Building Dataset and Morpheme Segmentation Model for Russian Word Forms

Bolshakova E.I.
Lomonosov Moscow State University
HSE, Moscow, Russia
eibolshakova@gmail.com

Sapin A.S.
Lomonosov Moscow State University
Moscow, Russia
alesapin@gmail.com

Abstract

The paper describes a way to generate a dataset of Russian word forms, which is needed to build an appropriate neural model for morpheme segmentation of word forms. The developed generation procedure produces word forms segmented into morphs that are classified by morpheme types, based on existing dataset of segmented lemmas and additional dictionary data, as well as fine-grained classification of Russian inflectional paradigms, which makes it possible to correctly process word forms with alternating consonants and fluent vowels in endings. The built representative dataset (more than 1,6 million word forms) was used to develop a neural model for morpheme segmentation of word forms with classification of segmented morphs. The experiments have shown that in detecting morphs boundaries the model has comparable quality with the best segmentation models for lemmas (98% of F-measure), slightly outperforming them in word-level classification accuracy (with score 91%).

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

DOI: 10.28995/2075-7182-2021-20-154-161

Построение датасета и модели морфемной сегментации для словоформ русского языка

Большакова Е.И.
МГУ имени М.В.Ломоносова,
НИУ ВШЭ, Москва, Россия
eibolshakova@gmail.com

Сапин А.С.
МГУ имени М.В.Ломоносова
Москва, Россия
alesapin@gmail.com

Аннотация

В статье описан способ генерации датасета с русскими словоформами, который необходим для построения соответствующей нейронной модели морфемной сегментации словоформ. Разработанная процедура генерации формирует словоформы, сегментированные на морфы, которые классифицируются по типам морфем, опираясь на существующий датасет сегментированных лемм и дополнительные словарные данные, а также дробную классификацию русских флективных парадигм, что позволяет правильно обрабатывать словоформы с чередованием согласных и беглыми гласными в окончаниях. Построенный представительный датасет (более 1,6 млн. словоформ) был применен для разработки нейронной модели морфемной сегментации словоформ с классификацией сегментированных морфов. Эксперименты показали, что при обнаружении границ морфов модель имеет сопоставимое качество с наилучшими моделями сегментации для лемм (98% F-меры), немного превосходя их по аккуратности классификации на уровне слов (91%).

Ключевые слова: морфологическая сегментация; морфемный анализ словоформ русского языка; нейросетевые модели морфологии; морфемная сегментация с классификацией

1 Introduction

Morpheme segmentation is a kind of morphological analysis, which implies breaking words into constituent morphs, the surface forms of morphemes (roots and affixes), for example, *taste-less*, Rus. *без-вкус-н-ый*. Morphemes are the smallest meaningful units of texts, so information about morphemic structure of words is helpful for various NLP problems, in particular, recognition of semantically related words: cognates with the same root, paronyms (words that have similar morphs but differ in meaning) and so on. In lexical semantics, morphemic structure of words may be exploited to overcome data sparseness inherent to natural languages. The work [2] shows that even simple subword information can improve distributional word vectors representations, therefore more accurate linguistic information about word structures is useful for deriving meaning of rare and out-of-vocabulary words.

The data sparseness problem is more complicated for languages with rich morphologies, such as Russian, which is a highly inflective language with many affixes (prefixes, suffixes, postfixes) of various types and meanings. Significantly varying word forms are present in Russian texts, among them unknown words are often encountered, and their lemmas are unknown. For morphology rich languages the task of morpheme segmentation of words is especially complicated task, as it requires not only splitting into morphs but also classification of resulted morphs by labeling their main types (Prefix, Root, Suffix, Ending), for example:

без:PREF/*вкус*:ROOT/*н*:SUFF/*ый*:END, *taste*:ROOT/*less*:SUFF.

The first works on automatic morpheme segmentation were pure statistical, either dictionary-based [8] or corpus-based [7]. For a long time, only unsupervised and semi-supervised machine learning techniques were applied for the task, because of absence of representative datasets with labeled segmented morphemes for training. The most known solutions were implemented in Morfessor system [7, 10], which performs only morpheme segmentation and exploits unsupervised machine learning methods to be trained on a large text collection, showing about 70-80% of F-measure for detected morpheme boundaries, for English, Finnish, and Turkish words.

Recently proposed work [9] presents a dictionary-based morpheme segmentation method supplemented by application of word vectors representations (word embeddings), but like the previous methods does not involve classification of segmented morphs and achieves no more than 85% F-measure for English words.

The problem of morpheme segmentation with classification of segmented morphs remained almost unexplored until recent works [4, 5, 11] undertaken for Russian, in which powerful supervised machine learning techniques were applied. The implemented methods consider the task of morpheme segmentation with classification as sequence labeling [12] and classify letters of words being segmented to main types of morphs. Relevant labeled data were exploited for training the segmentation models, and among them, the most volume dataset obtained from derivation dictionary [13], it contains about 96 thou. segmented words (lemmas, i.e. normalized forms of words). The trained high quality segmentation models rely on various approaches.

- Convolutional neural network model[1] (CNN) [11];
- Gradient boosted decision trees (GBDT) model[2] [4];
- Long short-term memory neural network model[3] (Bi-LSTM) [5].

Besides approaches, the implemented models differ in classification schemes: the CNN model was trained based on BMES labeling scheme with 22 classes of letters, accounting for beginning (B), middle (M), 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. Unlike the CNN model, in the works [4, 5] the number of letter classes was reduced to 10, since the set of BMES labels is redundant even for recognizing successive affixes and roots.

Evaluation of these CNN, GBDT and Bi-LSTM models trained on the same Russian datasets has shown their comparable quality [4, 5]: up to 98-99% of F-measure for morpheme boundaries (depending on training datasets and model hyperparameters), and about 96-98% of classification accuracy for letters and 87-88% for whole words. For now, these models present state-of-the-art methods for the considered task, outperforming the previously developed ones, both for morpheme segmentation and for segmentation with classification. However, they were developed only for segmenting lemmas, not for words in various grammatical forms encountered in texts. Meanwhile, for significantly varying Russian word forms (for verbs, up to 20 forms exist, differing in several affixes) the models for segmenting lemmas cannot work with similar quality. Thus, an appropriate morpheme segmentation mod-

[1] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

[2] https://github.com/alesapin/GBDTMorphParsing

[3] https://github.com/alesapin/RussianMorphParsing

el applicable for word forms is to be built, but there were no relevant datasets with segmented word forms.

In our work we have built a representative dataset of Russian word forms, which is suitable to train a supervised machine learning model for morpheme segmentation with classification. For this purpose, a generation procedure was developed, which produces word forms segmented into classified morphs, based on the available dataset with Russian lemmas from [13] and additional data from Russian morphological dictionaries. The latter includes, first and foremost, the fine-grained classification of Russian inflectional paradigms taken from the system [3], which makes it possible to correctly process many word forms with alternating consonants and fluent vowels in endings.

We have exploited the built dataset for training a neural model intended for morpheme segmentation of word forms, along with classification of segmented morphs. CNN architecture was chosen as a core of the model. Experiments have shown that the developed model achieves 98% of F-measure for detected morphs boundaries and also gives up to 91% of accuracy for classification of morphs in whole words, while the analogous CNN model trained on lemmas works poorly for word forms, giving only about 40% for word-level classification accuracy. Thereby, the quality of the developed model for word forms is comparable with the best supervised machine learning models for segmenting lemmas, even with slightly outperforming them in classification accuracy.

The paper first explains the generation procedure we have developed, as well as the generated dataset with segmented word forms and labeled morphs. Then the CNN model trained on this dataset is described, and the results of its experimental evaluation are reported and discussed. Finally, some conclusions are presented.

2 Generating Dataset with Segmented Word Forms

The developed procedure generates and segments word forms for given lemmas that are split into classified morphs and taken from the dataset[4] obtained from dictionary [13] (hereafter, Tikhonov's dataset), thus extending it. This dataset encompasses 96,046 words (lemmas) of main parts of speech (POS): nouns, adjectives, verbs, and adverbs. Segmented morphs of words are classified according to main morpheme types (Prefix, Root, Suffix, Ending, Postfix), hyphen (*чей-либо*) and also linking letter for multi-root words (e.g., *вод-**о**-наливной*); successive prefixes and suffixes (if any) are labeled, for example: the verb *полюбоваться* (*to admire*) is segmented and labeled as *по*:PREF/*люб*:ROOT/*ова*:SUFF/*ть*:SUFF/*ся*:POSTFIX.

The generation procedure depends on part of speech (POS) of input lemma and performs segmentation of its possible word forms, based on known segmentation of the lemma and grammatical information about Russian flexions and word formation suffixes. The significant problem was related to processing alternating consonants and fluent vowels in ending parts of word forms (in roots and end suffixes), mainly for nouns and verbs. Below we present examples.

Lemma *зверинец* – *звер:*ROOT/*ин:*SUFF/***ец***:SUFF
Word forms *звери**нца*** – *звер:*ROOT/*ин:*SUFF/***ц***:SUFF/*а*:END
*звери**нцу*** – *звер:*ROOT/*ин:*SUFF/***ц***:SUFF/*у*:END

Lemma *лечь* – *ле*:ROOT/*чь*:SUFF
Word forms *легла* – *лег*:ROOT/л:SUFF/а:END
но *ля**ж**ет* – *ля**ж***:ROOT/*ет*:END

For verbs, fluent vowels are encountered not only in endings but also in prefixes, e.g.:

Lemma *отмереть* – ***от***:PREF/***мер***:ROOT/*е*:SUFF/*ть*:SUFF
Word forms *отмер* – ***от***:PREF/***мер***:ROOT
но *отомрет* – ***ото***:PREF/***мр***:ROOT/*ет*:END

To overcome the problem and to correctly recognize morphs (prefixes, roots, suffixes) while segmenting, we have exploited not the known canonical inflection-class system for Russian [14], but the system of numerous inflectional classes from system CrossLexica [3]. CrossLexica's system includes 313 classes for nouns, 25 classes of adjectives (they are also encompass passive participles), and 289 classes for verbs. Approximately 35% of the classes describe words with alternating consonants or

[4] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

fluent vowels. Specification of each inflexion class includes endings (more precise, pseudo-flexions) of all word forms, and also an example of a word belonging to this class. Here is an example of noun inflexion class.

/ 96*/ {"ец","ца","цу","ец","цем","це","цу","цу"} ,* *\/*ранец*/*

The above mentioned words *зверинец* and *полюбоваться* have classes 96 and 34, respectively.

For our purposes, we have manually labeled boundaries between affixes in all endings of the classes, for example:

/ 96*/{ "ец-","ц-а","ц-у","ец-","ц-ем","ц-е","ц-у","ц-у"},* *\/*ранец*/*

Moreover, we have supplemented the specifications of the verbs classes with labeled gerund suffixes (*а/я, в,вши/ши)*, e.g.: *разлегшись − раз*:PREF/*лег*:ROOT/*ши*:SUFF/*сь*:SUFF, since they were not originally listed in CrossLexica's inflectional classes. For participles, the verbs classes specify only endings for active forms, whereas endings for passive participles are described in classes for adjectives. Here is an example of verb class.

/ 34*/ {"-ова-ть-ся","-ова-л-ся","-ова-л-а-сь","-ова-л-о-сь","-ова-л-и-сь",*
"у-ю-сь","у-ешь-ся","у-ет-ся","у-ем-ся","у-ете-сь","у-ют-ся",
"у-й-ся","у-йте-сь","-ова-вш-ий-ся","у-ющ-ий-ся", "у-я-сь","ова-вши-сь" },
*\/*жаловаться*/*

One can notice that labeled endings (pseudo-flexions) include word formation suffixes of verbs (*ова*/*ева*, *ыва*/*ива*, *вш/ш*, *уш/ющ*, *л*, etc.) and postfix (*ся*, *сь*). To enhance consistency in verbs forms, we have additionally replaced in the original dataset the last label SUFF in infinitives of verbs by label END (since there is no full agreement between linguists about classification of infinitive morphs *ть, ти, чь*), e.g., *от*:PREF/*мер*:ROOT/*е*:SUFF/*ть*:END.

While applying our generation procedure, all words (lemmas) from Tikhonov's dataset were considered. For each particular lemma, our procedure finds its inflectional class indicated in the CrossLexica's dictionary, generates all its word forms according to the endings of the found class and then segments each word form, based both on known labels of the lemma being processed and on data taken from the class specification. More precisely, the beginning part of the word form copies segmentation and labels of the lemma, while the rest part is segmented according to the ending from the class. The following pair of lemma and its word form illustrates the process:

Lemma *пожаловаться*: *по*:PREF/*жал*:ROOT/*ова*:SUFF/*ть*:END:/*ся*:POSTFIX

Word form *пожаловалась*: *по*:PREF/*жал*:ROOT/*ова*:SUFF/*л*:SUFF/*а*:END/*сь*:POSTFIX

If the given lemma to be segmented is absent in CrossLexica's dictionary, all necessary word forms are taken from Open Corpora dictionary[5] [1], and inflexion class is automatically restored by the following rule: the set of all endings for an assigned class should coincide with the endings set of OpenCorpora's word forms.

Lemmas absent both in Open Corpora and in CrossLexica's dictionary accounted for about 10% of Tikhonov's dataset. Nouns and adjectives were mostly also processed, but in semi-automatic manner: word forms were predicted by morphological processor CrossMorphy[6], their class was restored by the above-described rule, with the following manual validation and necessary correction. Some difficult cases of nouns and adjectives were discarded, as well as all verbs lemmas absent in both dictionaries (the most of them are very rare or even out of use, such as *каландрироваться, окулироваться*). Overall, only 1,950 lemmas of Tikhonov's dataset (less than 2 %) were omitted.

As a result, about 98% of lemmas from the source dataset were processed and a dataset with segmented and classified word forms was built, its total size is 1,613,047 elements: 28% nouns, 45% adjectives and participles, 27% verb forms, and 0.05% adverbs. The built dataset consists of groups, each group encompasses word forms for a particular processed lemma, hereafter we call such groups inflectional. Groups for nouns are relatively small (6 singular forms, 6 plural), and groups are larger for adjectives (24 elements) and for verbs (15-18 elements). A verb group includes all personal forms of present, future, past tenses, and imperative forms, as well as two forms of active participle and 1-3 forms of gerund. Below we present fragments of the group for verb *связать* (*to tie*):

[5] http://opencorpora.org

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

с:PREF/*вяз*:ROOT/*а*:SUFF/*ть*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*а*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*о*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*и*:END
с:PREF/*вяж*:ROOT/у:END
с:PREF/*вяж*:ROOT/*ешь*:END
... ...
с:PREF/*вяз*:ROOT/*а*:SUFF/*вш*:SUFF/*ий*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*в*:SUFF
с:PREF/*вяз*:ROOT/*а*:SUFF/*вши*:SUFF

While testing our generation procedure, we have manually verified some fragments of the resulting dataset, to make sure that it is correct. It has been observed quite many errors in labeling segmented words (lemmas) in the source Tikhonov's dataset, mainly in classification of root morphs. We have corrected more than 1,5 thou. errors, and the corrected version of the dataset with segmented lemmas is now freely available[7], as well as the created dataset[8] with word forms.

3 Neural Morpheme Segmentation Model for Word Forms

To build and evaluate a morpheme segmentation model based on the generated dataset with word forms, among the best approaches for morpheme segmentation, namely CNN, GBDT, and Bi-LSTM, we have chosen convolutional neural network (CNN), because CNN is much faster to train, without lose in quality. At the same time, we did not use the auxiliary correction procedure and ensembles of several models proposed for original CNN model [11], since in our work such techniques do not significantly improve quality of segmentation.

Our CNN model for segmenting word forms was implemented with Keras library [6] (based on Tensorflow). Any input word (word form) is represented as a vector: one-hot encoded letters concatenated with information about is a particular letter vowel or not, and also concatenated with POS tag of the word, which is taken from the morphological dictionaries. POS labels include nouns, adjectives, verbs (personal forms), participles (active forms), gerunds, and adverbs. Similar to works [5, 6] we apply simplified labeling scheme of letters, with 10 classes.

The resulted CNN model has three layers with 512 units in each one, dropout of 40%, and ReLU activation function. The last layer is fully connected and completed with a softmax activation function, which outputs a probability distribution over all possible letter classes. Preliminary experiments with various hyperparameters of the model have shown that additional layers do not significantly improve its quality (the model with three layers gives sufficient results, losing to four-layers network less than 1%). Among the gradient descent algorithms (Adam, RMSprop, SGD), the better results were shown by Adam with a fixed learning rate of 0.001.

For our experiments, the generated dataset was randomly divided in proportion 70:10:20 for training, validation, and testing, respectively. Two variants of random dividing the dataset and corresponding trained models were studied:

- Random mixing of all labeled word forms, with subsequent splitting them to training and testing subsets ― Model with Simple Mixing;
- Random mixing of inflectional groups (each group consists of all word forms corresponding to the same lemma), with subsequent splitting to training and testing subsets (thus, splitting does not divide the groups) ― Model with Group Mixing.

Besides these two implemented models, we have also trained CNN model (of the same architecture) only on lemmas taken from the generated dataset (more precise, from its training subset) − Model on Lemmas. Programming code of these implemented morpheme segmentation models is available at GitHub[9] (our training, testing and validation sets are fixed for reproducibility).

[7] https://github.com/cmc-msu-ai/NLPDatasets
[8] https://drive.google.com/file/d/1_-0zKmmr2MS8NhQee16dZcRWz7cAwkt2/view?usp=sharing
[9] https://github.com/alesapin/XMorphy/tree/master/scripts/rule

We have evaluated both the quality of segmentation and classification accuracy of our segmentation models, the results are given in Table 1 and Table 2, respectively. The last rows of the Tables correspond to Model only on Lemmas. All scores were computed twice: for all word forms and only for lemmas.

The quality of segmentation (cf. Table 1) was measured in precision (P) and recall (R) of morph boundaries and F-measure (computed as mean harmonic of the recall and precision, F1). One can see that Simple Mixing Model slightly outperforms its counterpart (Group Mixing) in all the scores for morphs boundaries, but both models for word forms are much better than Model on Lemmas in scores for word forms (99-98% compared with 86-90% of F1-measure). As for scores for lemmas, they are almost similar for Group Mixing Model and Model only on Lemmas.

Model:	**Word forms**			**Lemmas**		
Training Set	P	R	F1	P	R	F1
Simple Mixing	99.56	99.71	99.63	99.41	99.55	99.48
Group Mixing	98.22	99.05	98.63	98.16	98.95	98.55
Only Lemmas	86.56	90.67	88.57	98.06	98.53	98.30

Table 1: Evaluation of morpheme segmentation for word forms and lemmas (%)

Table 2 corresponds to classification accuracy of the segmented morphs, for letters and for whole words. The former is the ratio of correctly 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. Simple Mixing Model again outperforms its counterpart in all the scores, slightly for letters and significantly for words (97.34% and 91.06%). We can explain this as follows: since for the Simple Mixing Model inflectional groups may be divided while splitting to training and testing subsets, the testing subset may contain some word forms of the groups, whose elements are present in the training subset, and this improves evaluation results.

Thus, the group mixing is the more proper way of training and evaluating models on word forms, and scores of our Group Mixing Model are more adequate. This is additionally confirmed by comparing Group Mixing Model and Model only on Lemmas: their quality with respect to lemmas are almost similar, both in morpheme boundaries (Table 1, 98.55% and 98.30% of F1-measure) and accuracy in letters (Table 2, 97.54% and 97.13%); and at the same time these scores are highly close to those of the best neural morpheme segmentation models [5, 11].

Model:	**Word Forms**		**Lemmas**	
Training Set	Letters	Words	Letters	Words
Simple Mixing	99.42	97.34	99.15	96.40
Group Mixing	97.66	91.06	97.54	91.03
Only Lemmas	81.67	41.02	97.13	89.32

Table 2: Classification accuracy for word forms and lemmas (%)

It is surprising that our Group Mixing Model shows better results in classification accuracy, both for all word forms and for lemmas, than state-of-the-art results for lemmas: 91% for word-level accuracy compared with 87-88% [4, 5, 11]. In our opinion, the main reason is related to the learned knowledge: trained patterns for word forms help to more correctly parse lemmas. Other factors may also have influenced: exploiting corrected Tikhonov's dataset with lemmas and accounting for various POS during training. Apparently, these factors also improved quality of our Model only on Lemmas, for word-level classification accuracy: 89.32 % instead of 87-88% of state-of-the-art results [5, 11].

The last rows of Tables 1, 2 show that the Model only on Lemmas significantly loses when applied to word forms: much worse F1-measure on morpheme boundaries (88.57%) and even worse classification accuracy (81.67% for letters and 41.02% for words). It means, that for highly inflective languages, segmentation of word forms should be performed by models trained on relevant datasets.

4 Conclusions and Future Work

We have built the representative and volume dataset with Russian word forms split into morphs classified by main morpheme types. The rule-based generation procedure developed for this purpose relies on the analogous dataset containing only segmented lemmas, as well as on several known and proven morphological resources.

In our work, the built dataset was intended specifically to implement a neural morpheme segmentation model for word forms, which is important for morphologically-rich and highly inflective Russian. Experimental evaluation of implemented CNN models for segmenting word forms have shown their comparable quality with the state-of-the-art models for lemmas, and the properly trained model for word forms (Group Mixing Model) even outperforms the state-of-the-art results obtained for lemmas, giving about 91% in word-level classification accuracy.

The built dataset, programming code of the generation procedure, and the implemented neural models are of free access, as well as the corrected version of Tikhonov's dataset with lemmas. We hope they can be useful for other NLP tasks and experiments with Russian texts.

In our opinion, further progress in automatic morpheme segmentation for languages such as Russian may be achieved only after accounting some phonological features of words in training datasets, including introduction of iota sign [j] and two variants of consonants (hard and soft). Another interesting task for research involves creating a combined machine learning model performing both traditional inflectional morphological analysis and morpheme segmentation of a given word form.

Acknowledgements

We thank the author of CrossLexica for the provided dictionary resources.

References

[1] Bocharov V., et al. (2011), Quality assurance tools in OpenCorpora project [Instrumenty kontrolya kachestva dannyh v proekte Otkrytyj Korpus], Computational Linguistics and Intelligent Technologies: Papers from the Annual Int. Conference "Dialogue 2011", Bekasovo [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011"], Moscow, pp.101–109.

[2] Bojanowski P., Grave, E., Joulin, A., Mikolov, T. (2017), Enriching word vectors with subword information, Transactions of the Association for Comp. Linguistics, 5, pp. 135–146.

[3] Bolshakov I.A. (2013), CrossLexica – Universum of links between Russian words [CrossLexica – universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp.12–19.

[4] Bolshakova E., Sapin A. (2019), Comparing models of morpheme analysis for Russian words based on machine learning, Computational Linguistics and Intellectual Technologies: Proceedings of the Int. Conference "Dialogue 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2019"], Moscow, pp. 104–113.

[5] Bolshakova E., Sapin A. (2019), Bi-LSTM Model for Morpheme Segmentation of Russian Words // Artificial Intelligence and Natural Language: Proceedings of the conference AINL 2019, CCIS, vol. 1119. Springer, Cham, pp. 151–160.

[6] Chollet F. (2015), Keras: Deep learning library for theano and tensorflow. Access mode: https://keras.io/

[7] Creutz M., Lagus K. (2007), Unsupervised models for morpheme segmentation and morphology learning // ACM Transactions on Speech and Language Processing, 4 (1), Article 3.

[8] Harris S. Zellig (1967), Morpheme boundaries within words: Report on a computer test // Transformations and Discourse Analysis Papers, 73, pp. 68–77.

[9] Sakakini T., Bhat S., Viswanath P. (2017), MORSE: Semantic-ally Drive-n MORpheme SEgment-er // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 552–561.

[10] Smit P., Virpioja S., Gronroos S., Kurimo M. (2014), Morfessor 2.0: Toolkit for statistical morphological segmentation // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the ACL, Gothenburg, pp. 21–24.

[11] Sorokin A., Kravtsova A. (2018) Deep Convolution Networks for Supervised Morpheme Segmentation of Russian Language // Proceedings of the Conference on Artificial Intelligence and Natural Language, AINL 2018, St-Petersburg, Springer, Cham, pp. 3–10.

[12] Sutskever I., Vinyals O., Le Q. V. (2014), Sequence to sequence learning with neural networks // Advances in neural information processing systems, pp. 3104–3112.

[13] Tikhonov A.N. (1990), Word Formation Dictionary of Russian language [Slovoobrazovatel'nyj slovar' russkogo yazyka], Moscow, Russkij yazyk Publ.

[14] Zaliznjak A.A. (1977), Grammatical dictionary of Russian: Inflection. [Grammaticheskij slovar' russkogo yazyka], Moscow, Russkij yazyk Publ.