.00	100.0
	noun_subj_predicate_agreement_gender	70.50	72.00	74.90	44.90	79.00	78.80	86.30	88.20	90.00	98.00
	genitive_subj_predicate_agreement_gender	95.50	95.80	94.10	51.30	91.20	91.90	95.70	96.70	96.60	98.00
	clause_subj_predicate_agreement_gender	91.80	94.60	94.70	45.60	88.80	89.70	95.20	96.10	96.70	100.0
	subj_predicate_agreement_gender_attractor	69.80	73.10	75.80	51.90	81.40	81.20	84.90	87.60	88.00	98.00
	noun_subj_predicate_agreement_person	85.10	87.20	91.10	43.60	76.00	76.90	82.70	86.50	87.30	100.0
	genitive_subj_predicate_agreement_person	93.10	93.00	94.80	37.20	72.30	73.50	89.30	91.70	97.00	98.00
	clause_subj_predicate_agreement_person	96.70	97.00	97.70	30.00	80.80	80.80	94.70	96.30	97.20	97.96
Anaphor Agreement	anaphor_agreement_number	69.70	67.90	74.40	57.80	90.50	90.70	74.20	78.50	82.80	98.00
	anaphor_agreement_gender	69.80	63.80	71.80	73.60	95.10	96.20	22.50	33.50	61.40	98.00
Noun Phrase Agreement	np_agreement_number	75.00	78.80	83.80	56.80	92.30	93.00	91.80	94.50	94.80	100.0
	np_agreement_gender	76.40	76.00	80.10	55.80	88.40	90.60	81.10	84.20	87.80	98.00
	np_agreement_case	85.90	88.50	90.00	72.10	97.40	97.50	94.80	95.90	97.00	98.00
Floating Quant. Agreement	floating_quantifier_agreement_number	79.90	77.70	80.80	47.10	85.10	87.60	87.00	89.80	90.90	100.0
	floating_quantifier_agreement_gender	48.80	52.80	64.50	50.50	94.70	94.00	54.30	62.00	90.30	98.00
	floating_quantifier_agreement_case	71.90	74.50	80.80	44.10	94.10	91.50	75.00	79.10	82.30	100.0
Reflexives	external_posessor	13.10	19.30	23.30	64.10	82.90	90.80	84.80	89.00	32.40	98.00
Negation	negative_concord	98.30	98.80	99.00	75.50	100.00	100.00	99.50	99.60	99.80	100.0
	negative_pronoun_to_indefinite	12.40	16.10	11.70	2.60	21.90	23.30	20.80	25.20	19.60	100.0
	indefinite_pronoun_to_negative	84.90	88.60	94.50	85.80	100.00	100.00	96.90	97.20	97.80	100.0
Argument Structure	transitive_verb	74.80	74.80	76.20	28.10	83.50	83.00	81.70	85.90	87.20	100.0
	transitive_verb_subject	56.30	56.70	58.40	33.30	71.40	72.60	70.10	73.40	73.70	100.0
	transitive_verb_passive	56.30	54.30	60.10	54.10	89.90	91.60	90.30	92.40	91.10	100.0
	transitive_verb_object	34.80	38.90	43.50	38.60	88.00	87.20	82.30	83.90	83.20	100.0
	transitive_verb_iobject	46.70	46.40	50.00	36.90	80.30	81.20	75.40	79.80	81.20	100.0
Aspect	change_duration_aspect	50.00	48.20	53.20	74.10	91.00	90.40	81.40	87.00	84.60	100.0
	change_repetition_aspect	57.00	57.40	61.30	74.90	90.60	91.50	86.40	91.50	91.10	100.0
	deontic_imperative_aspect	57.00	47.80	50.70	76.80	89.80	88.90	75.80	78.00	84.10	100.0
Tense	single_verb_tense	75.50	76.60	85.80	49.70	71.40	75.00	78.90	84.40	84.80	100.0
	conj_verb_tense	73.50	75.60	83.70	50.60	87.20	88.50	86.90	91.30	92.40	100.0
	tense_marker	57.60	64.70	62.80	55.50	83.10	83.60	80.00	81.20	87.20	98.00
	Average	71.85	73.20	76.20	50.05	86.04	86.91	83.08	86.03	87.04	99.15

Table 10: Accuracy scores (%) for the monolingual LMs by paradigm (part 2). Random baseline is 50%.



(a) Wikipedia (b) Wikinews (c) Librusec

Figure 5: Results on RuBLiMP for the multilingual encoder-only LMs per domain grouped by seven quintiles of the length.



(a) Wikipedia (b) Wikinews (c) Librusec

Figure 6: Results on RuBLiMP for the multilingual decoder-only LMs per domain grouped by seven quintiles of the length.

G Multilingual Experiments

G.1 Experimental Setup

We evaluate 17 multilingual LMs on six benchmarks as shown in §3. The benchmarks can be characterized by the minimal pair generation method: (i) using a dictionary and linguistic templates (BLiMP), (ii) translating an English dictionary and adapting the linguistic templates (CLiMP, CLAMS), (iii) collecting examples from linguistic publications (JBLiMP), (iv) extracting sentences from a Universal Dependencies treebank and using linguistic templates (SLING), and (v) extracting sentences from open text corpora, using linguistic perturbations, and decontaminating test data (ours). The benchmark details are given below:

- BLiMP (Warstadt et al., 2020) comprises 67 paradigms for English, 1k minimal pairs each. It covers 12 representative phenomena in English, including anaphor agreement, argument structure, binding, control/raising, determiner-noun agreement, ellipsis, filler gap dependencies, irregular verb forms, island effects, NPI licensing, quantifiers, and subject-verb agreement.
- CLiMP (Xiang et al., 2021) includes 16 paradigms nine phenomena in Chinese, such as anaphor agreement, binding, argument structure, and classifier-noun agreement.
- SLING (Song et al., 2022) includes nine high-level linguistic phenomena in Mandarin Chinese, present in CLiMP (e.g., anaphor agreement, classifier-noun agreement, binding) and new ones (aspect, polarity items, relative clauses, and wh-fronting, among others).
- JBLiMP (Someya and Oseki, 2023) comprises 11 phenomena in Japanese: argument structure, binding, control/raising, ellipsis, filler gap dependencies, island effects, morphology, nominal structures, NPI licensing, verbal agreement, and quantifiers.
- CLAMS (Mueller et al., 2020) is a syntactic evaluation suite in five languages (English, Russian, French, German, and Hebrew) that covers different paradigms of subject-verb agreement.

G.2 Results

The results are summarized in Table 11. Overall, we find that `RemBERT` and `MDeBERTa` perform at the level of a random baseline on all benchmarks. We also observe an unsatisfactory performance of most decoder-only LMs on CLAMS (Hebrew) and JBLiMP (Japanese), with the scores ranging between approx. 50% (`xglm-1.7B`) to 70.6% (`xglm-4.5B`). No single LM performs consistently well in all languages.

Larger $\neq$ Better Similar to our findings on RuBLiMP (§5), the LMs' performance does not always improve with the number of parameters, e.g.: `XLMR` (BLiMP, SLING, CLAMS), `mGPT` (BLiMP, CLiMP, SLING), and `bloom` (CLiMP and SLING).

Sensitivity to Agreement For a more fine-grained analysis, we select AGREEMENT as one of the most well-represented phenomena in all considered benchmarks. We report the results in Table 12 and describe them by phenomenon and language. The general trend here is that model performance in a given language depends on the benchmark. In particular, the Δ-scores between the benchmarks for the SUBJECT-PREDICATE AGREEMENT in Russian can range from 2.4% (`distil-MBERT`) to 37% (`xglm-1.7B`). However, some LMS perform consistently w.r.t. this phenomenon on both RuBLiMP and BLiMP (e.g., `bloom`, `xglm`, `MBERT`). The LMs identify the ANAPHOR AGREEMENT contrast reliably on BLiMP and demonstrate lower performance on CLAMS, with the Δ-score in the range between 2.72% and 15.03% For Chinese, the Δ-scores vary between 4% and 21%. We assume that the result differences are attributed to the minimal pair generation method and quality, which is analyzed in detail for SLING and CLiMP (Song et al., 2022). We provide the results of the CLAMS' manual analysis below.

Now, we focus on the performance analysis for Chinese and Russian since both languages have benchmarks created through the translation of an English vocabulary and linguistic templates (CLiMP and CLAMS) and usage of open text corpora, linguistic resources, and linguistic perturbations (SLING and RuBLiMP).

CLiMP vs SLING We find that the decoder-only LMs generally perform worse on SLING, with the accuracy Δ-score of up to 12% (e.g., `xglm` and `bloom`). A high-level analysis indicates that SLING does overcome the limitations of CLiMP and represents a more challenging benchmark of linguistic minimal pairs for Chinese. We refer the reader to Song et al. (2022) for a detailed comparison of these two evaluation resources.

Model	RuBLiMP	BLiMP	CLiMP	SLING	JBLiMP	CLAMS					Avg.
	ru	en	zh	zh	ja	en	ru	de	fr	he	
distil-MBERT	70.65	66.49	69.07	75.25	59.72	69.42	73.52	84.50	75.04	65.12	70.88
MBERT	75.03	68.45	72.95	74.01	65.49	66.95	83.33	84.09	75.96	66.57	73.28
XLM-R$_{base}$	84.81	76.69	72.84	73.63	66.28	73.25	78.02	86.40	70.16	68.66	75.07
XLM-R$_{large}$	87.46	79.60	74.85	73.72	71.96	82.09	78.17	87.62	72.53	75.66	78.37
RemBERT	50.55	45.42	51.61	48.34	43.57	47.38	51.51	40.28	46.48	49.05	47.42
MDeBERTa	44.22	54.02	49.45	48.72	44.84	54.28	60.72	56.45	64.53	52.20	52.94
mGPT-1.3B	86.37	76.21	75.85	68.80	71.86	73.31	76.31	91.40	75.76	66.81	76.27
mGPT-13B	88.26	76.45	76.21	72.30	63.29	83.15	77.56	92.65	80.18	68.72	77.88
bloom-1b7	71.85	78.16	73.44	66.00	49.91	79.69	73.49	65.24	82.95	53.04	69.38
bloom-3b	73.20	78.69	74.25	64.67	60.47	81.87	76.18	67.33	83.85	58.50	71.90
bloom-7b1	76.20	79.54	73.66	66.80	65.83	84.70	79.72	69.53	86.37	56.08	73.84
xglm-1.7B	50.05	78.56	77.01	65.18	72.87	75.94	81.95	91.87	79.24	52.60	72.53
xglm-4.5B	86.04	77.94	76.07	67.36	71.06	75.18	83.53	91.75	81.60	70.58	78.11
xglm-7.5B	86.91	78.99	77.83	66.49	73.77	76.57	83.72	93.22	81.81	52.00	77.13
Llama-7b	83.08	79.46	63.89	74.75	70.36	78.14	80.73	89.52	84.39	53.96	75.83
Llama-13b	86.03	79.11	64.53	75.32	69.20	78.79	82.62	87.19	82.59	54.08	75.95
Mistral	87.04	80.66	72.03	79.51	69.15	86.01	87.04	84.01	82.91	56.78	78.51

Table 11: Accuracy scores (%) for the multilingual experiments on RuBLiMP, BLiMP, CLiMP, SLING, JBLiMP, and CLAMS. Random baseline is 50%. The line separates the encoder-only and decoder-only LMs.

CLAMS vs RuBLiMP We are interested in analyzing the LMs' performance differences on CLAMS and RuBLiMP in more detail. Three authors of this paper conduct a manual analysis of 50 random examples in CLAMS (approx. 17 examples per author) and the paper's appendices (Mueller et al., 2020). The results show there are:

1. 60% of plausible minimal pairs; the minimum length is 2 tokens (e.g., *Khudozhnik stariy/*stariye* "The painter is/*are old").

2. 20% of semantically implausible or uninterpretable pairs (e.g., *Vrachi, kotorykh lidery hotyat/*hochet, bol'schiye* "The doctors that the leaders want/*wants are big").

3. 15% of pairs do not isolate a target phenomenon, which means that the grammatical sentence is implausible or the ungrammatical sentence can have multiple errors. E.g., *Klienty govoryat i zhdali/*zhdal* "The clients are speaking and were/*was waiting". Here, the tense concord rules are violated in the grammatical sentence, which leads to the perturbation of both number and tense verb forms in the ungrammatical sentence).

4. 5% of pairs contain repetitive constructions or abruptly break off (e.g., *Senator lyubit smotret' teleperedachi and lyubit/*lyubyat smotret' teleperedachi* "The senator likes to watch TV and likes/*like to watch TV").

The primary reason behind these errors is that the word vocabulary is translated from English, and the contextual ambiguity is not controlled. There are 126 unique tokens (including the punctuation marks) in the 40.1k grammatical sentences in CLAMS, which significantly limits the diversity of the minimal pairs. Besides, some minimal pairs are plausible from the perspective of well-formedness and acceptability. However, a native Russian speaker – at least the authors performing the analysis – is unlikely to say or write a sentence this way. We conclude that these factors contribute to the performance differences.

Model	RuBLiMP			BLiMP			CLiMP	SLING	CLAMS					
	ru			en			zh	zh	en		ru	de	fr	he
	SPA	AA	NPA	SPA	AA	DNA	AA	AA	SPA	AA	SPA	SPA	SPA	SPA
distil-MBERT	75.90	52.35	79.43	75.30	94.70	87.11	83.40	83.37	67.83	85.26	73.52	84.50	75.04	65.12
MBERT	80.37	86.35	87.07	80.28	89.75	88.45	73.00	89.11	65.70	79.54	83.33	84.09	75.96	66.57
XLM-R$_{base}$	87.55	91.85	92.67	80.08	92.00	90.46	82.10	75.35	72.55	80.32	78.02	86.40	70.16	68.66
XLM-R$_{large}$	87.73	93.15	94.37	83.42	94.95	92.98	78.00	67.17	81.48	88.22	78.17	87.62	72.53	75.66
RemBERT	49.77	32.05	51.17	50.18	49.30	45.01	52.90	36.95	47.46	46.58	51.51	40.28	46.48	49.05
MDeBERTa	36.77	75.35	41.03	53.79	45.45	48.54	72.00	36.31	54.67	50.32	60.72	56.45	64.53	52.20
mGPT-1.3B	87.69	92.15	94.2	79.20	98.00	88.80	86.80	65.67	71.97	86.66	76.31	91.40	75.76	66.81
mGPT-13B	88.75	94.35	95.33	73.55	98.95	87.80	86.90	70.29	82.71	87.56	77.56	92.65	80.18	68.72
bloom-1b7	85.55	69.75	79.10	85.50	98.40	93.20	61.80	57.77	78.58	90.76	73.49	65.24	82.95	53.40
bloom-3b	86.66	65.85	81.10	84.80	98.80	91.90	62.40	56.27	81.02	90.40	76.18	67.33	83.85	58.50
bloom-7b1	88.90	73.1	84.63	85.80	99.30	93.50	62.10	60.07	84.03	91.41	79.72	69.53	86.37	56.08
xglm-1.7B	44.26	65.7	61.57	84.50	99.60	89.90	77.60	58.72	75.08	84.57	81.95	91.87	79.24	52.60
xglm-4.5B	82.70	92.8	92.70	84.30	99.10	89.80	78.30	60.69	74.19	85.11	83.53	91.75	81.60	70.58
xglm-7.5B	83.75	93.45	93.70	83.90	99.50	90.50	81.70	59.54	75.75	84.8	83.72	93.22	81.81	52.00
Llama-7b	89.45	48.35	89.23	74.50	99.45	91.15	63.30	79.48	76.52	94.40	80.73	89.52	84.39	53.96
Llama-13b	91.23	56.00	91.53	78.20	99.50	90.32	64.60	79.09	77.65	90.19	82.62	87.19	82.59	54.08
Mistral	92.99	72.10	93.20	76.60	99.55	91.39	91.00	86.87	85.47	91.41	87.04	84.01	82.91	56.78

Table 12: Results of the multilingual model evaluation on the agreement phenomena. **Phenomena**: SPA – Subject-Predicate agreement, AA – Anaphor Agreement, NPA – Noun-Phrase Agreement, DNA – Determiner-Noun Agreement.