https://doi.org/10.17323/jle.2024.22237

Automatic Morpheme Segmentation for Russian: Can an Algorithm Replace Experts?

Dmitry Morozov [1], **Timur Garipov** [1], **Olga Lyashevskaya** [2,3], **Svetlana Savchuk** [3], **Boris Iomdin** [4], **Anna Glazkova** [5]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] HSE University, Moscow, Russia
[3] Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia
[4] independent researcher
[5] University of Tyumen, Tyumen, Russia

ABSTRACT

Introduction: Numerous algorithms have been proposed for the task of automatic morpheme segmentation of Russian words. Due to the differences in task formulation and datasets utilized, comparing the quality of these algorithms is challenging. It is unclear whether the errors in the models are due to the ineffectiveness of algorithms themselves or to errors and inconsistencies in the morpheme dictionaries. Thus, it remains uncertain whether any algorithm can be used to automatically expand the existing morpheme dictionaries.

Purpose: To compare various existing algorithms of morpheme segmentation for the Russian language and analyze their applicability in the task of automatic augmentation of various existing morpheme dictionaries.

Results: In this study, we compared several state-of-the-art machine learning algorithms using three datasets structured around different segmentation paradigms. Two experiments were carried out, each employing five-fold cross-validation. In the first experiment, we randomly partitioned the dataset into five subsets. In the second, we grouped all words sharing the same root into a single subset, excluding words that contained multiple roots. During cross-validation, models were trained on four of these subsets and evaluated on the remaining one. Across both experiments, the algorithms that relied on ensembles of convolutional neural networks consistently demonstrated the highest performance. However, we observed a notable decline in accuracy when testing on words containing unfamiliar roots. We also found that, on a randomly selected set of words, the performance of these algorithms was comparable to that of human experts.

Conclusion: Our results indicate that although automatic methods have, on average, reached a quality close to expert level, the lack of semantic consideration makes it impossible to use them for automatic dictionary expansion without expert validation. The conducted research revealed that further research should be aimed at addressing the key identified issues: poor performance with unknown roots and acronyms. At the same time, when a small number of unfamiliar roots can be assumed in the test dataset, an ensemble of convolutional neural networks should be utilized. The presented results can be used in the development of morpheme-oriented tokenizers and systems for analyzing the complexity of texts.

KEYWORDS

automatic morpheme segmentation, Russian language morphology, machine learning, convolutional neural networks, dictionary expansion, morphological analysis, natural language processing, expert-level performance

Citation: Morozov D., Garipov T., Lyashevskaya O., Savchuk S., Iomdin B., & Glazkova A. (2024). Automatic Morpheme Segmentation for Russian: Can an Algorithm Replace Experts? *Journal of Language and Education, 10*(4), 71-84. https://doi.org/10.17323/jle.2024.22237

Correspondence:
Dmitry Morozov,
morozowdm@gmail.com

Received: August 18, 2024
Accepted: December 16, 2024
Published: December 30, 2024

INTRODUCTION

Morpheme segmentation of a word is the process of breaking down the word into its smallest meaningful units called morphemes, for example, prefixes, suffixes, and roots. Many spelling rules taught in school rely on the student's ability to identify a morpheme or determine the relative position of several morphemes (Bakulina, 2012). In Russian, such rules include spelling of voiceless and sonorous consonants in prefixes, spelling of -н-/-нн- at morpheme boundaries and within them, searching for cognates to determine which vowel to write in unstressed syllables, where several phonemes may be pronounced the same, etc.

Morpheme segmentation can also be used in developing tools for automatic language analysis, both in creating a feature-based description of text, for example, in text complexity assessment (Morozov et al, 2024), and in developing language models as an alternative to Byte-Pair Encoding (BPE) tokenizers, which can improve model quality (Matthews et al, 2018). However, the proportion of words not described in morpheme dictionaries is significant: in one of the largest such dictionaries of Russian, the “Word Formation Dictionary of Russian language” (Tikhonov, 1990), there are segmentations for about 150 thousand different lemmas, while in the Main Corpus of the Russian National Corpus (Savchuk et al, 2024), there are over 250 thousand unique lemmas. Therefore, developing algorithms for automatic analysis is an urgent task.

For the Russian language, morpheme segmentation is complicated by the lack of a unified approach to segmenting words into morphemes (Iomdin, 2019). Some authors use the so-called Vinokur criterion (Vinokur, 1946) as a guide for segmentation. In this case, to cut a morpheme from a word, it is necessary to present a word-forming chain, that is, to find a word that, when supplemented with this morpheme, coincides with the word under consideration, e.g. пис-а-ть ‘write’ + -тель = пис-а-тель ‘writer’. This approach, for instance, is adopted in the aforementioned “Word Formation Dictionary” and is often used within school education. A significant drawback of this approach is that words which are considered related by native speakers, may turn out to be unrelated morphologically. Thus, in the word неодобрительный ‘disapproving’, the root is claimed to be -одобр-, while in добро ‘good’ it is -добр-, meaning these words are not cognates.

Other researchers, like the authors of the “Dictionary of Russian Language Morphemes” (Kuznetsova & Yefremova, 1986), prefer a more granular approach to morphemes and rely on comparability of a word with other lexemes of similar structure. For example, the word улыбаться ‘to smile’ features the -лыб- root, as the structure is parallel to the other verbs with у- (cf. у-смех-а-ть-ся ‘to grin’), and some words are analyzed etymologically (на-сек-ом-ое ‘insect’, вос-точ-н-ый ‘eastern’). The borrowings are split into morphemes (eg. ре-волюц-и-я ‘revolution’, квит-анци-я ‘receipt’) if they have semantic parallels to other borrowings with a comparable structure (cf. э-волюц-и-я ‘evolution’, рас-квит-а-ть-ся ‘to get even’).

However, studying dictionaries reveals that in specific cases, authors make decisions that contradict the established paradigm, such as in the segmentation of suffixes, e.g. за-воева-тель-н-ый ‘aggressive’ vs за-град-и-тельн-ый ‘barrage’ in (Tikhonov, 1990). Thus, the rules of morpheme segmentation represent a loosely formalized area, which likely makes it impossible to devise an absolutely error-free algorithm.

Nevertheless, since the task has sufficient practical potential, there are many automatic approximate approaches presented. One of the most commonly used and extensively described is a family of algorithms based on the Morfessor algorithm (Creutz & Lagus, 2002). This algorithm belongs to language-independent unsupervised and semi-supervised machine learning methods to be trained on a large text collection. Among the most relevant modifications of the original algorithm, it is worth mentioning the approach by S.-A. Grönroos et al (2020), which explores the combination of Morfessor with EM+Prune. Significant progress in the quality of algorithms has been achieved during the SIGMORPHON 2022 competition (Batsuren et al, 2022), where several approaches were presented that significantly outperformed the baselines including Morfessor, ULM (Kudo, 2018) and WordPiece (Schuster, & Nakajima, 2012). Among the proposed architectures are those based on Transformer models (Zundi & Avaajargal, 2022; Peters & Martins, 2022), GRU models (Levine, 2022), neural hard-attention transducer models (Wehrlie et al, 2022), LSTM networks (Peters & Martins, 2022; Girrbach, 2022), and Hidden Markov models (Bodnár, 2022). The team DeepSPIN (Peters & Martins, 2022) achieved the best quality across all nine languages involved. Their solutions are based on LSTM networks with a specific loss function (DeepSPIN-1 and DeepSPIN-2) and the Transformer architecture (DeepSPIN-3).

In the near future, a rapid increase in the number of approaches utilizing large language models is to be expected. Pranjić et al (2024) proposed an algorithm based on the Glot500-m network (ImaniGooghari et al, 2023), representing a binary classifier for determining morpheme boundaries in a word. However, the limitations of the algorithm, namely, relatively low quality on the English, Finnish, and Turkish datasets, as well as extremely long processing time (as the algorithm checks each pair of neighboring letters in a word), do not currently allow this approach to be considered a priority.

For the Russian language, the most relevant solutions superior to the Morfessor algorithm are presented by A. Sorokin & A. Kravtsova (2018), A. Sapin & E. Bolshakova (2019a; 2019b). The authors introduce approaches based on convolutional neural networks, long short-term memory networks, and gradient boosting over decision trees. The results of comparing algorithms on two different datasets do not allow for a definitive conclusion regarding the superiority of one algorithm over the others. However, the quality they achieve (about 90% of completely correct segmentations) is quite high. The Russian language was also among the languages at SIGMORPHON 2022, and the DeepSPIN-3 model achieved the best quality.

At the same time, a number of questions in this area remain insufficiently explored. Garipov et al (2023) found that a model based on convolutional neural networks has a significant drawback: its quality sharply decreases when tested on words containing roots that were absent in the training set, with the percentage of fully correct segmentations dropping by 17-18%. It remains unclear whether a similar issue exists for other algorithms demonstrating high quality.

Additionally, when developing a new algorithm or conducting a competition, typically only one morpheme dictionary per language is considered, whereas it makes sense to consider more dictionaries for a more representative study. Comparing the algorithms presented in various papers is further complicated by the fact that researchers are actually addressing different tasks. In some cases (Sorokin & Kravtsova, 2018; Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b), the task specifically focuses on segmenting the original string into morphemes, while in the SIGMORPHON competition, the task involved reconstructing "standardized" forms of morphemes. Cotterell et al. (2016) describes the difference between these approaches: the so-called "surface" segmentation is a sequence of surface substrings whose concatenation is exactly equal to the original word, e.g., funniest → funn-i-est, while during "canonical" segmentation, the task is not only computing surface segmentation but also restoring standardized forms of morphemes, e.g., funniest → fun-y-est.

The third issue is the impact of internal inconsistency among dictionaries on the quality of the algorithm. It is impossible to determine whether the quality of the models has already reached the expert level, and the remaining errors can be explained by the internal inconsistency in the training dataset.

Thus, the purpose of this research is to compare various existing algorithms of morpheme segmentation for the Russian language and analyze their applicability in the task of automatic augmentation of various existing morpheme dictionaries. We seek to answer the following research questions:

RQ#1: Which of the presented algorithms achieve the best results for the Russian language based on various morpheme dictionaries annotated in different paradigms?

RQ#2: How well can the presented algorithms parse words containing roots that were not encountered in the training data?

RQ#3: How does the quality of annotation by algorithms compare to the quality of annotation by expert linguists?

METHOD

Datasets

In our study, we used morpheme dictionaries where each word is segmented into morphemes with the type of each one indicated. A total of seven morpheme types are used: PREF (prefix), ROOT (root), SUFF (suffix), END (ending), POST (postfix), LINK (linking vowel), and HYPH (hyphen). To ensure a high representativeness in the study, we utilized three morpheme segmentation datasets annotated in different paradigms:

(1) Morphodict-K: dataset based on the "Dictionary of Morphemes of the Russian Language" (Kuznetsova & Yefremova, 1986), used in the Main Corpus of the Russian National Corpus. Rules of segmentation is that of strong albeit not maximal splitting of morphemes and correspondences to other words with similar structure.

(2) Morphodict-T: dataset based on the "Word Formation Dictionary of Russian language" (Tikhonov, 1990). This dataset is used in the Educational Corpus of the Russian National Corpus. So-called Vinokur criterion is used as an algorithm for splitting words into morphemes. Morphemes in Morphodict-T are splitted in larger chunks than in Morphodict-K (улыб-а-ть-ся 'to smile', насеком-ое 'insect', восточ-н-ый 'eastern'), especially borrowings (революци-я 'revolution', квитанци-я 'receipt'). The vocabulary of the datasets also varies. For example, Morphodict-K dataset contains 75,649 words, of which only about 58,000 are present in the Morphodict-T one. Notably, Morphodict-T differs from the dataset utilized by A. Sorokin and A. Kravtsova (2018) in that it fixes many incorrect morpheme type annotations. Error detection and type correction were performed out by a team of three experts. A total of 31,468 segmentations were corrected. In cases of disagreement, the segmentations were discarded (27 cases in total).

(3) CrossLexica (Bolshakov, 2013): dataset used in (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). The rules of morpheme segmentation for this dataset are not described explicitly; however, in this small dataset there are differences from both Morphodict-K and Morphodict-T (Table 1). In the CrossLexica dataset, unlike the other two, there are no words with multiple roots, but there are a number of non-lemmatized words.

A brief description of the datasets is provided in Table 2.

Importantly, within the scope of the study, it was assumed that a word is exactly equal to the concatenation of its morphemes, which is generally incorrect. For example, the word горбунья 'female hunchback' can be parsed as горб*:ROOT/*ун*:SUFF/*ьj*:SUFF/*я*:END* (Kuznetsova & Yefremova, 1986) with an additional *-j-,* which is not written as a separate letter. In such cases, we modified segmentation: the *-j-* was excluded.

Table 1

Examples of markup differences between datasets

Word	Morphodict-K	Morphodict-T	CrossLexica
революция ‘revolution’	ре:PREF/волюц:ROOT/и:SUFF/я:END	революци:ROOT/я:END	ре:PREF/вол:ROOT/юци:SUFF/я:END
утверждать ‘to approve’	у:PREF/твержд:ROOT/а:SUFF/ть:END	утвержд:ROOT/а:SUFF/ть:END	у:PREF/твержд:ROOT/ать:END
собственник ‘owner’	соб:ROOT/ств:SUFF/енн:SUFF/ик:SUFF	собственн:ROOT/ик:SUFF	соб:ROOT/ств:SUFF/ен:SUFF/ник:SUFF

Table 2

Some characteristics of the datasets utilized

Characteristic	CrossLexica	Morphodict-T	Morphodict-K
Unique words	23426	95895	75649
Unique morphemes	2745	15899	8079
Unique roots	2256	15253	7148
Average morphemes per word	3.68	3.86	4.12
Average morpheme occurrence	25.14	23.29	38.56
Average root occurrence	8.31	7.54	12.24
Average root length	4.57	5.52	4.62

If after that -ь- became the only letter in the morpheme, we concatenated it to the previous morpheme. Therefore, in the considered case segmentation was simplified to горб*:ROOT/*унь*:SUFF/*я*:END*.

Another important feature of our work is that all the datasets utilized contain exclusively lemmata. This limits the applicability of the models trained during the experiments; however, it allows us to avoid spending resources on dealing with homonymy, as the homonymy of lemmata with different morpheme segmentation is a relatively rare occurrence in the Russian language.

Algorithms

Algorithms with Morpheme-Type Labeling

Among the algorithms with morpheme-type annotation, we selected three that showed the best quality in previous experiments: the convolutional neural networks ensemble (hereinafter CNN) (Sorokin & Kravtsova, 2018), the gradient boosting algorithm over decision trees (hereinafter GBDT), and long short-term memory network (hereinafter LSTM). Comparing these algorithms did not reveal a clear leader (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). To obtain a more comprehensive and objective comparison, we decided to replicate the experiment using the data from the three listed datasets. A small additional aspect of the study was the use of morphological features of words to improve the performance of the GBDT and LSTM algorithms by A. Sapin & E. Bolshakova (2019a; 2019b). We decided to investigate the impact of morphological features on the performance quality of these algorithms.

Thus, we investigated three morpheme segmentation algorithms with morpheme-type labeling:

(1) CNN. We used implementation from the original repository[1]. The model is an ensemble of three identical convolutional neural networks, each consisting of three layers with a window size of 5 and 192 filters. We trained the model for 25 epochs with early stopping set to 10.
(2) LSTM. We used implementation from the repository[2] without any changes.
(3) GBDT. We used implementation from the repository[3] without any changes.

Unfortunately, the required library versions were not specified in the repositories, so we were forced to use arbitrary ones.

[1] NeuralMorphemeSegmentation (Python library). A. Sorokin. https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

[2] RussianMorphParsing (Python library). A. Sapin. https://github.com/alesapin/RussianMorphParsing

[3] RussianMorphParsing (Python library). A. Sapin. https://github.com/alesapin/RussianMorphParsing

Each of the listed algorithms is a character-level classifier. Each character of the word is assigned a two-part label. The first part of the label indicates the position of the character within a morpheme: B for beginning (first but not last character in a morpheme), M for middle (neigther first nor last character in a morpheme), E for end (last but not first character in a morpheme), S for single (a single character in a morpheme). The second part of the label corresponds to the type of morpheme to which the character belongs. Thus, for слово 'word' with the segmentation слов*:ROOT/*о*:END*, the sequence assigned would be: B-ROOT, M-ROOT, M-ROOT, E-ROOT, S-END.

Segmentation-Only Algorithms

Most of the morpheme segmentation algorithms that have achieved high quality in the context of the Russian language are algorithms with morpheme-type labeling. However, in the SIGMORPHON competition in 2022 (Batsuren et al, 2022), morpheme segmentation was regarded as a task for nine languages, including Russian. The DeepSPIN team's algorithms (Peters & Martins, 2022) demonstrated the best quality, including for Russian, with the claimed approach quality being extremely high. At the same time, the dataset used in the competition largely consisted of word forms rather than lemmas, which could significantly impact the measured quality, especially because the training and test sets included word forms that differed only in endings. Additionally, the task was not about segmenting the provided string but about constructing the "canonical" segmentation, essentially involving the generation of a derivational chain from a base word. For example, for the word предугадывавшую 'foreseeing' in the dataset, a pseudo-segmentation "пред @@у @@гадать @@ывать @@вший @@ую" was assigned. Significant differences in the experimental setup and the dataset utilized make it impossible to compare the results of models presented in competitions with others. Therefore, we decided to study the performance quality of the best algorithm among those presented, the subword-level transformer model DeepSPIN-3, on our data.

Additionally, a model that extends the architecture from (Sorokin & Kravtsova, 2018) was presented by A. Sorokin (2022): instead of using character-level n-grams for word vectorization, pretrained subword embeddings from a BERT-like encoder are utilized. Direct comparison of the results of this model with the previously presented one is not feasible, as the model presented in the study lacks morpheme type annotation and has not been tested on Russian language data. To conduct a fair comparison, we decided to train the basic CNN ensemble model and BERT-extended one for tasks without morpheme-type labeling. Since the use of pretrained vectors could potentially help algorithm capture semantics, we hypothesized that this architectural modification would prevent a decrease in performance when tested on unfamiliar roots, as observed in (Garipov et al, 2023). Both of these algorithms, similar to the trio of morpheme type determination algorithms described above, classify individual characters without specifying the morpheme type.

Thus, we investigated three morpheme segmentation-only algorithms:

(1) DeepSPIN-3. We used implementation from the repository[4] without any changes. The vocabulary size was chosen as 4000 due to the insufficient amount of data. The remaining model hyperparameters were set according to the original paper.
(2) TorchCNN. CNN ensemble with n-grams. We used implementation from the original repository[5] without any changes.
(3) MorphemeBERT. CNN ensemble with subword BERT embeddings. We used implementation from the original repository[6] without any changes and the rubert-base-cased pretrained model as the source of embeddings (Kuratov & Arkhipov, 2019).

Experimental Setup

RQ1 Experiments

To address RQ1, we sequentially trained all models on all available datasets and measured their quality. To do this, we conducted five-fold cross-validation with random splitting. For the GBDT and LSTM models, three model variations were trained: 1) without using additional information apart from the word itself, 2) using parts of speech and lemmas, and 3) utilizing all available morphological information.

RQ2 Experiments

To address RQ2, we initially divided each of the available datasets into five approximately equal non-overlapping samples based on roots. To do this, we collected all roots present in the dataset and randomly splitted them into five groups. All words containing roots from Group 1 were included in Fold 1, and so on. Words with multiple roots were excluded from the dataset in advance. Subsequently, we conducted cross-validation of all models on this partitioning.

RQ3 Experiments

To tackle RQ3, we prepared four subsets of morpheme segmentations. The first and second subsets each included 50 random words from the Morphodict-T and Morphodict-K

[4] MorphemeSegmentation (Python library). J. Stephenson. https://github.com/joshstephenson/MorphemeSegmentation

[5] MorphemeBert (Python library). A. Sorokin. https://github.com/AlexeySorokin/MorphemeBert

[6] MorphemeBert (Python library). A. Sorokin. https://github.com/AlexeySorokin/MorphemeBert

datasets. This pair of dictionaries was selected because they differ most noticeably in annotation paradigm. In the third and fourth sets, we also included 50 words from the Morphodict-T and Morphodict-K dictionaries, but not randomly selected ones. Instead, we included words where the CNN model, trained on a random train-test split of the corresponding dataset, made errors in segmentation. Next, we asked four experts to parse each of these words according to the original annotation paradigm: words from the first and third sets according to the logic of the Word Formation Dictionary of the Russian language, and words from the second and fourth sets according to the logic of the Dictionary of Morphemes of the Russian Language. The experts were familiarized in advance with the Morphodict-K and Morphodict-T datasets and the principles of their compilation, but were not allowed to use additional sources of information during the annotation process. To achieve more objective results, random and potentially difficult-to-segment words were mixed, meaning Set 1 was mixed with Set 3, and Set 2 with Set 4. After the annotation, the sets were separated, and the results were calculated separately. We pairwise compared the annotations of the experts and the consistency of the experts' annotations with the dictionary version.

Metrics

To evaluate the quality of algorithms with morpheme-type annotation, we used metrics proposed in (Sorokin & Kravtsova, 2018): Precision, Recall, F-measure for morpheme boundary without considering their type, Accuracy for character annotation considering morpheme type and BMES annotation, and WordAccuracy — the proportion of fully correct segmentations. To evaluate the quality of solutions without morpheme-type annotation, we used character-level Accuracy and WordAccuracy. Additionally, for the DeepSPIN algorithm, we calculated the proportion of generated segmentations that do not match the original word after concatenation (for other algorithms, this metric is not meaningful as they involve character-level classification rather than sequence-to-sequence generation).

RESULTS

RQ1 Experiments

The results of evaluating LSTM and GBDT models are presented in Table 3. Here and further, for each metric (Accuracy, WordAccuracy), the maximum quality value obtained for each algorithm+dataset pair is typed in bold. It can be noticed that for the LSTM model, the use of additional information from all three datasets led to a decrease in quality. For the GBDT algorithm, the model quality improved, however, in two out of three cases, the improvement was very small. In addition, the model quality remained significantly lower than that of the LSTM algorithm. Since the use of additional morphological information did not lead to a significant change in the quality of the algorithms, further results are presented for LSTM and GBDT models without the use of additional morphological information.

The results of evaluating all six studied algorithms are presented in Tables 4 (algorithms with morpheme-type labeling) and 5 (algorithms without morpheme-type labeling; TCNN stands for the TorchCNN model, MBert stands for the MorpemeBERT model, DS-3 stands for the DeepSPIN-3 model). The results show that among the algorithms with morpheme-type labeling, an undisputed leader across all datasets and metrics is the CNN algorithm. In the case of algorithms without morpheme-type labeling, convolutional algorithms demonstrated similar results, but with an advantage for the MorphemeBERT algorithm. In 11-17% of cases, DeepSPIN-3 generated sequences that did not match the word after concatenation, and showed results 9-14% worse than CNN-based ones.

RQ2 Experiments

The results of evaluating algorithms with training data split by roots are presented in Tables 6 (algorithms with morpheme-type labeling) and 7 (algorithms without mor-

Table 3

Comparison of Quality of LSTM and GBDT Models with and without Additional Information

Metric	Variant	LSTM			GBDT		
		Morphodict K	Morphodict T	Cross Lexica	Morphodict K	Morphodict T	Cross Lexica
Accuracy	Base	**96.61**	**95.56**	**96.88**	88.84	86.88	92.26
	Lex+PoS	96.00	95.41	96.54	**88.96**	**87.29**	**92.37**
	Full	96.07	95.40	96.22	88.93	86.91	92.10
WordAccuracy	Base	**88.02**	**84.25**	**89.82**	64.43	58.63	75.25
	Lex+PoS	86.13	83.78	88.99	64.79	**60.01**	**75.62**
	Full	86.42	83.75	87.97	**64.84**	59.14	75.06

Table 4

Comparison of Quality of Models with Morpheme-Type Labeling in Five-Fold Cross-Validation with Random Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	**98.58**	98.00	91.88	**97.79**	97.22	89.62	**98.74**	98.03	93.50
Recall	**98.74**	98.30	94.69	**98.38**	97.54	93.34	**99.04**	98.33	96.85
F-measure	**98.66**	98.15	93.26	**98.09**	97.38	91.44	**98.89**	98.18	95.14
Accuracy	**97.40**	96.61	88.84	**96.61**	95.56	86.88	**98.10**	96.88	92.26
WordAccuracy	**90.82**	88.02	64.43	**88.49**	84.25	58.63	**93.60**	89.82	75.25

Table 5

Comparison of Quality of Models without Morpheme-Type Labeling in Five-Fold Cross-Validation with Random Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	12.22	-	-	11.32	-	-	17.02
Accuracy	97.43	**97.65**	86.07	96.80	**97.04**	86.21	98.01	**98.14**	81.83
WordAccuracy	89.42	**90.34**	80.89	86.00	**87.16**	78.28	91.99	**92.52**	78.43

Table 6

Comparison of Quality of Models with Morpheme-Type Labeling in Five-Fold Cross-Validation with Root-Based Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	95.35	93.91	90.79	94.46	93.89	88.61	94.67	93.95	90.25
Recall	95.04	94.32	92.61	94.96	93.21	92.09	95.68	94.09	93.33
F-measure	95.19	94.11	91.69	94.71	93.54	90.32	95.17	94.02	91.77
Accuracy	91.30	89.64	86.58	90.16	88.41	84.87	91.28	89.53	87.01
WordAccuracy	72.63	67.80	58.67	70.53	65.47	53.72	74.14	69.48	60.08
WA Drop	20.03%	22.97%	8.95%	20.30%	22.30%	8.37%	20.79%	22.64%	20.16%

Table 7

Comparison of Quality of Models without Morpheme-Type Labeling in Five-Fold Cross-Validation with Root-Based Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	74.52	-	-	56.06	-	-	84.49
Accuracy	92.03	92.37	22.41	91.99	91.90	39.09	92.45	93.32	13.73
WordAccuracy	69.69	71.03	14.55	67.63	67.24	25.59	71.03	74.03	9.20
WA Drop	22.06%	21.37%	73.96%	21.36%	22.85%	54.65%	22.79%	19.98%	83.22%

pheme-type labeling). An additional row indicates the decrease in quality based on the WordAccuracy metric compared to the random train-test split (in percentages, with quality under random train-test split taken as 100%). It can be seen that convolutional algorithms and LSTM decrease by 20-23%, GBDT decreases by 9-20%, and DeepSPIN-3 decreases significantly with a sharp increase in the invalid segmentations ratio. Comparing the decrease in quality between CNN and MBert, it can be observed that in two out of three cases, MBert decreased less, with the difference increasing as the training data decreased.

RQ3 Experiments

Tables 8-11 present the results of expert annotation for Samples 1-4, respectively. In each cell, the Accuracy and WordAccuracy metrics are separated by a delimiter |. The following observations are of particular interest:

1. The quality of expert annotation is comparable to the quality achieved by algorithms based on convolutional neural networks.
2. For all four samples, experts, ranked by quality relative to the benchmark, form a stable list: Expert 3 > Expert 4 > Expert 2 > Expert 1.
3. The agreement among experts is often lower than that with the benchmark, meaning that the differences from the benchmark vary among different experts.
4. The agreement between experts and the benchmark annotation depends much less on the source of a sample than on the word selection principle: for random words, the quality relative to the reference and the agreement among experts are significantly higher. Moreover, similar to automatic solutions, the quality is slightly higher for samples from Morphodict-K.

DISCUSSION

RQ1 Experiments

Since the best results for both types of algorithms were achieved by algorithms based on convolutional neural networks, we further examined the errors made by the CNN model.

It is worth noting that although the task with morpheme type identification is evidently more challenging than without it, this algorithm showed higher results in terms of Accuracy and WordAccuracy metrics compared to a similar architecture algorithm without morpheme type identification and its modification using BERT embeddings. We attribute this to two factors: firstly, the implementation of the algorithm from (Sorokin & Kravtsova, 2018) includes a set of heuristics that improve quality, and secondly, different frameworks (TensorFlow[7] in the first case, PyTorch[8] in the second one) and different library versions were used for the implementation in the original studies.

Earlier in Sorokin & Kravtsova (2018), it was found that some of the errors in the final algorithm were related to inconsistent labeling of training data and errors within them. This is confirmed by our observations. Studying cases where the model made errors, we found that the number of instances where the algorithm correctly identified morpheme boundaries but incorrectly selected their types is quite low — around 9% of all incorrect segmentations. These errors should primarily be attributed to the inconsistency in the dataset labels, as almost all of them occur in the choice between ROOT and PREF types in morphemes like ультра- 'ultra-', мега- 'mega-', супер- 'super-', and so on. In the Morphodict-K dataset, there are: seven cases of ультра*:PREF* and two cases of ультра*:ROOT*, six cases of мега*:PREF* and four cases of мега*:ROOT*, five cases of супер*:PREF* and 10 cases of супер*:ROOT*, and we could not justify the choice of a particular morpheme type based on the words. Thus, it can be considered that the task of determining morpheme types given the division of a word into morphemes can be solved with an accuracy close to 100%, provided there is consistency in the training dataset labels.

The need to increase consistency is also evidenced by errors related to the granularity of suffixes. Approximately 20% of cases show discrepancies between reference and generated segmentations where a pair of suffixes is combined into one, for example, н*:SUFF/*ик*:SUFF* versus ник*:SUFF*. Both variants are encountered in Morphodict-K, for instance, вечер*:ROOT/*ник*:SUFF* 'party', о*:PREF/*город*:ROOT/*ник*:SUFF* 'gardener', борт*:ROOT/*ник*:SUFF* 'beekeeper', and еже*:PREF/*год*:ROOT/*н*:SUFF/*ик*:SUFF* 'yearbook', не*:PREF/*год*:ROOT/*н*:SUFF/*ик*:SUFF* 'scoundrel', при*:PREF/*кла*:ROOT/*д*:SUFF/*н*:SUFF/*ик*:SUFF* 'applied scientist'. Therefore, it is necessary to address such inconsistencies in the dataset.

As in Sapin & Bolshakova (2019a), the errors in some cases can be addressed by using simple heuristics based on automatically identified morphology. For example, replacing the selected morpheme type END with SUFF for invariable parts of speech helped increase WordAccuracy by approximately 0.2%. However, the use of morphological information is unlikely to be considered a promising way to significantly improve quality. This is evidenced by the results of experiments with LSTM and GBDT models, where the use of morphological information led to a noticeable increase in quality only in the case of the GBDT model and the Morphodict-T dataset, while in other cases, it either had a weak impact or resulted in a slight decrease in quality.

[7] TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. A. Martin et al. https://www.tensorflow.org/

[8] PyTorch. A. Paszke et al. https://pytorch.org/

Table 8

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to The Reference Sample and Each Other. Sample 1: Morphodict-T, Random Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	90.79 | 70	90.79 | 72	97.05 | 92	96.69 | 88
Expert 1	90.79 | 70	-	89.87 | 66	89.69 | 68	92.82 | 72
Expert 2	90.79 | 72	89.87 | 66	-	89.69 | 70	93 | 76
Expert 3	97.05 | 92	89.69 | 68	89.69 | 70	-	94.66 | 84
Expert 4	96.69 | 88	92.82 | 72	93 | 76	94.66 | 84	-

Table 9

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample and Each Other. Sample 2: Morphodict-T, “Complex” Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	78.18 | 36	83.64 | 44	95.35 | 86	92.12 | 74
Expert 1	78.18 | 36	-	83.84 | 52	78.99 | 38	82.42 | 44
Expert 2	83.64 | 44	83.84 | 52	-	85.05 | 52	84.65 | 52
Expert 3	95.35 | 86	78.99 | 38	85.05 | 52	-	88.89 | 68
Expert 4	92.12 | 74	82.42 | 44	84.65 | 52	88.89 | 68	-

Table 10

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample And Each Other. Sample 3: Morphodict-K, Random Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	88.71 | 60	91.88 | 68	97.82 | 90	97.03 | 86
Expert 1	88.71 | 60	-	92.28 | 70	89.11 | 62	89.11 | 62
Expert 2	91.88 | 68	92.28 | 70	-	92.08 | 70	92.08 | 72
Expert 3	97.82 | 90	89.11 | 62	92.08 | 70	-	97.23 | 88
Expert 4	97.03 | 86	89.11 | 62	92.08 | 72	97.23 | 88	-

Table 11

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample And Each Other. Sample 4: Morphodict-K, “Complex” Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	82.05 | 46	82.69 | 44	95.51 | 86	94.02 | 80
Expert 1	82.05 | 46	-	76.71 | 32	80.77 | 46	85.04 | 54
Expert 2	82.69 | 44	76.71 | 32	-	81.62 | 42	83.97 | 50
Expert 3	95.51 | 86	80.77 | 46	81.62 | 42	-	88.68 | 62
Expert 4	94.02 | 80	85.04 | 54	83.97 | 50	88.68 | 62	-

A significant number of model errors are related to incorrectly defined word semantics and processing of abbreviations and acronyms (e.g., за*:PREF/*влаб*:ROOT* compared to the reference зав*:ROOT/*лаб*:ROOT* from заведующий лабораторией ‘head of laboratory’, во*:ROOT/*ен*:SUFF/*к*:SUFF/*ом*:SUFF* compared to the reference во*:ROOT/*ен*:SUFF/*ком*:ROOT* from военный коммисар ‘military commissar’). Interestingly, in some cases, the segmentations are linguistically valid, for example, пере*:PREF/*дом*:ROOT* can be derived from дом*:ROOT* ‘home, house’ like пере*:PREF/*груз*:ROOT* ‘overload’ from груз*:ROOT* ‘cargo’ (correct segmentation should be перед*:ROOT/*ом*:SUFF* ‘in front’), не*:PREF/*суш*:ROOT/*к*:SUFF/*а*:END* can be derived from суш*:ROOT/*к*:SUFF/*а*:END* ‘drying’ like не*:PREF/*у*:PREF/*вер*:ROOT/*енн*:SUFF/*ость*:SUFF* ‘uncertainty’ from у*:PREF/*вер*:ROOT/*енн*:SUFF/*ость*:SUFF* ‘confidence’ (correct segmentation should be нес*:ROOT/*ушк*:SUFF/*а*:END* ‘laying hen’). Errors related to the identification of the root boundaries constitute the majority also in Bolshakova & Sapin (2019a) and Bolshakova & Sapin (2019b). It is logical to assume that addressing these shortcomings can be partially achieved by using models of semantic vectors pretrained on large text corpora. This is supported by the comparison of the TorchCNN and MorphemeBERT models. With identical architectures, MorphemeBERT showed results 0.5-1% higher in terms of WordAccuracy metric on each dataset, which is consistent with the results obtained in Sorokin (2021) for six other languages.

Among other noteworthy results, it is important to highlight the significantly lower performance of LSTM and GBDT models compared to the original reports (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). In our case, the LSTM architecture did not outperform the CNN ensemble on any of the datasets. Another distinction was that the use of morphological features directly in the model had little impact on the quality of the labeling. We believe that, similar to the comparison of models based on convolutional networks, the reason may lie in the unfixed versions of the libraries used in the original repository. At the same time, as in Bolshakova & Sapin (2019a) and Bolshakova & Sapin (2019b), the quality of automatic segmentation on the CrossLexica dataset is higher than on the dataset based on Word Formation Dictionary of Russian language. Thus, despite some differences, our results align quite well with the previously obtained results, generalizing them to a larger number of algorithms and datasets.

The quality obtained by the DeepSPIN-3 algorithm also indicates significantly lower quality of generated parses. This is primarily attributed to substantial differences in dataset construction principles: in the SIGMORPHON competition, the dataset for the Russian language was approximately 10 times larger than Morphodict-K, but a significant percentage consisted not of lemmas but word forms, with different forms of the same word potentially appearing in both the training and test sets. The choice of this dataset construction approach might prove effective for using models as tokenizers, but it is not entirely clear whether it can be applied to expanding morpheme dictionaries. In the future, we plan to conduct additional research in this direction, supplementing our data with automatically collected and annotated word forms.

RQ2 Experiments

Analysis of the quality of algorithms with root-based train-test split showed that all considered algorithms experience a significant loss in quality in this setup, which is critical for an automatic expansion of a morpheme dictionary. This is consistent with the results obtained in Garipov et al. (2023) for the CNN ensemble and extends them to several algorithms that were previously unexplored from this perspective. The errors made by the CNN model in this scenario differ from those in the case of random splitting, as expected: in some cases the model attempts to identify known morphemes, leading to additional segmentation of the reference root in many cases, e.g. при*:PREF/*бран*:ROOT/*н*:SUFF/*ый*:END* compared to the reference при*:PREF/*бр*:ROOT/*а*:SUFF/*нн*:SUFF/*ый*:END* ‘tidy’ with instances of the root -бран- in the training set, such as in не*:PREF/*воз*:PREF/*бран*:ROOT/*н*:SUFF/*ый*:END* ‘unrestricted’. Hopes may lie in the use of pretrained language models, especially when dealing with small training dataset sizes.

RQ3 Experiments

To the best of our knowledge, there have been no previous comparisons of automatic morpheme annotation with expert annotation on Russian language data, so we conducted a detailed analysis of errors made by experts. This analysis revealed that in most cases, experts could have arrived at the reference segmentation through a combination of their annotations: at least two out of four experts produced a segmentation matching the reference in 45 out of 50 cases for Sample 1, 36 out of 50 cases for Sample 2, 45 out of 50 cases for Sample 3, and 40 out of 50 cases for Sample 4. However, in only six out of 200 cases did none of the experts provide a segmentation matching the reference: усердн*:ROOT/*ый*:END* ‘diligent’, чет*:ROOT/*в*:SUFF/*ер*:SUFF/*ич*:SUFF/*н*:SUFF/*ый*:END* ‘quaternary’ (Sample 1), о*:PREF/*свежева*:ROOT/*нн*:SUFF/*ый*:END* ‘skinned’, чет*:ROOT/*в*:SUFF/*ер*:SUFF/*ик*:SUFF* ‘quadruple’ (Sample 2), короб*:ROOT/*чат*:SUFF/*ый*:END* ‘box-shaped’, не*:PREF/*про*:PREF/*долж*:ROOT/*и*:SUFF/*тельн*:SUFF/*ый*:END* ‘short-lived’ (Sample 4). It is worth noting that errors in the reference annotation are possible in the mentioned cases: excessive granularity of the root in the case of чет*:ROOT/*в*:SUFF/*ер*:SUFF/*ич*:SUFF/*н*:SUFF/*ый*:END* ‘quaternary’ and чет*:ROOT/*в*:SUFF/*ер*:SUFF/*ик*:SUFF* ‘quadruple’, insufficient granularity of the root in the case of усердн*:ROOT/*ый*:END* ‘diligent’ (see усердие ‘diligence’ with no suffix -н-), a single suffix in the case of короб*:ROOT/*чат*:SUFF/*ый*:END* ‘box-shaped’ and не*:PREF/*про*:PREF/*долж*:ROOT/*и*:SUFF/*тельн*:SUFF/*ый*:END* ‘short-lived’ (despite the existence of variants -ч*:SUFF/*ат*:SUFF-* and -тель*:SUFF/*н*:SUFF-* in Morphodict-K, e. g. in сум*:ROOT/*ч*:SUFF/*ат*:SUFF/*ый*:END* 'marsupial' and у*:PREF/*по*:PREF/*доб*:ROOT/*и*:SUFF/*тель*:SUFF/*н*:SUFF/* ый*:END* 'similising').

Having classified the differences between expert and reference segmentations, we identified the following most common types of errors (with the number of such differences in parentheses):

- Sample 1 (Morphodict-T, random cases): root vs root+suff (9), root vs pref+root (8), root granularity (6), suff vs suff+suff (5)
- Sample 2 (Morphodict-T, "complex" cases): root vs root+suff (29), root granularity (14), root vs pref+root (11), suff vs suff+suff (10)
- Sample 3 (Morphodict-K, random cases): suff vs suff+suff (21), root vs root+suff (13), root vs pref (4)
- Sample 4 (Morphodict-K, "complex" cases): root vs root+suff (23), suff vs suff+suff (12), root vs pref+root (8), root vs root+link (7)

Here, **root vs root+suff** refers to cases where segmentations differ in the additional suffix extracted from the root, in **root vs pref+root** the additional prefix is extracted from the root, in **root vs root+link** a linking vowel is concatenated to the root, in **suff vs suff+suff** a suffix is splitted into two, root granularity refers to cases where segmentations differ in dividing a long root into multiple (>2 morphemes), **root vs pref** refers to cases where segmentations differ in the choice of PREF or ROOT morpheme type. The results confirm the conclusion drawn earlier from model error analysis: the rules for the granularity of root and suffix extraction are poorly formalized and contribute to discrepancies. The most frequent discrepancies, such as -тель*:SUFF/*н*:SUFF-* vs -тельн*:SUFF-*, -н*:SUFF/*ик*:SUFF-* vs -ник*:SUFF-*, -ич*:SUFF/* а*:SUFF-* vs -ича*:SUFF-*, lack consistent resolutions in both datasets and among experts.

Notably, the proportion of words marked as unknown by the experts was too small to draw conclusions about the quality of expert annotation in the case of unknown roots. In the future, we plan to conduct an additional experiment aimed at evaluating the quality in such cases.

Limitations

The main limitation of the study is the use of dictionaries containing exclusively or almost exclusively lemmata, rather than word forms. This is due to the fact that we were unable to find morpheme dictionaries of word forms of sufficient volume for training models. However, in applied tasks, it is often necessary to analyze word forms. Consequently, it seems necessary to search for or create a morpheme dictionary of word forms and re-evaluate the algorithms on its material.

Additionally, we were unable to compare the performance of the algorithms and experts on words containing unfamiliar roots, as we could not find enough words in the dictionaries utilized with roots unfamiliar to the experts.

CONCLUSION

Morpheme segmentation is in demand for language learning and natural language processing tasks. In last decades many algorithms for morpheme segmentation have been proposed. However, comparing the quality of different approaches is challenging due to differences in data and experimental setups. In our study, we conducted a comprehensive comparison of six state-of-the-art algorithms for the Russian language using three morpheme dictionaries with different segmentation paradigms. This allowed us to obtain representative results and determine how the quality of the algorithms relates. To assess the potential for improvement in the existing algorithms and understand the limitations imposed by inconsistencies in morpheme dictionaries, we compared the quality of the algorithms with that of expert annotations. Additionally, we investigated the previously identified significant drawback — a sharp decline in the quality of the algorithm when handling words with roots missing in the training dataset.

We found that the best performance across all datasets is achieved using an ensemble of convolutional neural network algorithms, and its quality can be enhanced by utilizing BERT embeddings. Error analysis of this algorithm revealed that many errors are related to inconsistent segmentation and labeling of morpheme types in the training set; handling of abbreviations and acronyms, ignoring word semantics. It has been confirmed that the performance quality of all examined algorithms significantly decreases when dealing with unknown roots, making it challenging to use these algorithms for automatic expansion of existing morpheme dictionaries.

The results obtained indicate that on a random sample of words, algorithms reach parity with expert markup in terms of quality. Errors made by experts are typically related to making localized decisions about the degree of granularity in segmentation, which, in our view, illustrates that morpheme segmentation for the Russian language is often precedent-based, relying on previously annotated cases, and cannot be unambiguously derived solely from the declared paradigm of morpheme segmentation.

Therefore, in the future, the focus should not be on increasing the average quality of the algorithms, but on addressing the key identified issues: poor performance with unknown roots, abbreviations, and acronyms. It is likely that considering word semantics and recognizing abbreviations can be achieved using language models pretrained on large text corpora. We plan to explore this possibility further. In addition, future research should explore the performance of the algorithms examined not only on lemmata but also on word forms of the Russian language. Currently, this is hindered by the limited availability of datasets for experimentation; however, recent works enable progress in this direction.

ACKNOWLEDGMENTS

We are grateful to Dmitry Sichinava for his advisory assistance and to Sofia Chizhevskaya for help with proofreading the text.

DECLARATION OF COMPETITING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Dmitry Morozov: Conceptualization, methodology, software, investigation, writing - original draft preparation.

Timur Garipov: Methodology, software.

Olga Lyashevskaya: Data curation, investigation, writing - reviewing and editing.

Svetlana Savchuk: Data curation.

Boris Iomdin: Data curation, writing - reviewing and editing.

Anna Glazkova: Methodology, validation, writing - reviewing and editing.

REFERENCES

Bakulina, G. A. (2012). Morfemnyy razbor slova: novye podkhody — novye vozmozhnosti [Morpheme segmentation: new approaches - new opportunities]. *Nachal'naya shkola*, 4, 29–32.

Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K., Giunchiglia, F., Cotterell, R., & Vylomova, E. (2022). The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 103–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.11

Bodnár, J. (2022). JB132 submission to the SIGMORPHON 2022 shared task 3 on morphological segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 152–156). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.17

Bolshakov, I.A. (2013). Krossleksika: universum sviazi mezhdu russkimi slovami [Crosslexica: a universe of links between russian words]. *Biznes-informatika*, *3*(25), 12–19.

Bolshakova, E., Sapin, A. (2019). Bi-LSTM model for morpheme segmentation of russian words. In Ustalov, D., Filchenkov, A., Pivovarova, L. (Eds.), *Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science* (pp. 151-160). Springer. https://doi.org/10.1007/978-3-030-34518-1_11

Bolshakova, E., Sapin, A. (2021). Building a Combined Morphological Model for Russian Word Forms. In Burnaev, E. et al. (Eds), *Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science* (vol. 13217, pp. 45-55). Springer. https://doi.org/10.1007/978-3-031-16500-9_5

Bolshakova, E.I., & Sapin, A.S. (2019). Comparing models of morpheme analysis for Russian words based on machine learning. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 104-113). Russian State University for the Humanities.

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 21–30). Association for Computational Linguistics. https://doi.org/10.3115/1118647.1118650

Cotterell, R., Vieira, T., & Schütze, H. (2016). A joint model of orthography and morphological segmentation. In *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. https://doi.org/10.18653/v1/N16-1080

Garipov, T., Morozov, D., & Glazkova, A. (2023). Generalization ability of CNN-based morpheme segmentation. In *2023 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 58–62). IEEE https://doi.org/10.1109/ISPRAS60948.2023.10508171

Girrbach, L. (2022). SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 124–130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.13

Grönroos, S.-A., Virpioja, S., & Kurimo, M. (2020). Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3944–3953). European Language Resources Association.

Imani, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., & Schütze, H. (2023). Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 1082–1117). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.61

Iomdin, B. L. (2019). How to define words with the same root? *Russian Speech*, (1), 109–115. https://doi.org/10.31857/S013161170003980-7

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 66–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1007

Kuratov, Y. & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 333–339). Russian State University for the Humanities.

Kuznetsova, A. I. & Efremova, T. F. (1986). *Dictionary of morphemes of the Russian language*. Russkii yazyk.

Levine, L. (2022). Sharing data by language family: Data augmentation for romance language morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 117–123). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.12

Matthews, A., Neubig, G., & Dyer, C. (2018). Using Morphological knowledge in open-vocabulary neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1, pp. 1435–1445). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1130

Morozov, D. A., Smal, I. A., Garipov, T. A., & Glazkova, A. V. (2024). Keywords, morpheme parsing and syntactic trees: Features for text complexity assessment. *Modeling and Analysis of Information Systems*, *31*(2), 206–220. https://doi.org/10.18255/1818-1015-2024-2-206-220

Peters, B. & Martins, A. F. T. (2022). Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 131–138). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.14

Pranjić, M., Robnik-Šikonja M., & Pollak, S. (2024). LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (pp. 10665–10674). ELRA and ICCL.

Savchuk, S. O., Arkhangelskiy, T., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Yu. N., Lyashevskaya, O. N., Orekhov, B. V., & Podryadchikova, M. V. (2024). Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, *2,* 7–34. https://doi.org/10.31857/0373-658X.2024.2.7-34

Schuster, M. & Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 5149–5152). IEEE. https://doi.org/10.1109/ICASSP.2012.6289079

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

Sorokin, A. (2022). Improving morpheme segmentation using BERT embeddings. In E. Burnaev, D. Ignatov, S. Ivanov, M. Khachay, O. Koltsova, A. Kutuzov, S.Kuznetsov, N. Loukachevitch, A. Napoli, A. Panchenko, P. Pardalos, J. Saramäki, A. Savchenko, E. Tsymbalov, & E. Tutubalina, (Eds.), *Analysis of images, social networks and texts* (pp. 148-161). Springer. https://doi.org/10.1007/978-3-031-16500-9_13

Tikhonov, A. N. (1990). *Slovoobrazovatel'nyi slovar' russkogo yazyka* [Word Formation Dictionary of Russian language]. Russkiy yazyk.

Vinokur, G. O. (1946). *Zametki po russkomu slovoobrazovaniyu* [Notes on Russian word formation]. *Izvestiya Akademii nauk SSSR. Seriya literatury i yazyka*, *V*(4), 317-317.

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

Zundi, T. & Avaajargal, C. (2022). Word-level Morpheme segmentation using Transformer neural network. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 139–143). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.15