An Experimental Study of Neural Morpheme Segmentation Models for Russian Word Forms

Elena Bolshakova[1][0000−0002−8659−5978]
and Alexander Sapin[2][0000−0002−9532−132X]

[1] Lomonosov Moscow State University,
National Research University Higher School of Economics, Moscow, Russia
eibolshakova@gmail.com
[2] Lomonosov Moscow State University, Moscow, Russia
alesapin@gmail.com

Abstract. Morphemic structure of words is useful for various NLP problems, in particular, for deriving a meaning of unknown words in languages with rich morphology, such as Russian. For Russian, several machine learning models for automatic morpheme segmentation of words were built, but only for parsing their lemmas. Meanwhile, significantly varying word forms are present in texts, among them unknown words are often encountered, and their lemmas are unknown. The paper reports on experiments for comparing two ways to automatically segment Russian word forms, both ways involve splitting into morphs and classification of resulted morphs. The former is based on a neural model trained on a data set automatically augmented with segmented word forms, the latter produces segmentation through predicted lemma and a pre-trained neural morpheme segmentation model for lemmas. It was shown that the models have comparable quality in morpheme segmentation with classification, and the model based on the augmented dataset slightly outperforms in word-level classification accuracy.

Keywords: morphological segmentation, morpheme analysis of Russian word forms, neural network models for morphology, morpheme segmentation with classification

1 Introduction

Morpheme segmentation as a kind of morphological analysis implies splitting words into constituent morphs, which are the surface forms of morphemes (roots and affixes), for example: *без-вкус-н-ый*, *taste-less*. Though the task of automatic morpheme segmentation was studied in early years of natural language processing (NLP), significant progress in its solution has appeared in recent years, when various machine learning techniques began to be applied.

Since morphemes are the smallest meaningful language units, information about morphemic structure of words is already in use in various NLP applications and auxiliary tasks, including machine translation [2], recognition of semantically related words (cognates, paronyms, etc.), creating derivational trees of words [10], constructing word embeddings [3] for handling rare and out-of-vocabulary words (by deriving their meaning based on distributional word vector representations) and so on.

Morpheme segmentation is especially topical and at the same time more difficult for languages with rich morphologies (such as Russian or Finnish). For morphologically rich languages with many affixes of various types and meanings, a more complicated task is relevant, which involves besides segmentation classification of segmented morphs. The main types of morphemes are Prefix, Root, Suffix, Ending, for example: *без*:PREF/*вкус*:ROOT/*н*:SUFF/*ый*:END, *taste*:ROOT/*less*:SUFF.

The first works on morpheme segmentation were pure statistical and dictionary-based [10]. Since during a long time only a small amount of words with labeled segmented morphemes was available for training, only unsupervised and semi-supervised machine learning techniques were applied, the most known solutions are implemented in Morfessor system [8, 12].

The task of morpheme segmentation with classification of segmented morphs remained almost unexplored until recent works [4, 5, 13] undertaken for Russian, due to powerful supervised machine learning techniques applied to relevant labeled data, first of all, the dataset from Tikhonov's derivation dictionary [15]. These works presented various supervised models with open-source code:

- Convolutional neural network (CNN) model[3] [13];
- Gradient boosted decision trees (GBDT) model[4] [4];
- Bidirectional long short-term memory (Bi-LSTM) neural model[5] [5].

The implemented methods consider the task of morpheme segmentation with classification as sequence labeling [14] and classify letters of words according to main types of morphs. As showed by comparative evaluation of the models, which was undertaken in the works [4, 5], they all achieve F-measure about 98–99% for detecting morpheme boundaries and they also show high accuracy of morpheme classification: up to 96–98% for letters, and about to 87–89% for whole words (depending on training datasets and model hyper parameters). Therefore, these models present state-of-the-art (SOTA) methods for the task of morpheme segmentation with classification.

However, these SOTA models for Russian were developed only for morpheme segmentation of lemmas (normalized forms of words), so far as only lemmas are present in the existing labeled datasets. Meanwhile, for morphologically rich and highly-inflecting Russian language, significantly varying word forms are present in texts, in particular, for verb *успеть* (to be in time) more than 15 its forms may be used: *успеют*, *успел*, *успели* and so on. Among various word forms, unknown ones are often encountered, and their lemmas are unknown. Since it turned out that the developed SOTA models work poorly for word forms, giving

[3] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation
[4] https://github.com/alesapin/GBDTMorphParsing
[5] https://github.com/alesapin/RussianMorphParsing

only about 30% for classification accuracy, we aimed to research segmentation methods applicable for word forms.

In the paper we describe and experimentally compare two ways to automatically segment Russian word forms, both ways involve splitting into morphs and classification of resulted morphs. The former is based on a neural model trained on a dataset automatically augmented with segmented and labeled word forms, the latter produces segmentation through predicted lemma and a pre-trained segmentation model for lemmas. It is unclear a priory, which of the ways is preferable, and to evaluate them, we have chosen CNN model as a core for both ways and have exploited an available dataset containing about 90,000 segmented words (lemmas) from Tikhonov's dictionary [15]. To train the model on word forms, we have extended this dataset by segmented word forms generated by an augmentation procedure we have developed.

Experimental evaluation has shown that the model trained on the augmented dataset (hereafter, model on word forms) and the model trained on lemmas and supplemented by the rules for segmenting the word form based on its segmented lemma (hereafter, hybrid model) have comparable quality in morpheme segmentation with classification (as well as comparable with quality of SOTA methods), while the model on word forms slightly wins in word-level classification accuracy with score 88%.

The paper starts with an overview of the main works on the morpheme segmentation, followed by explanation of our augmentation procedure and the resulted augmented dataset. Then our CNN model architecture and key issues of training the model on word forms are described, and the results of experiments with the compared models are reported and discussed. Finally, we present some conclusion.

2 Related Work

The earliest method of morpheme segmentation was proposed by Z. Harris in [9], it detects morpheme boundaries by letter variety statistics (LVS) [7]. Despite that the method showed only 61% of precision (tested on a small English dictionary), the statistics was useful in many subsequent researches of the task, in particular [4, 11].

In the next years, the most known solutions for morpheme segmentation were implemented in Morfessor system [8, 12], which exploits unsupervised machine learning methods to be trained on a large unlabelled text. The pure unsupervised method and its semi-supervised version that uses some labeled data in addition to the text collection give about 70–80% of F-measure for detected morpheme boundaries (tested on English, Finnish, and Turkish words).

Another kind of semi-supervised machine learning for morpheme segmentation [11] was based on conditional random fields (CRF), the task was considered as sequential classifying and labeling letters of a given word. Besides LVS values and features of letters, the developed CRF classifier exploits some data obtained by Morfessor, thus increasing F-measure on morpheme boundaries to 84–91%.

A pure supervised method with significantly better quality for the twofold task of morpheme segmentation with classification was proposed in [13], it was effective due to applying convolutional neural network (CNN) and training on the representative labeled data of Tikhonov's dictionary [15]. The task is considered as sequence labeling by classifying letters with 22 classes based on BMES labeling scheme: the classes account for beginning (B), middle (M), and ending (E) positions of a letter in the corresponding affix (prefix, root, suffix, postfix), as well as single (S) letter variants of affixes, and also hyphen and linking letter in multi-root and hyphenated words. The trained CNN model is supplemented with post editing of predicted classes by an auxiliary correcting procedure, which fixes some wrong sequences of classes, according to their probabilities. The model outperforms all previous morpheme segmentation models, giving F-measure up to 98% on morpheme boundaries and also achieving classification accuracy of 96% for letters and 88% for whole words.

Two more supervised machine learning models for morpheme segmentation with classification were developed for Russian words in recent works [4, 5]: the first is based on decision trees with gradient boosting (GBDT), while the second applies Bi-LSTM neural network. In both models, unlike the CNN model, the number of letter classes was reduced to 10, since the set of BMES labels is redundant even for recognizing successive affixes and roots. The GBDT classifier takes into account features of the letter (in particular, its position in the word and LVS values), features of its word (some morphological tags), and also window of 5 previous and 5 subsequent letters. The Bi-LSTM model [5] has three LSTM layers, the input includes one-hot encoded letters and also some morphological tags of the word being segmented. Both GBDT and Bi-LSTM morpheme segmentation models were trained and evaluated on two different datasets of Russian words segmented into labeled morphs, including Tikhonov's dataset.

Evaluation of these CNN, GBDT and Bi-LSTM models trained on the same Russian datasets has showed their comparable quality, about 98–99% of F-measure on morpheme boundaries and 96–98% of classification accuracy for letters and about 87–89% on words [4, 5]. For now, they are SOTA methods outperforming the previously developed ones, both for morpheme segmentation and for segmentation with classification. However, they were developed for segmenting lemmas (normalized word forms), not for various word forms encountered in texts. Therefore, it seems reasonable to study possible ways to build a more broad supervised model, and for this purpose, a dataset with word forms splitted into morphs is needed.

3 Data Augmentation

In order to build a dataset augmented with segmented word forms and thus suitable for training, we have developed a procedure that produces necessary segmentation of word forms based on known segmentation of the corresponding lemmas along with grammatical information about Russian word formation suffixes and about specific features of Russian inflection for words of various part of speech [16].

The dataset[6] based on Tikhonov's dictionary was the source of segmented and labeled lemmas, and various word forms for a particular lemma were taken from Open Corpora dictionary[7] [1]. The dataset encompasses 96,046 words (lemmas) of main part of speech: nouns, adjectives, verbs, adverbs. Segmented morphs of words are classified according main morpheme types of Russian language (prefix, root, suffix, ending, postfix), and successive prefixes and suffixes (if any) are labeled, for example, the verb *смазываться* (to lubricate) is segmented and labeled as *с*:PREF/*маз*:ROOT/*ыва*:SUFF/*ть*:SUFF/*ся*:POSTFIX.

While applying our augmentation procedure, all lemmas from Tikhonov's dataset were considered and their corresponding word forms from OpenCorpora were processed, but those dataset elements that are absent in Open Corpora dictionary were discarded (approximately, 5 thous. words, the most of them are very rare, such as *гофмейстерский*, *яспис*, *спассеровать*).

For a given word form to be segmented and its segmented lemma, the procedure applies segmenting rules depending on the part of speech and its subclass. For most nouns, adjective, and participles the rules are quite simple: in the general case, the given word form and lemma have some common beginning, and if the rest part of the lemma is labeled as ending, the rest part of the word form is also annotated as ending, whereas its common part copies segmentation and labels of the lemma. The following word pair illustrates the rule:

Lemma:	*разрумяненный*	*раз*:PREF/*румян*:ROOT/*енн*:SUFF/*ый*:END
Word form:	*разрумяненному*	*раз*:PREF/*румян*:ROOT/*енн*:SUFF/*ому*:END

However, for some subclasses of nouns and adjectives (words with a final yota: *ковбой – ковбоя*, *соболий – собольего*), short adjectives (*послушный – послушен*) nouns (words with fugitive vowels: *день – дня*, *замочек – замочка*), as well as for verbs, more difficult segmenting rules were elaborated.

Specifically, to segment personal verbal forms and gerund (e.g., *увидевши* – *у*:PREF/*вид*:ROOT/*е*:SUFF/*вши*:SUFF), after detection of the common part with infinitive form, the segmenting rules sequentially try to recognize and to label word-formative suffixes (*ова, ева, ыва, ива, вши, ев, ен, в, л*, and so on) and postfix (*ся*, *сь*) in the mismatching part of the given word form, and its rest part (if any) is classified as ending. Here is an example:

Lemma	*выходить*	*вы*:PREF/*ход*:ROOT/*и*:SUFF/*ть*:SUFF
Word form	*выходила*	*вы*:PREF/*ход*:ROOT/*и*:SUFF/*л*:SUFF/*а*:END

In such a way, our augmentation procedure has processed about 92% of wordforms. Some rare difficult cases were discarded, in particular, consonant alternation, but such discarding does not impact on the result of comparing. The resulted dataset augmented with segmented and classified word forms has a total size of 1,130,359 elements: 34% nouns, 32.35% adjectives and participles, 33.56% verbal forms, and 0.07% words of other POS.

[6] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation/tree/master/data

[7] http://opencorpora.org

The augmented dataset consists of inflectional paradigms for the processed lemmas (hereafter, inflectional groups), each group encompasses word forms for a particular lemma. Groups for nouns and adjectives are relatively small, while for verbs, a group includes all forms of present, future, and past tense, gerund forms, up to 31 elements. Here is a fragment of inflectional group for verb *обсыпать* (*to strew*):

обсыпать	*об:PREF/сып:ROOT/а:SUFF/ть:SUFF*
обсыпал	*об:PREF/сып:ROOT/а:SUFF/л:SUFF*
обсыпала	*об:PREF/сып:ROOT/а:SUFF/л:SUFF/а:END*
обсыпало	*об:PREF/сып:ROOT/а:SUFF/л:SUFF/о:END*
обсыпали	*об:PREF/сып:ROOT/а:SUFF/л:SUFF/и:END*
обсыплю	*об:PREF/сып:ROOT/л:SUFF/ю:END*
обсыпем	*об:PREF/сып:ROOT/ем:END*
обсыплем	*об:PREF/сып:ROOT/л:SUFF/ем:END*
обсыпешь	*об:PREF/сып:ROOT/ешь:END*
обсыплешь	*об:PREF/сып:ROOT/л:SUFF/ешь:END*
обсыпете	*об:PREF/сып:ROOT/ете:END*

4 Model Architecture

For our study of segmenting word forms and building morpheme segmentation models, among three SOTA models for morpheme segmentation, namely CNN, GBDT, and Bi-LSTM we have chosen convolutional neural network (CNN), because CNN is training much faster than others, and at the same time does not lose in quality. For simplification of experiments, in all our segmentation models we did not use the auxiliary correction procedure proposed for the original CNN model, as well as ensembles of several models [13]. Though such techniques improve quality of segmentation, but not significantly (1–2%), moreover, their application is not necessary for correct comparison of our model on word forms and hybrid model, as they use the same neural architecture.

All our trained CNN models for segmenting words (word forms) were implemented with Keras library [6] (based on Tensorflow). As model input we use letters represented in one-hot encoding format, complementing them with information about is a particular letter vowel or not, and also with POS tag of the word, which are taken from morphological analyzer, one-hot encoded and concatenated with letter vectors. To align all words to the same fixed length (20 letters), we evidently exploit padding, but with masking residual letters <PAD>(by excluding them while calculating errors), in order to avoid their influence on gradient descent. Thereby, one word is represented as an 1120 dimensional vector.

The model has several layers, the last layer is fully connected and completed with a softmax activation function, which outputs a probability distribution over all possible letter classes. The resulted classes of letters are obtained from probability distribution with argmax function. Similar to works [4, 5] we apply simplified (i.e., BE) labeling scheme of letters, with 11 classes.

Various hyperparameters of our CNN model were experimentally tested in preliminary experiments. The resulted model has four layers with 512 filters in each layer, dropout of 40%, ReLU activation function and kernel size of 5. More filters in a layer slightly improve the quality (less than 0.5%), but the model became too heavy both for training and for evaluation. As for additional layers, they also do not significantly improve quality: the model with three layers gives sufficient results, losing to four-layer network only about 1–2%. Among the gradient descent algorithms (Adam, RMSprop, SGD), the better results were shown by Adam.

5 Models on the Augmented Dataset

For all our experiments, the data sets (original Tikhonov's dataset and the augmented one) were randomly divided in proportion 70:10:20 for training, validation, and testing, respectively; the training subset of the augmented dataset includes 791 thous. word forms. After tuning the model with random splits, for correct evaluation of the models, we have fixed our training, testing and validation sets for reproducibility. All trained and evaluated models are freely available[8].

In experiments with training our CNN model on the augmented dataset, two different variants of random dividing the dataset were studied:

- Random mixing of labeled word forms and then splitting them to training and testing subsets;
- Random mixing of inflectional groups (each group consists of all word forms corresponding to the same lemma); and after that splitting to training and testing subsets is performed (thus, splitting does not divide the groups).

Thereby we have obtained two trained models, namely, the model on word forms with simple mixing and the model on word forms with group mixing, the results of their evaluation are presented in Tables 1, 2. Table 1 shows quality of only segmentation measured in precision, recall, and F-measure (computed as mean harmonic of the recall and precision).

Table 1. Evaluation of morpheme segmentation for models on word forms

Model:	Word Forms			Lemmas		
Training set	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Simple Mixing	**99.40**	**99.65**	**99.52**	**98.82**	**99.32**	**99.07**
Group Mixing	97.76	98.65	98.20	97.04	98.17	97.60
Only Lemmas	89.60	89.44	89.52	96.95	98.14	97.54

Table 2 corresponds to classification accuracy of the segmented morphs, for letters and for whole words, respectively. The former is the ratio of correctly

[8] https://github.com/alesapin/XMorphy

recognized classes of letters to the number of all letters, the latter estimates the ratio of completely correctly segmented words with true classes of all their letters.

For comparison, in the last lines of the Tables we have added scores of the CNN model trained only on lemmas taken from the augmented dataset (more precise, from its training subset). The scores show that this model significantly loses when applied to word forms: much worse F-measure on morpheme boundaries (89.52%) and even worse classification accuracy (81.19% and 34.30% for letters and word, respectively). At the same time, almost similar scores for lemmas confirm consistency of experimental settings.

Table 2. Classification accuracy for models on word forms

Model:	Word Forms		Lemmas	
Training set	**Letters**	**Words**	**Letter**	**Word**
Simple Mixing	**99.26**	**96.75**	**98.53**	**94.46**
Group Mixing	96.94	88.89	96.00	86.36
Only Lemmas	81.19	34.30	95.98	86.07

As for our models on word forms, the model with simple mixing outperforms its counterpart in all the scores (slightly on morphs boundaries and significantly in classification accuracy for words). The explanation is simple: since inflectional groups may be divided while mixing and splitting to training and testing subsets for the model with simple mixing, the testing subset can contain some word forms of the groups, whose elements are present in the training subset, and this improves evaluation results. At the same time, the quality of the model with group mixing is comparable with SOTA morpheme segmentation models built on lemmas. Therefore, it is not quite correct to compare the model with simple mixing with our hybrid model for segmenting word forms, and we have compared only the model with group mixing.

6 Comparison with the Hybrid Model

The hybrid model implements another way to segment word forms, which implies the following steps:

1. converting a given word form into its lemma;
2. segmenting the latter by the model trained on lemmas (in our experiments, by the model already learned and indicated in the last lines of Tables 1, 2);
3. transforming the resulted segmented lemma into a segmented word with the aid of the procedure and segmenting rules described in section 3.

Using the model already trained on lemmas, we have evaluated the proposed hybrid model, with precision, recall and F-measure on morph boundaries (for segmentation, the scores are given in Table 3), and also accuracy both for letters and whole words (see Table 4). For comparison, in these Tables we repeat scores of the model on word forms (trained with group mixing). It is important, that CNN network of the hybrid model was trained on the lemmas taken from the training dataset for the model on word forms, and it was evaluated on the same testing set.

Table 3. Evaluation of morpheme segmentation for hybrid and word forms model

Model:	Word Forms			Lemmas		
Training set	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Hybrid Model	97.37	98.44	97.90	96.95	98.14	97.54
Model on Word forms	97.76	98.65	98.20	97.04	98.17	97.60

One can notice that two our evaluated models for segmenting Russian word forms have highly close scores for morpheme segmentation, while for classification (Table 4), the model on word forms (group mixing) slightly wins both for letters and words.

Table 4. Classification accuracy for hybrid model and model on word forms

Model:	Word Forms		Lemmas	
Training set	**Letters**	**Words**	**Letters**	**Words**
Hybrid Model	96.51	87.28	95.98	86.07
Model on Word forms	**96.94**	**88.89**	96.00	86.36

Additionally, we have evaluated ratio of various errors in morpheme segmentation, depending on wrong boundaries between morphemes of various types, the results are presented in Table 5. In both models under comparison, the most frequent errors are related with wrong boundaries between roots and suffixes, almost half of the errors (column ROOT-SUFF in Table 5). Another types of frequent errors are wrong recognition of boundary between prefix and root (PREF-ROOT) and erroneous segmentation of successive roots (ROOT-ROOT) or suffixes (SUFF-SUFF). Below we present some examples of these types. In general, the presented statistics of errors are about the same, with rare errors of segmenting word endings.

- Root and suffix(ROOT-SUFF) – for verb *перетлевать*, the incorrectly segmented word form *пере*:PREF/*тл*:ROOT/*е*:SUFF/*ва*:SUFF/*ешь*:END instead of correct *пере*:PREF/*тле*:ROOT/*ва*:SUFF/*ешь*:END;
- Prefix and root (PREF-ROOT) – for adjective *подоблачный*, erroneous *под*:ROOT/*о*:PREF/*блач*:ROOT/*н*:SUFF/*ою*:END instead of the correct segmentation *под*:PREF/*облач*:ROOT/*н*:SUFF/*ою*:END;

- Successive roots and suffixes (ROOT-ROOT, SUFF-SUFF) – for adjective *трегубный*, instead of correct *тр*:ROOT/*е*:LINK/*губ*:ROOT/*н*:SUFF/*ого*:END, wrong segmentation variant: *тре*:ROOT/*губ*:ROOT/*н*:SUFF/*ого*:END.

Table 5. Types of errors in morpheme segmentation (%)

Model	**PREF-**	**PREF-**	**ROOT-**	**ROOT-**	**SUFF-**	**SUFF-**	**ROOT-**	**Other**
	PREF	**ROOT**	**ROOT**	**SUFF**	**SUFF**	**END**	**END**	
Hybrid	0.06	26.52	10.46	**51.36**	10.3	0.61	0.15	0.54
On Word forms	0.06	27.42	8.0	**49.02**	10.3	3.33	0.91	0.96

7 Conclusion and Future Work

We have developed and evaluated two models of morpheme segmentation with classification, which were proposed specifically for word forms and are important for morphologically rich and highly-inflective languages, such as Russian. The first model is purely supervised and built on the augmented dataset with labeling of constituent morphs, the second is the hybrid one combining both the supervised model based on lemmas and rules for segmenting word forms. For augmentation of existing dataset with labeled Russian lemmas we have created the rule-based procedure generating segmented word forms.

The quality of the developed models turned out to be comparable, and the model based on the augmented dataset is slightly better in word-level accuracy. This means, that both models can be used in various NLP experiments with Russian text. At the same time, the choice of the model may depend on its computational complexity important in particular applications. For some applied tasks, a three-layer CNN model instead of our four-layers CNN (as a core of the hybrid model) is more preferred, as it is faster to train and takes less memory.

Our future work implies:

- To resolve some inconsistencies and errors in the original Tikhonov's dataset, which have been observed while experimenting with it, in order to increase the quality of the built models for word forms;
- To elaborate additional segmenting rules for some unconsidered cases of word forms, such improvement of our augmentation procedure may be useful not only for improving the morpheme segmentation models, but also for other tasks.

References

1. Bocharov, V., Bichineva, S., Granovsky, D., Ostapuk, N., Stepanova, M.: Quality assurance tools in the OpenCorpora project. In.: Computational Linguistics and Intelligent Technologies: Papers from the Annual Int. Conference ”Dialogue 2011”, Issue 10, 101–109, Moscow (2011).

2. Botha, J., Blunsom, P.: Compositional morphology for word representations and language modelling. In: Proceedings of the 31th International Conference on Machine Learning (ICML), 1899–1907 (2014).
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146 (2017).
4. Bolshakova, E., Sapin, A.: Comparing models of morpheme analysis for Russian words based on machine learning: In.: Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conference "Dialogue 2019", Moscow, RGGU (2019).
5. 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. Proceedings of the Int. Conference AINL 2019, CCIS, 1119, 151–160. Springer, Cham (2019).
6. Chollet, F.: Keras: Deep learning library for theano and tensorflow, https://keras.io/, last accessed 2020/12/9
7. Çöltekin, Ç.: Improving Successor Variety for Morphological Segmentation. Lot Occasional Series, 16, 13–28. University of Groningen (2010).
8. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning, ACM Transactions on Speech and Language Processing, 4(1), Article 3 (2007).
9. Harris, S. Zellig: Morpheme boundaries within words: Report on a computer test. Transformations and Discourse Analysis Papers, 73, 68–77 (1967).
10. Lango, M., Žabokrtský, Z., Ševčíková, M.: Semi-automatic construction of word-formation networks. Language Resources & Evaluation (2020). https://doi.org/10.1007/s10579-019-09484-2
11. Ruokolainen, T. et al. : Painless semi-supervised morphological segmentation using conditional random fields. In: Proceedings of the 14th Conference of the European Chapter of the ACL, Short Papers, 84–89 (2014).
12. Smit, P., Virpioja, S., Gronroos, S., Kurimo, M.: Morfessor 2.0: Toolkit for statistical morphological segmentation. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the ACL, Gothenburg, 21–24 (2014).
13. Sorokin, A., Kravtsova, A.: Deep Convolution Networks for Supervised Morpheme Segmentation of Russian Language. In: Ustalov, D. et al.(eds) Artificial Intelligence and Natural Language. Proc. of the Int. Conference AINL 2018, CCIS, 930, 3–10. Springer, Cham (2018).
14. Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th Int. Conference on Neural Information Processing Systems, 2, 3104–3112 (2014).
15. Tikhonov, A.N.: Word Formation Dictionary of Russian language. Moscow, Russkij Yazyk Publ. (1990).
16. Zaliznjak, A.A.: Grammatical dictionary of Russian: Inflection. Moscow, Russkij Yazyk Publ. (1977).