DOI: 10.14529/jsfi250305

Can LLMs Get to the Roots? Evaluating Russian Morpheme Segmentation Capabilities in Large Language Models

Dmitry A. Morozov[1], *Anna V. Glazkova*[2], *Boris L. Iomdin*[3]

Automatic morpheme segmentation, a crucial task for morphologically rich languages like Russian, is persistently hindered by a significant drop in performance on words containing out-of-vocabulary (OOV) roots. This issue affects even state-of-the-art models, such as fine-tuned BERT models. This study investigates the potential of modern Large Language Models (LLMs) to address this challenge, focusing on the specific task of root identification in Russian. We evaluate a diverse set of eight state-of-the-art LLMs, including proprietary and open-weight models, using a prompt-based, few-shot learning approach. The models' performance is benchmarked against strong baselines – a fine-tuned RuRoberta model and a CNN ensemble – on a 500-word test set. Our results demonstrate that one model, Gemini 2.5 Pro, surpasses both baselines by approximately 5 percentage points in root identification accuracy. An examination of the model's reasoning capabilities shows that while it can produce logically sound, etymologically-informed analyses, it is also highly prone to factual hallucinations. This work highlights that while LLMs show significant promise in overcoming the OOV root problem, the inconsistency of their reasoning presents a significant obstacle to their direct application, underscoring the need for further research into improving their factuality and consistency.

Keywords: morpheme segmentation, tokenizers, large language models, Russian language.

Introduction

Word segmentation into minimal meaningful substrings, morphemes, is important for morphologically rich languages. The construction of morpheme segmentations can be applied in language learning, for building hypotheses about possible word etymology, or as a subword tokenizer. In the latter capacity, morpheme-oriented tokenizers can be used for low-resource languages as an alternative to common BPE tokenizers [2, 9, 11, 14].

The need for automation in this task arises from the incompleteness of existing morpheme dictionaries. For instance, in Russian, the largest dictionaries contain no more than half of the words found in the Russian National Corpus [5]. At the same time, algorithmic approaches today have an accuracy that, on average, is not inferior to expert annotation [13]. However, existing algorithms have a significant drawback: a low quality of performance with roots not encountered in the training sample [7, 13]. This problem can be partially solved by using pre-trained BERT-like models, but the quality of annotation for words containing out-of-vocabulary (OOV) roots is still significantly lower than the average, and the vast majority of annotation errors are specifically related to identifying root boundaries in such words [12].

A potential solution to the problem of identifying OOV roots could be the use of large language models (LLMs) for word segmentation and root finding [1, 17]. However, the applicability of LLMs to morpheme segmentation remains insufficiently investigated.

In this work, we focus on Russian, the language that is both well-represented in the training corpora of state-of-the-art LLMs and well-studied within the context of morpheme segmentation. Since the primary challenge for existing methods is the identification of OOV roots, we decided to concentrate on the applicability of LLMs specifically to the task of word root identification.

[1] Novosibirsk State University, Novosibirsk, Russian Federation
[2] University of Tyumen, Tyumen, Russian Federation
[3] Käthe-Kollwitz-Gymnasium, Berlin, Germany

The rationale is that if the root can be successfully identified, the remainder of the word can be segmented with very high accuracy using established algorithms.

Our main contributions are as follows:

- LLMs can outperform other approaches in identifying word roots, yet the quality of segmentation using them is still far from ideal;
- examining the content of the reasoning fields from a linguistic perspective showed that in a number of cases, the segmentation variant proposed by the model is logically justified and more suitable than the dictionary-based one.

The rest of the paper is structured in the following way. Section 1 contains a brief review of related work on automatic morpheme segmentation. Section 2 describes the dataset and the models utilized for the experiments. Section 3 presents the experimental results and discussion. Section 3 concludes this paper, summarizing the study and pointing directions for further work.

1. Related Work

Automatic morpheme segmentation comprises two main paradigms: surface segmentation, where a word is segmented into its constituent substrings (e.g., funniest → funn-i-est), and canonical segmentation, which aims to restore the underlying forms of the morphemes (e.g., funniest → fun-y-est) [6]. For both tasks, machine learning-based algorithms have demonstrated the highest efficacy.

The task of surface segmentation has been extensively studied for the Russian language. The problem is typically framed as a character-level classification task, where each character is assigned a two-part label. The first part of the label indicates the character's position within a morpheme, while the second specifies the morpheme type. Among traditional approaches not leveraging pre-trained models, strong results have been achieved using Convolutional Neural Networks (CNNs) [20] and Long Short-Term Memory (LSTM) networks [4], with the performance of CNNs on random word samples being comparable to expert-level annotation. The use of convolutional networks has also shown strong results for other languages [15, 19].

Fine-tuning BERT-like models on Russian data has further improved performance, achieving a character-level accuracy exceeding 97% and a perfect-segmentation rate for words over 92% on random samples [12]. This approach has also yielded strong results for other Slavic languages, including Belarusian and Czech. Furthermore, using pre-trained models partially addresses the key limitation of CNN and LSTM architectures: the sharp decline in performance on words containing roots not seen in the training data. Nevertheless, the challenge of OOV roots remains critical, with a performance drop of over 15% in word-level accuracy for such cases [13].

For canonical segmentation, state-of-the-art algorithms were benchmarked in the SIGMORPHON 2022 shared task [3]. The top-performing systems were developed by the DeepSPIN team using LSTM and Transformer-based models [16], closely followed by the CLUZH team with models based on neural transducers [22]. During testing on nine languages (English, Spanish, Hungarian, French, Italian, Russian, Czech, Latin, and Mongolian), participants achieved morpheme-level F1-scores exceeding 93.5 for all languages, with scores surpassing 99 for three languages, including Russian. However, subsequent testing of the DeepSPIN approach confirmed that, as with surface segmentation, OOV roots remain the primary challenge [12].

Notable existing methods utilizing LLMs include the LLMSegm algorithm [17]. In this framework, a pre-trained Glot500 model [8] performs binary classification for each potential boundary position within a word. Despite its significant computational complexity (requiring $N-1$ LLM inferences for a word of length N), this method has outperformed previous techniques for several low-resource South African languages. Conversely, end-to-end generation of segmentations for another low-resource language, Bribri, using an LLM (Claude Sonnet 3.7), while more efficient on average, proved to be less effective for multi-morphemic words than a non-pretrained algorithm [1]. Therefore, the application of LLM-based approaches to morpheme segmentation is currently under-explored.

2. Data and Models

2.1. Data

The task of morpheme segmentation for the Russian language is complicated by the absence of a unified approach among linguists for defining what constitutes a morphemic analysis. In practice, the choice of a dataset for experiments determines the segmentation paradigm, and an algorithm trained on one dataset will inherently produce segmentations that are incorrect from the perspective of another. At the same time, previous studies have shown that the performance of algorithms is generally similar across different datasets [7, 13]. The largest machine-readable datasets for Russian morpheme segmentation currently available are Morphodict-K (based on the "Dictionary of Morphemes of the Russian Language" (ed. A. I. Kuznetsova and T. F. Efremova) [10]), Morphodict-T (based on the "Word Formation Dictionary of the Russian Language" (ed. A. N. Tikhonov) [21]), and the Russian dataset from the SIGMORPHON 2022 Shared Task on Morpheme Segmentation [3]. In the present study, we chose to use the Morphodict-K dataset. This decision was based on two reasons:

- The Morphodict-K and the Morphodict-T datasets use the same notation: words are segmented into morphemes with an indication of the type of each of them, in total, 7 types of morphemes are used: PREF (prefixes), ROOT, SUFF (suffixes), END (endings), LINK (connecting vowels), POST (postfixes), HYPH (hyphens). This makes it easy to move from experiments with one dataset to experiments with another. However, the segmentation in Morphodict-K is based on etymology and features a high degree of morpheme granularity. Although this paradigm is not strictly formalized, it allows for significantly less subjectivity than the paradigm underlying the Morphodict-T dataset. The latter employs a non-obvious criterion of the transparency of derivational chains in modern Russian. This leads to the identification of different roots in words, which relatedness is obvious to native speakers (e.g., in *dobro* 'goodness', the root is identified as -*dobr*-, whereas in *odobrenie* 'approval', the root is identified as -*odobr*-). The use of such a criterion reduces the internal consistency of the annotation, which is also reflected in the lower performance of automated methods on this dataset.
- Upon closer inspection, the SIGMORPHON dataset is annotated in a highly inconsistent manner. Specifically, there is no uniform approach to segmenting the infinitive suffix -*t'*- or the adjectival ending -*yy*-; in about half of the cases, they are merged with the preceding morpheme. This internal inconsistency, along with the absence of documented segmentation rules, makes this dataset a poor candidate for our research.

Therefore, within the scope of our study, we aimed to identify the etymological root specifically. This approach may be more valuable for creating morpheme-oriented tokenizers for training language models, as it offers more fine-grained segmentation and results in a smaller token vocabulary.

The utilized dataset was split by word roots into training and test sets at an approximate 4:1 ratio. Words containing multiple roots were preliminarily removed from the dataset. However, due to financial constraints, the LLM testing was not performed on the entire test set, but on a random sample of 500 words from it. For baseline models, we calculated the quality both on the entire test sample and on the selected 500 words. A summary of the characteristics of the dataset and the samples is provided in Tab. 1.

Table 1. Brief characteristics of the datasets

Dataset	**Morphodict-K**	**Train set**	**Test set**	**500 test words**
Unique words	75 649	51 620	12 401	500
Unique morphemes	8 079	5 606	1 662	434
Unique roots	7 148	4 768	1 192	275
Avg morphemes per word	4.12	3.95	3.98	4.07
Avg morpheme occurrence	38.56	36.36	29.71	4.69
Avg root occurrence	12.24	10.83	10.40	1.82
Avg characters in root	4.62	3.64	3.74	3.73

2.2. Prompt-based models

To ensure the representativeness of the study, we used a variety of state-of-the-art multilingual general-purpose LLMs developed and trained by different teams. The access to the models was provided via the OpenRouter API.

Some of the models used are proprietary, accessible only through API, and architecture and training details are unknown. Using such models inevitably leads to a worse understanding of the reasons for the effectiveness of a particular model, but the high performance of these models in independent benchmarks necessitates their inclusion for a comprehensive evaluation. These models include:

1. Claude Sonnet 4[4];
2. Gemini 2.5 Pro and Gemini 2.5 Flash Lite[5];
3. Mistral Medium 3.1[6];
4. GPT 5 Chat[7] (we used this model instead of the base GPT 5 because GPT 5 was still not available in the OpenRouter API at the time of our experiments.).

We compared these models to the models whose weights are available on HuggingFace:

1. Llama 4 Maverick[8], an auto-regressive language model that uses a mixture-of-experts (MoE) architecture with 128 experts and has 17B activated parameters (400B parameters in total);
2. gpt-oss-120b[9] with 117B parameters and 5.1B active parameters;
3. Qwen3-235B-A22B[10], MoE model with 128 experts (8 activated experts) and 235B parameters in total (22B parameters activated).

[4] https://www.anthropic.com/claude/sonnet

[5] https://deepmind.google/models/gemini/

[6] https://mistral.ai/news/mistral-medium-3

[7] https://openai.com/index/introducing-gpt-5/

[8] https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct

[9] https://huggingface.co/openai/gpt-oss-120b

[10] https://huggingface.co/Qwen/Qwen3-235B-A22B

It should also be noted that the models used differ significantly in the costs of API requests. This is due to three reasons: first, different tokenization of requests and responses, second, different token prices, and third, the presence of reasoning, which significantly affected the volume of generated text. For the Claude and Gemini models, we forced the reasoning mode to be enabled. The other models used do not have such a setting via the OpenRouter API, but we recorded non-empty reasoning in the response in the case of gpt-oss-120b (for all 500 words) and Qwen3-235B-A22B (for 35 words). The cost of a request varied from 10^{-4} to 10^{-2} USD. In total, less than 25 USD were spent on experiments (including preliminary ones). A temperature value of 0 was used during generation for all models.

2.3. Prompt

To construct the prompt, we used the segmentation paradigm description from the original dataset[11] and a series of examples from the training set.

During preliminary experiments with models Gemini 2.5 Flash Lite, Mistral Medium 3.1, and gpt-oss-120b, we tested different strategies for selecting examples presented to the model in the prompt. In particular, we attempted to use randomly selected examples from the training set, as well as words similar to the analyzed word by a substring at the beginning or end of the word. We also tried different numbers of examples in the prompt (from 10 to 100). These strategies turned out to be insufficiently effective. Increasing the number of examples with a random selection strategy improved the quality only slightly, and when adding words with a similar substring, the models tended to overfit.

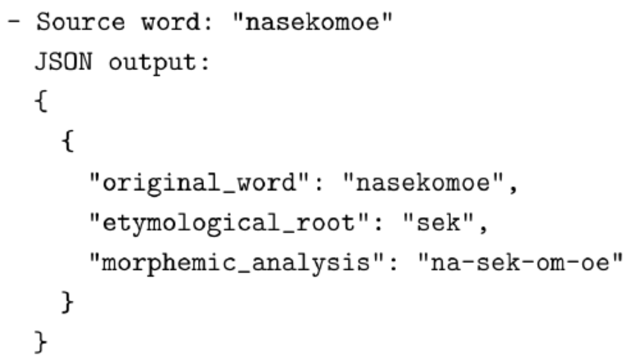
The best result was achieved by adding a small number of examples to the prompt, illustrating various features of the segmentation paradigm: consideration of etymology and a high degree of morpheme granularity (in comparison, for example, with the approach of the "Word Formation Dictionary of the Russian Language" (ed. A. N. Tikhonov) [21]. Examples were selected iteratively; the final prompt included 13 words from the training sample: *"nasekomoe"* ('insect', root -*sek*-), *"ulybat'sya"* ('to smile', root -*lyb*-), *"revolyutsiya"* ('revolution', root -*volyuts*-), *"vostochnyy"* ('eastern', root -*toch*-), *"obidet'sya"* ('to be offended', root -*obid*-), *"pozlashchat'"* ('to gild', root -*zlashch*-), *"obratit'"* ('to turn', root -*obrat*-), *"bytovoy"* ('domestic, household', root -*by*-), *"nenastnyy"* ('inclement', root -*nast*-), *"izverzhenie"* ('eruption', root -*verzh*-), *"uproshchat'sya"* ('to be simplified', root -*proshch*-), *"truzhenichestvo"* ('hard work', -*truzh*-), *"annulirovanie"* ('cancellation', root -*nul*-). An example of word formatting is shown in Fig. 1.

Additionally, the prompt included requirements describing the surface segmentation procedure: the concatenation of morphemes must exactly match the original word, and in cases of alternations, the root must be extracted in the exact form in which it appears in the word. Finally, during preliminary experiments, we determined that the segmentation quality improves when the response includes a complete segmentation. The final prompt was written in Russian. In Fig. 2 we provide a translation of the original prompt into English

2.4. Baselines

As baselines we considered two approaches, originally developed for constructing surface segmentation with prediction of the type of each morpheme. The task of constructing morpheme

[11] `https://ruscorpora.ru/en/page/instruction-derivation`



```
- Source word: "nasekomoe"
  JSON output:
  {
    {
      "original_word": "nasekomoe",
      "etymological_root": "sek",
      "morphemic_analysis": "na-sek-om-oe"
    }
  }
```

Figure 1. An example of word formatting

segmentation is considered as a task of character-level classification. The choice of these models was based on their high quality in previous studies [12, 13, 20]. In our study, we also trained models on the surface segmentation task, and then extracted the morpheme marked by the model as the root of the word. The baselines are:

1. **Convolutional neural network ensemble** [20]. The ensemble consists of three identical convolutional networks trained independently. We used the original implementation of the algorithm. The number of convolutional layers in each convolutional network was 3, the number of filters was 192. Each of the models was trained for 25 epochs on an AMD Ryzen 5 5600X CPU.
2. **Fine-tuned RuRoberta model** [12]. We fine-tuned `ruRoberta-large`[12] (355M parameters) [23] for token-classification task. The input sequence consisted of the lemma itself and sequence of the separated word letters. The output sequence is ‘0’ for the lemma and letter class for each letter. For implementation we used the `simpletransformers`[13] framework [18]. The batch size during training was set to 16, and the learning rate was set to 4e-6. The values of the remaining parameters were set to default. We fine-tuned the model for 30 epochs on an Nvidia RTX 4090 GPU.

3. Results and Discussion

The results we obtained are presented in Tab. 2 below, where the LLMs are ordered by the decreasing number of correctly identified roots. The performance of the baseline models is also included for comparison.

Notably, Gemini 2.5 Pro was the only model to outperform the baselines, achieving a root accuracy nearly 5% higher. The remaining LLMs underperformed compared to the baseline models. Furthermore, we observed that proprietary models generally surpassed open-weight models.

Our analysis of LLM errors revealed no single, consistent pattern of incorrect root identification; the models tended to err on different words rather than the same ones (Fig. 3). For instance, only 185 words were correctly segmented by all eight models, while 474 words were segmented correctly at least once (meaning 26 words were never segmented correctly by any model).

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

[13] `https://simpletransformers.ai/`

```
You are a linguistic expert specializing in morphemic and etymological
   analysis of the Russian language. Your task is to conduct morphemic
   analysis of words, strictly following the principles of the "Dictionary
   of Morphemes of the Russian Language" by A. I. Kuznetsova and T. F.
   Efremova. Your answer should always be in JSON format.

Your task is to perform a morphemic analysis of a word and identify its
   etymological root. Follow these rules:

0. **Exact match:** The morpheme segmentation of the word MUST be a letter-
   for-letter match with the original word and must not change any letters.
1. **Principle of granularity:** Divide the word into morphemes (prefixes,
   root, suffixes, endings) in as much detail as possible.
2. **Historical root:** Identify historical and etymological roots, even if
    their connection to the modern meaning is not obvious to a native
   speaker.
3. **Structural correlation:** Identify morphemes (including the root) if
   there are other words in the language with a similar structure or
   morphemes, even if the word is not used without them (e.g., "u-lyb-at'
   sya" by analogy with "u-smekh-at'sya").
4. **Analysis of loanwords:** Segment borrowed stems if there are other
   lexemes in the Russian language with similar structural elements (e.g.,
   "re-volyuts-iya" and "e-volyuts-i-ya").
5. **Handling of the soft sign:** If a soft sign follows the root, include
   it in the root (e.g., "kol'-ts-o").
6. **Alternations:** If an alternation is allowed in the root, you must
   choose the spelling that is present in the word (e.g., in the word "
   pozlashchat'" the root is "zlashch").

Here are some reference examples:

[EXAMPLES]

Now, please process the following word:

Source word: "{word_to_analyze}"

Provide the answer strictly in the following JSON format, without including
    any other explanations. The etymological root MUST be part of the full
   morpheme segmentation.
{{
  "original_word": "<the word being processed>",
  "etymological_root": "<the etymological root>",
  "morphemic_analysis": "<the full morpheme segmentation with hyphens>"
}}
```

Figure 2. Translation of the used prompt into English

Table 2. Experiment results

Model	Correct roots	Root-level accuracy	Fully correct segmentations	Word-level accuracy
Gemini 2.5 Pro	430	0.86	392	0.78
Mistral Medium 3.1	391	0.78	349	0.69
Claude Sonnet 4	386	0.77	294	0.59
Gemini 2.5 Flash Lite	355	0.71	279	0.56
GPT 5 Chat	343	0.69	251	0.50
Llama 4 Maverick	341	0.68	305	0.61
Qwen3 235B A22B	335	0.67	242	0.48
gpt-oss-120b	334	0.67	228	0.46
fine-tuned ruRoberta-large	401	0.80	387	0.77
CNN ensemble	406	0.81	358	0.72

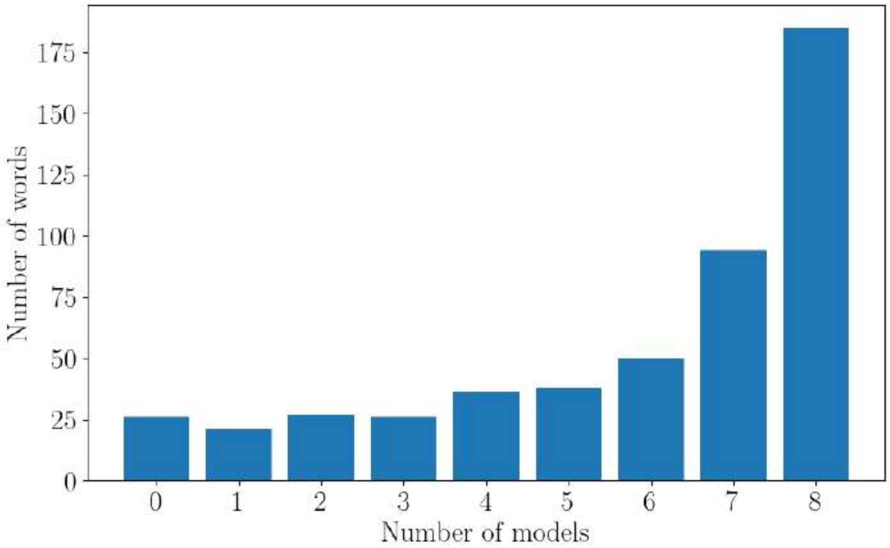


Figure 3. Number of words for which the root was correctly identified by N models

The primary challenge for the LLMs was the root *-sta-* in words such as *zastava* 'outpost', *nastavlyat'* 'to instruct', and *predstavitel'skiy* 'representative'. In these cases, all LLMs incorrectly identified the root as *-stav-*. However, the correct root *-sta-* was successfully identified by at least some models in words like *perestavat'* 'to cease' and *ostanovit'sya* 'to stop'. The current experimental design does not allow us to draw general conclusions about the relationship between specific root features and the quality of their identification. We plan to replicate this experiment with a larger dataset in the future to investigate this issue further.

Interestingly, even the top-performing model, Gemini 2.5 Pro, incorrectly processed some words for which the other seven models provided the correct answer, such as *"vaflya"* ('waffle') and *"prorab"* ('foreman'). However, an analysis of the responses showed that once the model had violated the response format: for *"vaflya"*, it identified the root *-vafl'-*, where the soft sign might be included from a phonetic standpoint but is incorrect within a surface segmentation paradigm.

Since Gemini 2.5 Pro was the only model to surpass the baseline approach, we decided to focus our error analysis on this model as the most promising. A preliminary analysis revealed that in several cases, a discrepancy between the predicted root and the reference root might not indicate a model error but rather an inaccuracy in the dictionary or the possibility of a different interpretation. Consequently, we conducted a detailed manual analysis of 70 words for which the root identified by the model differed from the reference. We evaluated the model's responses based on two criteria:

1. **Is the model's answer more suitable than the reference?** A score of 2 indicated the model's answer was better, 1 meant it was difficult to choose the better option, and 0 meant the reference was better.
2. **Is the model's reasoning factually and logically sound?** A score of 2 meant the reasoning represented a well-conducted word-formation analysis containing only correct etymological facts; 1 indicated that the reasoning contained a mix of correct statements and hallucinations; and 0 meant the reasoning consisted entirely of hallucinations.

The analysis showed that the root predicted by the model was more suitable than the reference in 13 out of 70 cases, and in another 11 cases, choosing the better option was difficult. Cases where the model's prediction was more accurate included words such as:

- *“anonimnost’”* ‘anonymity’, reference root -*nim*-, predicted -*onim*-: in this case, -onim- is better suited as a root, since words such as *“anonim”* ‘anonymous’ and *“sinonim”* ‘synonymous’ contain the borrowed ancient Greek root $-o\nu o\mu\alpha-$, that is, -*o*- cannot be considered either a prefix or a linking vowel;
- *“oblechennyy”* ‘endowed’, reference root -*oblech*-, predicted -*lech*-: from the point of view of the segmentation paradigm used in the Morphodict-K dataset, it is reasonable to isolate the historical prefix -*ob*-;
- *“podkrylok”*‘fender liner’, reference root -*kry*-, predicted -*kryl*-: in the Proto-Slavic language, this root was supposedly contained in a form containing -*l*- (**kridlo*, **skrdlo*), so its isolation here is incorrect.

The model provided adequate reasoning in 12 cases and partially correct reasoning in 29. However, it is important to note that correct reasoning coincided with a better-identified root in only 6 of these cases (Tab. 3).

Table 3. Analysis of predicted roots different from reference ones for the Gemini 2.5 Pro model

Root correctness \ Reasoning validity	0	1	2
0	27	16	3
1	1	7	3
2	1	6	6

On three occasions, correct reasoning led to a root that was less suitable than the reference:

- *“ukomplektovyvat’sya”* ‘to be staffed/equipped’ and *“doukomplektovat’sya”* ‘to become fully staffed/equipped’, reference root -*komplekt*-, predicted -*plekt*-: despite the fact that etymologically this substring contains the prefix -*kom*- (in Latin *“completus”* ‘complete’), this root was borrowed into the Russian language in its current form through Polish and, before that, German, and there are no other Russian words with substring *-plekt-*, so the rule about lexemes with similar structural elements does not apply here;

- *"sovetizirovat'sya"* 'to become sovietized', reference root *-vet-*, predicted *-sovet-*: from the point of view of the segmentation paradigm used in the Morphodict-K dataset, it is reasonable to isolate the historical prefix *-so-*.

In some instances, the model's reasoning arrived at the correct root, yet a different option was chosen for the final answer (e.g., for *"rabota"* 'work' the reasoning pointed to the root *-rab-*, but the answer given was *-rabot-*).

The majority of Gemini 2.5 Pro's errors relate to the incorrect handling of root alternations (e.g., *-treb-/-trebl-*, *-yav-/-yavl-*), insufficient consideration of etymology (for instance, the model failed to connect the words *"stat'"* ('to become/stand') and *"stavit'"* ('to put/place')), or the excessive segmentation of loanwords in cases where there is no basis for segmentation in the Russian language (*interes* → *inter-es*, *krendel* → *krend-el*, or the aforementioned *komplekt* → *kom-plekt*). Furthermore, the reasoning often contains fabricated facts or flawed logical transitions (even when the identified root is correct), which should be considered a significant drawback for the potential integration of the model into lexicography for automating dictionary creation.

In addition to analyzing the model's errors, we conducted a zero-shot evaluation using Gemini 2.5 Pro to evaluate the example selection strategy. For this, we removed the examples from the prompt, leaving the rest of the instructional text unchanged. We then generated roots for the test set. The model produced 420 correct roots and 387 fully correct analyses. This resulted in a correct root rate of 84%, compared to 86% achieved with the few-shot approach. However, we observed an increased proportion of incorrectly formatted responses among the model's errors in the zero-shot setting. The model often appended hyphens or extraneous explanations to the root, and in two instances, it generated an empty response. Therefore, despite the small difference in metrics, the few-shot approach enhances the stability of the response format, which can be critical for practical applications.

Conclusion

A key challenge for state-of-the-art automatic morpheme segmentation algorithms is their poor performance on words containing roots that were not present in the training data. This paper presents an investigation into the potential of using Large Language Models (LLMs) to overcome this limitation. For our experiments, we utilized the Russian-language Morphodict-K dataset and a range of multilingual, general-purpose LLMs, including the most current proprietary models. The Russian language was selected because it is, on the one hand, well-represented in the training corpora of these LLMs and, on the other, a well-studied language for the morpheme segmentation task.

We compared the efficacy of LLMs against two strong baselines – a fine-tuned BERT-like model and an ensemble of convolutional neural networkson the specific task of word root identification. Using the Gemini 2.5 Pro model, we successfully surpassed the baselines by 5 percentage points in accuracy. A subsequent linguistic analysis of this model's errors revealed that in several instances, the root predicted by the LLM was more suitable than the reference one from the dataset. An examination of the model's reasoning fields showed that it is sometimes capable of justifying its choices with factual evidence. However, it frequently generates reasoning that contains hallucinations and fabricated facts. This should be considered a significant drawback for the potential integration of the model into lexicography for automating dictionary creation.

The limitations of this study include the relatively small size of the test set (500 words), the focus on a single target language, and the limited number of prompting strategies explored. In addition, a significant limitation of the overall approach lies in its speed and cost: although we do not query the LLM for every possible boundary position in a word, each word is processed using a separate query to the model via the API. Despite these constraints, our approach managed to outperform state-of-the-art baselines, which underscores the need for more extensive and larger-scale research in this domain.

References

1. Anderson, C., Nguyen, M., Coto-Solano, R.: Unsupervised, semi-supervised and LLM-based morphological segmentation for Bribri. In: Mager, M., Ebrahimi, A., Pugh, R., *et al.* (eds.) Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). pp. 63–76. Association for Computational Linguistics, Albuquerque, New Mexico (May 2025). `https://doi.org/10.18653/v1/2025.americasnlp-1.7`

2. Asgari, E., Kheir, Y.E., Javaheri, M.A.S.: MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies (2025), `https://arxiv.org/abs/2502.00894`

3. Batsuren, K., Bella, G., Arora, A., *et al.*: The SIGMORPHON 2022 shared task on morpheme segmentation. In: Nicolai, G., Chodroff, E. (eds.) Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 103–116. Association for Computational Linguistics, Seattle, Washington (Jul 2022). `https://doi.org/10.18653/v1/2022.sigmorphon-1.11`

4. Bolshakova, E., Sapin, A.: Bi-LSTM model for morpheme segmentation of Russian words. In: Ustalov, D., Filchenkov, A., Pivovarova, L. (eds.) Artificial Intelligence and Natural Language. pp. 151–160. Springer International Publishing, Cham (2019). `https://doi.org/10.1007/978-3-030-34518-1_11`

5. Bonch-Osmolovskaya, A., Gladilin, S., Kozerenko, A., *et al.*: Russian National Corpus 2.0: corpus platform, analysis tools, neural network models of data markup. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (01 2025). `https://doi.org/10.28995/2075-7182-2025-23-57-73`

6. Cotterell, R., Vieira, T., Schütze, H.: A joint model of orthography and morphological segmentation. In: Knight, K., Nenkova, A., Rambow, O. (eds.) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 664–669. Association for Computational Linguistics, San Diego, California (Jun 2016). `https://doi.org/10.18653/v1/N16-1080`

7. Garipov, T., Morozov, D., Glazkova, A.: Generalization ability of CNN-based Morpheme Segmentation. In: 2023 Ivannikov Ispras Open Conference (ISPRAS). pp. 58–62 (2024). https://doi.org/10.1109/ISPRAS60948.2023.10508171

8. Imani, A., Lin, P., Kargaran, A.H., *et al.*: Glot500: Scaling multilingual corpora and language models to 500 languages. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1082–1117. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.61

9. Kildeberg, M.W., Schledermann, E.A., Larsen, N., van der Goot, R.: From Smør-re-brød to Subwords: Training LLMs on Danish, One Morpheme at a Time (2025), https://arxiv.org/abs/2504.01540

10. Kuznetsova, A.I., Efremova, T.F.: Dictionary of Morphemes of the Russian Language. Russkii yazyk, Moscow (1986)

11. Matthews, A., Neubig, G., Dyer, C.: Using morphological knowledge in open-vocabulary neural language models. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1435–1445. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1130

12. Morozov, D., Astapenka, L., Glazkova, A., Garipov, T., Lyashevskaya, O.: BERT-like models for Slavic morpheme segmentation. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6795–6815. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.337

13. Morozov, D., Garipov, T., Lyashevskaya, O., *et al.*: Automatic morpheme segmentation for Russian: Can an algorithm replace experts? Journal of Language and Education 10(4), 71–84 (Dec 2024). https://doi.org/10.17323/jle.2024.22237

14. Nzeyimana, A., Niyongabo Rubungo, A.: KinyaBERT: a morphology-aware Kinyarwanda language model. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5347–5363. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.367

15. Olbrich, M., Žabokrtský, Z.: Morphological segmentation with neural networks: Performance effects of architecture, data size, and cross-lingual transfer in seven languages. In: Ekštein, K., Konopík, M., Pražák, O., Pártl, F. (eds.) Text, Speech, and Dialogue. pp. 275–286. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-02551-7_24

16. Peters, B., Martins, A.F.T.: Beyond characters: Subword-level morpheme segmentation. In: Nicolai, G., Chodroff, E. (eds.) Proceedings of the 19th SIGMORPHON Workshop on

Computational Research in Phonetics, Phonology, and Morphology. pp. 131–138. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.14

17. Pranjić, M., Robnik-Šikonja, M., Pollak, S.: LLMSegm: Surface-level morphological segmentation using large language model. In: Calzolari, N., Kan, M.Y., Hoste, V., *et al.* (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 10665–10674. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.933/

18. Rajapakse, T.C., Yates, A., de Rijke, M.: Simple transformers: Open-source for all. In: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. pp. 209–215. SIGIR-AP 2024 (2024). https://doi.org/10.1145/3673791.3698412

19. Sorokin, A.: Improving Morpheme Segmentation Using BERT Embeddings. In: Burnaev, E., Ignatov, D.I., Ivanov, S., *et al.* (eds.) Analysis of Images, Social Networks and Texts. pp. 148–161. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-16500-9_13

20. Sorokin, A., Kravtsova, A.: Deep convolutional networks for supervised morpheme segmentation of Russian language. In: Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) Artificial Intelligence and Natural Language. pp. 3–10. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01204-5_1

21. Tikhonov, A.N.: Word Formation Dictionary of the Russian language [Slovoobrazovatelnyi slovar russkogo yazyka]. Russkiy yazyk, Moscow (1990)

22. Wehrli, S., Clematide, S., Makarov, P.: CLUZH at SIGMORPHON 2022 shared tasks on morpheme segmentation and inflection generation. In: Nicolai, G., Chodroff, E. (eds.) Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 212–219. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.21

23. Zmitrovich, D., Abramov, A., Kalmykov, A., *et al.*: A family of pretrained transformer language models for Russian. In: Calzolari, N., Kan, M.Y., Hoste, V., *et al.* (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 507–524. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.45/