A Comparative Study of Synonym Generation in Russian with Instruction-Tuned Large Language Models in a Zero-Shot Setting

Maria Khokhlova
St Petersburg State University
Russia
m.khokhlova@spbu.ru

Abstract

The article presents the results of applying modern large language models to the task of generating synonyms for Russian verbs. The aim of the study is to evaluate the quality and characteristics of the generated synonyms depending on the model type and decoding parameters. Two models were employed for synonym generation: the general-purpose multilingual model Mistral and the Russian-language model T-lite.

The experimental material consisted of 18 motion verbs, for each of which synonyms were generated using an instruction-based approach. Multiple runs were performed at different temperature settings. The linguistic analysis of the results was complemented by the computation of quantitative metrics, including the average number of generated synonyms, diversity and stability coefficients, as well as the average frequency of lexical item repetition across runs.

The findings indicate a clear dependence of generation outcomes on temperature. At low temperature values, the models exhibit high stability but limited diversity of the generated synonyms. In contrast, increasing the temperature leads to greater variability and a reduction in repetition. A comparative analysis shows that T-lite produces significantly less noise, with generated lexemes being semantically closer to the target verbs, and better captures the specific features of the Russian language. The results can be applied to tasks such as lexical substitution and the automatic expansion of lexical resources.

Keywords: synonym generation, large language models (LLMs), lexical substitution, Russian language

1 Introduction

Large language models (LLMs) are applied to a wide range of tasks and demonstrate strong performance, particularly in generating texts. At the same time, this raises the question of their applicability to more specialized tasks, such as automatic synonym generation. Synonymy is one of the key lexical relations connecting linguistic units within a language system. The results produced by such models may have broad applications in linguistics, for example in the creation of exercises for language learners, text simplification, style transfer, summarization, or paraphrase generation. The aim of this paper is to compare the performance of different types of LLMs on the task of synonym generation focusing on the quality of the output.

The task of synonym generation belongs to a relatively recent direction in natural language processing and involves automatic selection of words or multi-word expressions that are semantically close to a given item. This naturally raises the first question of whether such models can be effectively applied to synonym generation, thereby supporting lexicographic work. This issue is particularly relevant for morphologically rich languages such as Russian. In addition, the presence of extensive polysemy and homonymy may introduce further challenges and errors in automatic synonym generation. A second research question concerns whether different models produce different results, and if so, what the nature of these differences is.

The paper is organized as follows. The first section motivates the relevance of the study and formulates its objectives. The second section provides a brief review of related work. The third section discusses the obtained results. The final section summarizes the study and outlines directions for future research.

2 Related Work

A great deal of attention has been paid to synonymy in linguistic research, as its accurate description is of fundamental importance both theoretically and practically. Synonym generation is closely related to the historically earlier task of lexical substitution, i.e., finding an appropriate replacement for a word in a given context. Within the SemEval-2007 competition, the task formulation and dataset description were introduced in (McCarthy and Navigli, 2007; McCarthy and Navigli, 2009). Participants were asked to replace target ambiguous words in English contexts, performing three subtasks: (1) selecting the most appropriate substitution; (2) identifying the top 10 candidate substitutions; and (3) determining whether a word is part of a multiword expression. The results produced by systems were compared against annotations provided by human annotators. It was shown that human-generated substitutions are more accurate than those produced by automated systems.

The development of transformer-based language models substantially changed the lexical substitution task. The BERT model (Devlin et al., 2019) showed that bidirectional contextual representations capture semantic information much more precisely than distributional approaches, enabling better contect-based prediction of lexical alternatives. Zhou et al. (2019) proposed one of the first BERT-based methods for lexical substitution, showing that pretrained contextual language models can generate high-quality substitute candidates.

With the emergence of neural network technologies, new models have been proposed for substitution tasks in text, which necessitates the evaluation of more recent algorithms. In particular, a comparison of approaches based on context2vec, ELMo, BERT, and XLNet is presented in (Arefyev et al., 2020). The results show the promise of neural models, while the authors also analyze types of semantic relations between a target word and its proposed substitutions.

A further breakthrough came with the emergence of large autoregressive language models. Brown et al. (2020) demonstrated that LLMs are capable of solving a wide range of downstream tasks with zero-shot and few-shot settings using only prompts, without task-specific fine-tuning. This paradigm shift laid the foundation for subsequent research on instruction tuning and prompted the use of LLMs for various tasks, including lexical generation.

Further studies made it possible to directly compare instruction-tuned models without additional task-specific training (Shi et al., 2024). Wei et al. (2021) demonstrated that instruction tuning substantially improves model performance on previously unseen tasks in the zero-shot setting. At the same time, instruction-tuned models remain sensitive to the formulation of the prompt, and their performance may degrade considerably even when semantically equivalent instructions are used (Sun, Shaib and Wallace, 2023).

Recent surveys emphasize the growing role of LLMs in semantic processing and lexical-semantic tasks (Mao et al., 2024). At the same time, the evaluation of lexical abilities of LLMs remains an active research area, with the benchmarks proposed for assessing lexical instruction following and lexical control (Ren et al., 2026).

In general, synonym selection can be viewed as the task of identifying words that can replace a given lexical unit in a specific context. The primary requirement for candidate words is the preservation of the meaning of the “corrected” text. At the same time, a narrower formulation of the task may reduce it to finding synonyms for isolated words, i.e., without contextual information. In this case, synonym identification may rely on synonym sets or hyponymy relations. Particular attention should be paid to the evaluation of proposed candidate words, which can be carried out using lexicographic resources (e.g., dictionaries or the WordNet thesaurus), expert annotation, or distributional models (vector representations) based on semantic similarity.

3 Methodology

3.1 Data

As the dataset, 18 motion verbs (denoting movement in space) were selected from the Academic Grammar of the Russian Language (Shvedova, 1980). These verbs are also attested in dictionaries and exhibit a large number of related lexical units forming synonymic sets with them. The study considers verbs characterized by single occurrence and unidirectional motion. These include: «бежать» (*bezhat’*) ‘to run’, «брести»

(*bresti*) ‘to trudge/wade’, «везти» (*vezti*) ‘to transport (by vehicle)’, «вести» (*vesti*) ‘to lead’, «гнать» (*gnat’*) ‘to drive/chase’, «гнаться» (*gnat’sya*) ‘to chase’, «ехать» (*ekhat’*) ‘to go (by vehicle)’, «идти» (*idti*) ‘to go/walk’, «катить» (*katit’*) ‘to roll (transitive)’, «катиться» (*katit’sya*) ‘to roll (intransitive)’, «лезть» (*lez’t’*) ‘to climb/creep’, «летать» (*letat’*) ‘to fly (habitual)’, «нести» (*nesti*) ‘to carry’, «нестись» (*nestis’*) ‘to rush’, «плыть» (*plyt’*) ‘to swim/sail’, «ползти» (*polzti*) ‘to crawl’, «тащить» (*tashchit’*) ‘to drag/carry’, «тащиться» (*tashchit’sya*) ‘to trudge/go slowly’.

3.2 Models

In this study, two models were used. First, a general-purpose multilingual model, Mistral (version Mistral-7B-Instruct-v0.3[1]). Second, we used a modern instruction-tuned model, T-lite (version T-lite-instruct-0.1[2]), which is oriented on processing and generating texts in Russian. The choice was motivated by the fact that models from the T-lite family demonstrate strong performance across a range of tasks. They are based on the Qwen 2.5 architecture[3] and were further fine-tuned on large Russian-language datasets, and therefore take into account language-specific characteristics.

A comparison of these two types of models makes it possible to identify the potential impact of fine-tuning on the quality of synonym generation, as well as to evaluate differences between them in the task of identifying semantically similar words.

3.3 Prompting

In the experiments, a zero-shot approach with instruction prompting was used, in which the models were given the task of generating synonyms for verbs from the list without providing any examples, but with an explicit specification of the output format and constraints:

“Generate synonyms for the verbs from the list. The number of synonyms for each verb is 10. The synonyms must not be repeated; they must correspond to verbs and be provided in their base (infinitive) form, separated by commas, in descending order of semantic similarity. Output the result in the following format:

verb1: synonym1, synonym2, …

verb2: synonym1, synonym2, …”.

In total, 8,217 synonyms were obtained during the experiments.

3.4 Hyperparameters of Inference

In the experiments, the following inference hyperparameters were used: temperature, *top_p*, *max_new_tokens*, and *repetition_penalty*. The temperature parameter controls the “creativity” of generation and is sensitive to its value, which affects whether the outputs are more predictable or, conversely, more diverse in nature. The following values were used in this study: 0.0, 0.3, 0.5, 0.7, and 1.0.

According to the hypothesis, at values close to 0, the model tends to generate more probable lexical items that align with dictionary entries, whereas values close to 1 lead to the appearance of less expected (though still semantically appropriate) units and may introduce additional “noise”. The *top_p* parameter (nucleus sampling) controls randomness in token selection based on cumulative probability mass, which must exceed a predefined threshold (0.7 in our experiments). The *max_new_tokens* parameter defines the maximum number of new tokens the model is allowed to generate, i.e., the length of the output (set to 50 in the experiments). The *repetition_penalty* parameter is used to penalize repeated output units (set to 1.05).

Due to the stochastic nature of generation in LLMs (when temperature is greater than zero), multiple runs were performed for each verb. This is motivated by the fact that the model can produce different outputs for the same prompt; therefore, repeated generation allows for more robust and stable results. In the experiments, five runs were conducted for each configuration (model, temperature, and verb), after which the results were aggregated using frequency-based statistics.

3.5 Evaluation Metrics

To evaluate the results, descriptive statistics and derived metrics were used to assess the diversity of the generated outputs.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[2] https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1

[3] https://huggingface.co/collections/Qwen/qwen25

Diversity is defined as the ratio of the number of unique synonyms obtained by merging the results of all model runs (i.e., the cardinality of the union of sets) to the total number of generated synonyms:

$$Diversity = \frac{|U_{i=1}^{n} w_i|}{\sum_{i=1}^{n} |w_i|},$$

where n is the number of model runs; $|U_{i=1}^{n} W_i|$ is the number of unique synonyms across all runs; $|W_i|$ is the number of synonyms generated in the i-th run.

The metric demonstrates how many different words the model generates at a given temperature across multiple runs. A higher score indicates fewer repeated outputs across model runs performed under identical generation settings, reflecting a greater variety of generated synonyms.

Stability was evaluated as the mean value of the Jaccard coefficient computed over all pairs of runs. For each pair, the ratio of the cardinality of the intersection of the generated synonym sets to the cardinality of their union was calculated, after which the resulting values were averaged:

$$Stability = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{|w_i \cap w_j|}{|w_i \cup w_j|},$$

where n is the number of model runs; W_i is the set of synonyms generated in the i-th run; and W_j is the set of synonyms generated in the j-th run.

The coefficient corresponds to the overlap (intersection) of results across different generations. A higher coefficient score indicates greater output stability, meaning that the model produces more consistent (i.e., more similar or overlapping) results across repeated runs.

Diversity and stability were selected as the first step for evaluation and the primary criteria because they directly quantify variability and consistency of model outputs. In addition, the *average repeat frequency* of individual synonyms across runs was computed. For each verb, the number of runs in which a given generated lexical item appears was calculated, and these values were then averaged.

All generated items were included in the calculations regardless of their semantic correctness. Consequently, the reported metrics characterize the overall generation behaviour of the models rather than the quality of the generated synonyms. The synonym lists generated by the two models were further subjected to qualitative analysis, providing an expert linguistic evaluation of outputs.

4 Results

4.1 Mistral

The results produced by Mistral differ considerably from those generated by T-lite, both in lexical quality and in the interpretation of the synonym generation task.

Table 1 presents the values of the statistics for the Mistral model.

	0.0	**0.3**	**0.5**	**0.7**	**1.0**
average number of synonyms	9.89	8.62	9.10	8.67	9.28
diversity	0.18	0.42	0.51	0.59	0.61
stability	1.00	0.40	0.30	0.25	0.19
average repeat frequency	5.97	2.45	2.04	1.84	1.64

Table 1: Performance statistics (Mistral).

The model does not consistently generate exactly 10 lexical items despite the instruction explicitly specified in the prompt. The proportion of runs in which the required number of synonyms was produced is 19.8%. In many cases, the generated synonyms exceed a single word in length (e.g., «бегать быстро» [*begat' bystro*] 'to run fast', «бежать на большую дистанцию» [*bezhat' na bol'shuyu distantsiyu*] 'to run a long distance', «катиться по кругу» [*katit'sya po krugu*] 'to roll in a circle', «лежать неподвижно» [*lezhat' nepodvizhno*] 'to lie motionless').

In some cases, instead of a synonym, the model provides explanatory paraphrases: for instance, «ползти» [*polzti*] is glossed as 'to move using arms and legs' or 'to move on all fours', while «тащиться» [*tashchit'sya*] is rendered as 'to move along an inconvenient path' or 'to move over an uneven surface'. Occasionally, the proposed word is accompanied by clarification (e.g., «совершать» [*sovershat'*] with the sense of 'to do').

At the same time, the model does not demonstrate high lexical diversity: a large number of repetitions are observed, and the generated variants often differ only in their prepositional phrase components (e.g., «поддерживать на плаву» [*podderzhivat' na plavu*] 'to keep afloat', «поддерживать в состоянии» [*podderzhivat' v*

sostoyanii] ‘to maintain in a state’, «поддерживать в равновесии» [*podderzhivat’ v ravnovesii*] ‘to keep in balance’, «поддерживать в жизни» [*podderzhivat’ v zhizni*] ‘to sustain in life’, or «поддерживать в действии» [*podderzhivat’v deistvii*] ‘to keep in operation’).

Another characteristic feature of Mistral is the production of malformed lexical items. For example, such items as «агать» [*agat’*], «ваться» [*vat’sia*], «довать» [*dovat’*], «вать» [*vat’*] and «ься» [*’sia*] are not valid Russian lexical forms but parts of words. Such outputs most likely result from subword tokenization errors or incomplete decoding and illustrate one of the weaknesses of the model for lexical generation in Russian.

Another notable phenomenon is the generation of English examples despite the instruction requesting Russian synonyms. The output also contains invented lexical items, such as «спринтировать» (*sprintirovat’*, from English ‘to sprint’), as well as items that do not correspond to verbs (e.g., the noun «погонщик» (*pogonshchik*) ‘herder/driver’). It is noteworthy that for the verb «ползти» (*polzti*) ‘to crawl’, at temperature settings 0.00, 0.03 and 0.05, only English-language examples were produced (“crawl”, “slither”, “scuttle”, “wriggle”, “slink”, “scramble”, “scurry” and “skulk”), whereas Russian synonyms appeared only at temperature settings 0.07 and 1.00.

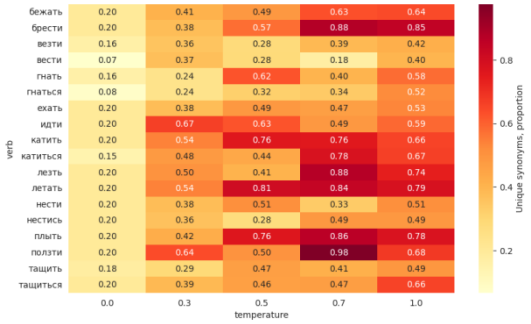

Figure 1: A heatmap of diversity by verbs (Mistral).

The heatmap (see Fig. 1) shows that, in general, the model demonstrates a tendency to generate more diverse synonyms as temperature increases. However, in some cases (e.g., for the verbs «лезть» [*lez’t’*] ‘to climb/creep’ and «летать» [*letat’*] ‘to fly’), maximum diversity is already achieved at t=0.7 and even decreases at t=1.0.

4.2 T-lite

The model generally succeeds in generating verbs belonging to the same lexical field. For example, for the verb «бежать» [*bezhat’*] ‘to run’, T-lite frequently produces the lexical items «бегать» [*begat’*] ‘to run habitually’, «мчаться» [*mchat’sya*] ‘to race’, «нестись» [*nestis’*] ‘to speed’, «убегать» [*ubegat’*] ‘to run away’ and «спешить» [*speshit’*] ‘to hurry’, that denote rapid movement. For the verb «брести» [*bresti*] ‘to trudge’, the model generates such lexical items as «плестись» [*plestis’*] ‘to plod’, «ковылять» [*kovyliat’*] ‘to hobble’, «тащиться» [*tashchit’sya*] ‘to drag oneself’ and «волочиться» [*volochit’sya*] ‘to trail along’, which share the sememes of slow or difficult movement. The results suggest that the model captures fine-grained semantic distinctions related to the manner of motion rather than relying solely on general semantic similarity.

In some cases, during test runs, the model generated additional verbs that were not specified in the original prompt, which required post-processing of the results.

Table 2 presents the values of the evaluation metrics for the T-lite model.

	0.0	**0.3**	**0.5**	**0.7**	**1.0**
average number of synonyms	10.00	12.96	11.00	10.93	11.75
diversity	0.17	0.52	0.61	0.74	0.88
stability	1.00	0.27	0.24	0.13	0.06
average repeat frequency	5.88	1.94	1.72	1.35	1.14

Table 2: Performance statistics (T-lite).

In contrast to Mistral, the model generates a more diverse set of synonyms, which in most cases do not exceed one word in length. However, in a number of cases, the model appears to generate descriptive expressions rather than lexical synonyms, effectively explicating the meaning of the source verb. Examples include «идти пешком» [*idti peshkom*] ‘to walk on foot’, «нести в руках» [*nesti v rukakh*] ‘to carry in one’s hands’, «вести за собой» [*vesti za soboy*] ‘to lead along’ and «плыть по течению» [*plyt’ po techeniiu*] ‘to drift with the current’. Although these phrases are semantically meaningful, they do not satisfy

the definition of single-word lexical substitution adopted in this study.

Another recurring phenomenon is the generation of text continuations rather than synonyms. For instance, the results include «вести к» [*vesti k*] ‘to lead to’, «вести через» [*vesti cherez*] ‘to lead through’, «взбираться на» [*vzbirat’sya na*] ‘to climb onto’ and «подняться в небо» [*podniat’sya v nebo*] ‘to rise into the sky’, suggesting that the model interprets the task as one of text continuation rather than lexical substitution.

T-lite occasionally produces semantically inappropriate items. For example, for the verb «везти» (*vezti*) ‘to transport; to carry by vehicle’, T-lite occasionally generates «говорить» (*govorit’*) ‘to speak’, which is unrelated to the semantics of motion. Another example is the colloquial verb «тащиться» (*tashchit’sya*) ‘to drag oneself; to trudge’ which produces «врубаться» (*vrubat’sia*) ‘to get into; to understand (slang)’, reflecting the secondary colloquial meaning of «тащиться» (‘to enjoy’ or ‘to be enthusiastic about’) rather than its literal meaning dealing with motion. This observation demonstrates that the model is sensitive to highly frequent colloquial senses and fails to correctly disambiguate polysemous verbs.

The model adheres more strictly to the prompt taking into account the number of items: the majority of runs (72.3%) produced exactly 10 synonyms; however, increasing temperature leads to an increase in the volume of generated outputs.

Fig. 2 presents a heatmap showing the results for the T-lite model according to the diversity metric. The model consistently produces more diverse outputs as temperature increases. It is important to note that, overall, the generated synonyms are semantically close to the corresponding verbs. For example, for «брести» (*bresti*) ‘to trudge’, the model produces: «ковылять» (*kovylyat’*) ‘to hobble’, волочить ноги (*volochit’ nogi*) ‘to drag one’s feet’, «ползти» (*polzti*) ‘to crawl’, «шаркать» (*sharkat’*) ‘to shuffle’, «плестись» (*plestis’*) ‘to plod’.

The model did not generate outputs at t=0.0 for the verbs «везти» (*vezti*) ‘to transport (by vehicle)’ and «нестись» (*nestis’*) ‘to rush’. One possible explanation is the use of greedy decoding, whereby the model deterministically selects the most probable token sequence. In some cases, this decoding strategy may fail to generate a complete response, resulting in empty outputs.

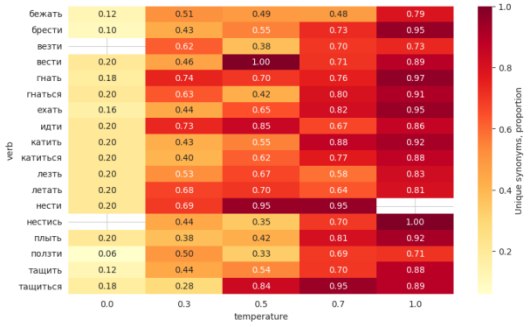

Figure 2: A heatmap of diversity by verbs (T-lite).

In addition, no substitutions were produced for the verb «нестись» (*nestis’*) at t=1.0, which may be explained by the stochastic nature of the model. At higher temperature values, this stochasticity can manifest as “noise” and generation errors.

4.3 Discussion

The analysis of the results reveals notable differences in the candidate synonyms generated by the two models. Overall, T-lite predominantly generates lexical synonyms, whereas Mistral frequently fails to adhere to the lexical substitution task, instead producing meaning descriptions, text continuations, parts of the original prompt or explanatory reasoning accompanying its responses. These findings suggest that the differences between the models are not limited to the number of correct responses but also involve distinct strategies for interpreting the instruction. While T-lite generally treats the prompt as a lexical substitution task, Mistral often interprets it as a more general text generation task.

For both models, the proportion of unique synonyms increases with rising temperature, i.e., the diversity of generated outputs becomes higher. However, this is also accompanied by an increase in “noise” and erroneous outputs, particularly in the case of Mistral. Conversely, the stability metric demonstrates the robustness of both models at t=0.0, where the results are nearly identical across runs, whereas at t=1.0 maximum variability is observed, reflecting increased stochasticity in generation.

The T-lite model exhibits a lower degree of overlap between runs as temperature increases

compared to Mistral; at the highest temperature, the intersection between runs is almost negligible. This also indicates greater generative variability and creativity of the model.

The average repeat frequency metric shows that at low temperature values, synonyms that are consistently reproduced across all runs dominate, whereas with increasing temperature, the proportion of unique lexical items that appear only once increases.

5 Conclusion

The analysis of the generated lexical items shows that instruction-tuned LLMs capture semantic relations between Russian motion verbs, but also exhibit some limitations associated with the complex nature of lexical semantics of Russian verbs, including aspectual characteristics, polysemy or stylistic variation.

The results confirm the feasibility of using LLMs for generating synonyms for a given lexical unit; however, the selected parameter settings significantly influence the obtained outputs. Varying the temperature parameter allowed control over the stochasticity and diversity of the generated items. At low temperature values, the systems exhibit a high rate of repetition among outputs, i.e., generation is largely deterministic, whereas higher temperature values lead to more variable outputs, which are nevertheless characterized by instability across runs. At intermediate values, a balance between stability and diversity of generation is achieved.

The T-lite model demonstrated a lower number of errors and irrelevant outputs compared to the multilingual Mistral model. The latter exhibits weaker adherence to the constraints (including lexical ones) imposed by the prompt. Our observations suggest that instruction-tuned LLMs perform lexical generation rather than dictionary-based synonym selection, showing both the flexibility and some limitations of zero-shot approach.

In future work, it is planned to include a larger number of models fine-tuned on Russian-language data for comparative evaluation. Another direction of research may involve assessing the extent to which modern Russian instruction-tuned LLMs are capable of reproducing synonymic relations documented in lexicographic resources.

Acknowledgments

The author acknowledges the financial support of Russian Science Foundation (project No. 22-18-00189-P “Structure and Functionality of Stable Multiword Units in Russian Everyday Speech”).

References

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, pages 48–53.

Diana McCarthy and Roberto Navigli. 2009. The English Lexical Substitution Task. In *Language Resources and Evaluation*, 43(2), Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond, Agirre E., Màrquez L., Wicentowski R. (eds.), pages 139–159. Springer.

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.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based Lexical Substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. A Comparative Study of Lexical Substitution Approaches Based on Neural Language Models. *arXiv preprint arXiv:2006.00031 [cs]*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Ning Shi, Bradley Hauer, and Grzegorz Kondrak. 2024. Lexical Substitution as Causal Language Modeling. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 120–132, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jiuding Sun, Chantal Shaib, and Byron C. Wallace. 2023. Evaluating the Zero-shot Robustness of Instruction-tuned Language Models. *arXiv preprint arXiv:2306.11270*

Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2024. A survey on semantic processing techniques. *Information Fusion*, Vol. 101, Art. 101988.

Huimin Ren, Yan Liang, Baiqiao Su, Chaobo Sun, Hengtong Lu, Kaike Zhang, and Chen Wei. 2026. LexInstructEval: Lexical Instruction Following Evaluation for Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(30), pages 25117–25123.

Natalia Yu. Shvedova (ed.). 1980. *Russkaya grammatika (AG-80) [Russian Grammar (AG-80)]*, Vol. 1: Phonetics. Phonology. Stress. Intonation. Word Formation. Morphology. Moscow: Nauka, 789 p.