A Morphological Processor for Russian with Extended Functionality

Elena I. Bolshakova[1,2](✉) and Alexander S. Sapin[1]

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

Abstract. The paper presents an open-source morphological processor of Russian texts recently developed and named CrossMorphy. The processor performs lemmatization, morphological tagging of both dictionary and non-dictionary words, contextual and non-contextual morphological disambiguation, generation of word forms, as well as morphemic parsing of words. Besides the extended functionality, emphasis is put on linguistic quality of word processing and easy integration into programming projects. CrossMorphy is fully implemented in C++ programming language on the base of OpenCorpora vocabulary data. To clarify the reasons of its development, a comparison of several freely available morphological processors for Russian is given, across their linguistic and some technological properties. The experimental evaluation shows that CrossMorphy ensures rather high quality of word processing.

Keywords: Morphological tagging
Morphological parsers for Russian
Functionality of morphological processors
Morphological disambiguation · Morphemic parsing

1 Introduction

Morphological analysis of texts is a traditional task of computational linguistics and natural language processing (NLP). Almost any NLP system needs lemmatization and morphological tagging of word forms. For Russian language, methods of formal description of Russian morphology have been long known, and main problems of automatic morphological analysis are considered principally solved. However, some attendant and related problems are not fully solved, in particular, automatic morphemic analysis and morphological disambiguation. The latter is more complicated for languages with rich morphologies, such as Russian, and needs to be further investigated, since rather few topical works are known [1,9–12,14].

W. M. P. van der Aalst et al. (Eds.): AIST 2017, LNCS 10716, pp. 22–33, 2018.
https://doi.org/10.1007/978-3-319-73013-4_3

Nowadays, more than a dozen morphological processors for Russian are known, including freely available ones: AOT[1], Mystem[2], TreeTagger[3], Pymorpy2[4]. The processors differ in their functionality and also in technological features. Most processors are appropriate for majority of NLP researches and applications that do not require any deep analysis of texts (e.g., categorization of texts). Nevertheless, for more complicated applied tasks (such as information extraction or question answering), morphological parsers with a specific combination of properties are needed. From this point of view, the set of freely available morphological processors for Russian is not complete.

In our research project on lexico-syntactic patterns language [4] intended to build various NLP applications on the basis of surface syntactic analysis, we used open source processor AOT[1]. Unfortunately, it is not supported now and has some weak spots. The rest freely available morphological parsers are also not suitable: we lack a processor with the particular functionality and at the same time with the ability to integrate it in our project. So we were forced to begin development of our own morphologic processor, with emphasis to extended functionality, linguistic quality of word processing, and easy integration into programming projects. We suppose that yet another open source module with the particular linguistic and technological properties will be useful not only for us, and our efforts are a step towards collection of high-quality open-source morphological tools for Russian useful for various applications.

In this paper we present the developed morphological processor CrossMorphy[5] that is open source software fully implemented in C++ language and based on freely available data of Open Corpora[6]. To clarify CrossMorphy's peculiarities and reasons of its development, we begin with a comparison of the most used and freely available morphological parsers for Russian. We consider their properties including pure linguistic (such as lemmatization, morphological tagging, generation of word forms) and also some technical features important for our purposes (such as ability to integrate source code of processor to NLP programming project or to connect a specific dictionary). Then we explain main decisions undertaken while developing CrossMorphy and describe its functionality, which encompasses, besides main linguistic properties, morphological disambiguation and morphemic parsing. Evaluation of the described CrossMorphy's functionality shows sufficient quality of performed word processing.

2 Comparison of Morphological Parsers for Russian

We consider the most popular morphological processors that are freely available (so they can be tested) and are also frequently used in research projects,

[1] http://aot.ru/docs/rusmorph.html.
[2] https://tech.yandex.ru/mystem/doc/.
[3] http://corpus.leeds.ac.uk/mocky/.
[4] http://pymorphy2.readthedocs.io/en/latest/index.html.
[5] https://github.com/alesapin/XMorphy.
[6] http://opencorpora.org.

namely: AOT, Mystem, TreeTagger, Pymorpy2. Meanwhile, they demonstrate existing variety in functionality and approaches to build morphology models. The standard functionality of morphologic processors encompasses:

– lemmatization or/and stemming of a given word form;
– tagging its morphological features, first of all, POS (part of speech) and also gender, case, person, time, etc.;
– sufficient coverage of lexicon, which depends on used morphology model; for dictionary models it involves an ability to classify unknown (non-dictionary) words;
– generation of necessary word forms (or the whole word paradigm) for a given lemma.

We should note that in last two decades a trend appeared and settled to additionally provide parsers with properties that were earlier implemented by separate modules, namely:

– preliminary tokenization (and even sentence segmentation) of the text to be morphologically analyzed;
– morphological disambiguation of output parsing variants.

The reasons are obvious: traditional stages of text analysis, such as text segmentation, morphological analysis, syntax parsing correspond to language levels, which are internally interconnected. The results and quality of morphological analysis often strongly depend on text segmentation results: in Russian, typical examples are hyphen words, including specific terms, e.g. *α-редукция (α-reducing)*, *интернет-новости (internet news)* – the hyphen is often omitted and should be restored. Specific writing forms of numerals, e.g., *3-й*, *32-ая (3rd, 32nd)*, are also need special rules of processing, which are easier to implement as initial step of morphological analysis. As for morphological disambiguation, it facilitates subsequent syntactic analysis.

Besides above-mentioned pure linguistic properties of morphological processors, several more technological features are no less important for research projects and development of particular NLP applications. By technological features we mean:

– tools for modifying or/and extending morphological dictionary, as a rule, it means certain ability to connect a specific dictionary of your own;
– open source code, which makes it possible to integrate the source code of morphological parser into NLP programming project.

Both linguistic and technological properties are important to make an appropriate choice of tools for morphological processing in a particular application.

Comparing parsers AOT, Mystem, Tree Tagger, Pymorpy2 across the linguistic features, one can see that all of them perform lemmatization, full morphological tagging, and processing of non-dictionary words. Almost all compared parsers (except TreeTagger) are built on dictionary morphology models, *while TreeTagger is built by training on tagged corpus* [13]. At the same time, the dictionary-based parsers differ in accepted model of Russian morphology involving syntactic classes of words (such as POS), and as a result, they have different systems of morphological tags and rules of lemmatization [8]. In particular, Mystem partially retains the canonical morphological paradigm inherited from Zaliznyak's grammar dictionary [16], and for word form *пропил (drunk)* it gives lemma with changed verb aspect *пропивать (drink)* instead of expected lemma *пропить (drink away)*, which is output by AOT and Pymorpy2.

The differences also concern processing of new (non-dictionary) words. In the parsers under comparison, prediction of lemma and morphological tags are based on various heuristics rules, so the results may essentially vary, for example, from four parsing variants for *Пикачу* (*Pikachu*) in Pymorpy2 and only one variant in Mystem.

Morphological disambiguation is important for Russian, since morphological homonymy is a hard problem for all higher flexional languages: in Russian texts, for almost each word form it is necessary to choose from 2-5 parsing variants generally differing in part-of-speech (POS), lemma and grammatical properties. Morphological disambiguation is absent in AOT parser, the other parsers implement various methods: non-contextual disambiguation in Pymorphy2 and more reliable statistical contextual disambiguation in MyStem and TreeTagger.

Generation of correct word forms is a more rare function of the parsers (it is not required in many NLP tasks), it is incorporated in AOT and Pymorphy2, while absent in MyStem and TreeTagger.

As for technological features, only two parsers, AOT and Pymorphy2 have an open source code, and only MyStem permits connection of a specific dictionary (by replacing the main dictionary, which is often not suitable).

Thus, the parsers under comparison vary in linguistic and technological features, and choice of a parser adequate for a particular NLP task may be difficult because of absence of nessasary functionality. For development of our project based on lexico-syntactic patterns for building information extraction applications on the basis of surface syntactic analysis [4], we need an open source dictionary-based morphological processor with the main linguistic functions (lemmatization, morphological tagging, disambiguation), and also with stemming and word generation (in order to extract word phrases in correct grammatical form). Morphemic parsing of words are needed for our purposes as well, this makes it possible to recognize semantically close words (with the same root and some different affixes), such as *сахарный* and *сахаристый* (*sugar* and *sugary*), as well as words having different POS but indicating the same concepts, such as *компиляция* and *компилятор* (*compilation* and *compiler*).

Initially, we used processor AOT with open source code in our project. However, it is not supported now, its dictionary contains many obsolete words whereas does not include many new words, moreover, it does not provide morphological disambiguation. Among the other considered parsers, Pymorphy2 [7] has nearly sufficient functionality, but it provides simplest tokenization (it outputs only Russian words, the other tokens are skipped), it does not perform contextual disambiguation, and it is implemented in interpretive programming language Python, which complicates its integration into projects in other programming languages. For these reasons, morphological processor CrossMorphy with the desired functionality has been built.

3 CrossMorphy: Key Decisions and Main Functions

Key decisions involve the choice of a computer model for Russian morphology and corresponding vocabulary data. Among the known models, the dictionary models based on large lists of possible word forms or word stems are traditionally used because of their linguistic quality (in particular, Zaliznjak's canonical model and dictionary [16] were implemented in almost all first known morphological analyzers including Russian version of Microsoft Word).

Initially, we made an attempt to build morphological processor based on the morphological model used in CrossLexica system [3], since it encompasses wide Russian lexicon significantly renewed at last decades. However, CrossLexica proposes too few morphological and lexical tags of words, so we decide to lean on vast and freely available OpenCorpora dictionary [2] with the detailed system of lexical and grammatical tags.

Thus, based on Open Corpora data, CrossMorphy's dictionary of word forms ($\sim$ 2 mln forms) was developed, taking the form of directed acyclic word graph (DAWG) [5], or acyclic finite state automaton. This effective data structure was proposed for storing word forms for highly flectional languages, and the same structure was applied in Pymorphy2 parser [7]. Therefore, CrossMorphy mainly inherits the system of lexical and morphological tags of OpenCorpora. Several rare grammatical cases were excluded (in particular, the second genitive), the traditional denotation of instrumental case was restored. Some modifications of OpenCorpora dictionary data were also made: several errors were fixed, analyses of single letters were excluded, lemma for personal pronouns was corrected, and links between adverbs and comparative adjectives were added, e.g., *дорого* – *дороже* (*expensive* - *more expensive*).

CrossMorphy performs both lemmatization and full morphological parsing, it is capable to find all interpretations of a given word forms. Stemming (that is splitting a given word form into pseudo flexion and pseudo stem and then outputting the latter) is incorporated into the processor as well (e.g., *носок, носками* – *нос*).

To estimate coverage of Russian lexicon, we have experimentally compared the rate of dictionary word forms processed by Mystem, Pymorphy, and CrossMorphy in vast text collection Librusec[7], the results are 97.2%, 96.5%, and 96.6% correspondingly, which evidences the sufficient coverage. In comparison with MyStem, CrossMorphy proposes (on average) more parsing variants for homonymous word forms, in particular, for word *улыбающийся* (*smiling*) it gives 4 variants whereas MyStem has 2 variants.

CrossMorphy can generate both paradigm or particular word forms for a given lemma or word form. More precise, if input set of tags for a given word

[7] http://lib.rus.ec/.

is incomplete, the processor produces all possible word forms. For example, for input word *шарами* and the given tag of grammatical number (single), CrossMorphy outputs the following forms: *шар, шара, шару, шаром, шаре.*

Functionality of CrossMorphy also includes no less useful auxiliary function of preliminary tokenization of texts and classifying tokens into words, numbers, punctuation, separators, and hieroglyphs. Each class of tokens has own additional tags, such as Cyrillic or Latin for words.

Handling of non-dictionary words and morphological disambiguation are also incorporated into CrossMorphy.

4 Processing of Non-dictionary and Hyphen Words

For handling new (non-dictionary) words and predicting their morphological features, CrossMorphy applies three general heuristic methods.

Prediction according word flexion (ending) is based on the well-known principle of analogy used in almost in all parsers for Russian with dictionary morphology. As a rule, the same word endings (1-5 last letters) correspond to the same syntactic class, so morphological tags (and lemma) of unknown word may be predicted by the final letters. The implementation of the principle varies in morphological processors, giving different numbers of resulting variants.

We propose the following prediction version with a reasonable number of answers. Statistics on all word endings (1 to 5 letters long) are collected for the dictionary, rare endings encountered less than 3 times are excluded, and the most frequent POS (part of speech) are determined for any particular ending. Then all the morphological interpretations for the endings with the determined POS are considered as the result.

Prediction according prefix is the second method, it involves cutting of possible prefix and then parsing the rest of the word form. Unlike AOT and Pymorphy2, we take into account only known prefixes (the built-in list of 207 prefixes compiled in open Russian Wiki-dictionary[8] is used) – this makes it possible to avoid errors in prediction of some words (e.g., for word *вейпер*).

For handling unknown hyphen words several rules are employed, accounting for cases with several hyphen (e.g., *фолк-панк-рок – folk-punk-rock*), words with digital and Latin letters constituents (*Рубин-5, S-выражение – Ruby-5, S-expression*), words with a single inclined constituent (*веб-инструктор – web instructor*), and with both inclined constituents (*человек-гора – man-mountain*).

It is important that three described methods are applied independently, and as a result, parsing of some words (such as *авторша*) is successive whereas in Pymorphy2 it fails.

To estimate processing of non-dictionary words, we used tagged corpus of NCRL (National Corpus of Russian Language)[9] with $\sim$ 1 mln word forms. Table 1 presents comparative data counted for three parsers: the total number of

[8] https://ru.wiktionary.org/wiki/.
[9] http://ruscorpora.ru/.

encountered non-dictionary tokens, percentages of tokens with correct resulted lemma, POS, and full tag parses accordingly (all the parsers performed morphological disabiguation). One can see that CrossMorphy wins in POS accuracy, exceeds Pymorphy2's scores for full tags, but loses to Mystem in lemma and full set of tags.

Table 1. Accuracy of parsing non-dictionary words

Processor	Total #	Lemma (%)	POS (%)	Full tags (%)
Mystem	11478	66.20	72.58	56.51
Pymorphy	15024	60.43	67.15	35.71
CrossMorphy	15030	59.68	85.60	41.13

5 Morphological Disambiguation

To now, the problem of POS classification for wordforms is well investigated for many languages, and disambiguation accuracy is near 98%. One of the first work for Russian [11] proposed the statistical method with the accuracy 97.42%, while MorphoRuEval-2010 evaluation [9] reported 94-95% obtained by rule-based methods.

For Russian, the challenging task is full morphological disambiguation, i.e. assignment of lemma and all meaningful grammatical tags (POS, case, gender, person, etc.) to word token. In the recent work [10] CRF (conditional random field) statistical method for morphological disambiguation was investigated and resulted in the accuracy up to 94, 95%.

We should note that all indicated evaluation rates are relative, since they depend on several factors including not only the applied method, but also the set of used morphological tags and the size of text corpora for training and testing. In all the works mentioned above, these factors differ, and at the same time all of them use reduced tag sets, as well as relatively small test corpora. In overall, evaluation and comparison of disambiguation method is complicated by the fact that there is neither standard of Russian morphology tagging, nor gold standard corpora for evaluation. Besides, the real problem is some incompatible tags used in morphological parsers.

In CrossMorphy two methods of statistical morphological disambiguation are implemented, contextual and non-contextual. The latter ranks parsing variants for a processed word form, according to frequency statistics of all parsing variants for the corresponding lemma. The statistics are gathered on the tagged corpus of NCRL (National Corpus of Russian Language)[10]. The possibility of a particular parsing variant is calculated according the formula

$$P(t|w) = \frac{Fr(w,t)}{Fr(w)}$$

[10] http://ruscorpora.ru/.

where w is a word form, t is a set of morphological tags, $Fr(w)$ and $Fr(w,t)$ are frequencies of w and w with its tags t in the corpus. Similar to the other parsers, CrossMorphy outputs the calculated probabilities of homonymous variants.

MyStem and Pymorphy2 use another methods to compute non-contextual scores of parsing variants, but it makes no sense to compare them, since they only rank the parsing variant, and the resulted ranks are similar.

In CrossMorphy, the non-contextual method is considered as auxiliary for contextual disambiguation. For the latter, CRF++ method is applied, so far as it presents results close to the state of the art for POS tagging for flexional languages [10,12]. Specifically, we use Limited-memory BFGS version of CRF. Since our classification task involves too many features (POS, lemma and all Russian obligatory grammatical tags), four CRF classifiers are sequentially applied.

First, the POS classifier is used, among accounted features are the token being processed and possible POS variants in the form of binary vector. The next CRF classifier is responsible for gender recognition, and accounted features are lemma, POS determined by the previous classifier, and also possible variants of gender (masculine, feminine, neutral). In similar way, subsequent CRF classifiers for number and case work. After all the classifying procedure, rare homonymous variants could still remain (for example, concerning animacy), in this case the non-contextual disambiguation is applied to choose an adequate variant.

An example of disambiguation for Russian word form *мыла* (*washed* or *soap*?) is shown in Fig. 1.

мыла	мыло	NOUN,inan,neut,nomn,plur
мыла	**мыло**	**NOUN,gent,inan,neut,sing**
мыла	мыло	NOUN,accs,inan,neut,plur
мыла	мыть	VERB,femn,impf,indc,past,sing,tran

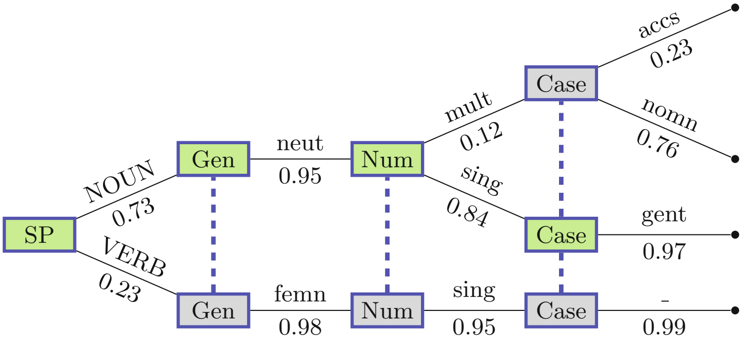

Fig. 1. Disambiguation of word form *мыла*

To estimate CrossMorphy's disambiguation model, we have performed two experiments with training the model on tagged data and its cross-validation (10%). We first used the corpus of NCRL with $\sim$ 1 mln word forms, and then Syntagrus and GICRL (General Internet Corpus of Russian Language) tagged data with more than 2 mln tokens. The latter corpora were obtained within MorphoRuEval-2017 competition [15] for comparing various disambiguation methods on the basis of large tagged Russian text corpora and the system of Universal Dependency tags[11]. In both experiments we had to convert morphological tags: NCRL tags into CrossMorphy's tags, and CrossMorphy's tags into UD tags (in the latter case we had to resolve some mismatches of the tag sets, concerning, in particular, restoring the difference between the comparative adjectives and comparative adverbs).

Accuracy rates achieved in the experiments after each step of the overall CRF classification procedure are presented in Table 2 (the last column also indicates full tag disambiguation). The rates of POS classification are better than in [9,11], but the final full tag rates (90-93%) are slightly less than in [12,15]. The best result achieved in the closed track of MorphoRuEval-2017 [15] is 93.39 (while the open track gives 97.11 due to training on large corpora and using neural net models). Thus, a more tricky procedure should be further developed for CrossMorphy. Our experiments also showed that another sequences of classifying gender, number and case do not improve final accuracy of diambiguation.

Table 2. Accuracy of sequentially applied CRF classifies (%)

Corpora / CRF	POS	Gender	Number	Case
NCRL	97.94	97.03	96.61	93.42
Syntagrus+GICRL	98.12	96.32	94.23	90.53

Accuracy of disambiguation showed by CrossMorphy indirectly evidences the quality of its dictionary and procedures for handling non-dictionary words. What is important for us, that CrossMorphy demonstrates about similar behavior on various testing corpora, containing news, fiction, and texts from internet social networks. Taking into account that conversion of morphological tags, which is needed for training, may lead to inevitable loss of significant information, we think that there is a reserve to improve overall quality of CrossMorphy parsing, in particular, disambiguation accuracy.

6 Morphemic Parsing

Additional functionality supported by CrossMorphy is automatic morphemic parsing (segmentation), that is dividing words into their morphs (root and affixes), e.g. *под - ковер - н - ый*, *в - брас - ыв - ать - ся*, *ин - дукц - и - я*. Clearly, it is reasonable to store morphemic structure of words in the dictionary, but there exist significant problems. First, there are no full dictionaries with morphemic segmentation of words (and many words of OpenCorpora and CrossMorphy are

[11] http://universaldependencies.org/u/overview/morphology.html.

absent in the known dictionaries). Second, there is no agreement between linguists about rules of morphemic segmentation for Russian words (apart from another languages with rich morphologies, there are many affixes of various types and behavior in Russian). And finally, the task cannot be automatically solved with high accuracy because of similarity of morphs.

Unlike the works [1,6,14], for automatic morphemic parsing we use supervised machine learning, specifically, CRF method. Morphemic segmentation is considered as classification of letters by recognizing their morphemic classes (Prefix, Root, Suffix, Ending). As accounted features we take the letter itself, is it a vowel, lengths of the word and its stem, POS of the word, its morphological tags, and also Harris's features [6] (local maximums of letter frequencies counted for various positions within words). An example of resulted classification is showed in Fig. 2.

$$\text{индукция} \rightarrow \begin{array}{|c|c|c|c|c|c|c|c|} \text{И} & \text{Н} & \text{Д} & \text{У} & \text{К} & \text{Ц} & \text{И} & \text{Я} \\ \hline \text{P} & \text{P} & \text{R} & \text{R} & \text{R} & \text{R} & \text{S} & \text{E} \end{array}$$

Fig. 2. Morphemic parsing of word *индукция*

Morphemic classification models were obtained by training on two tagged data taken correspondingly from CrossLexica system [3] (23426 parsed words) and Russian Wiki dictionary[12] (94485 parsed words). We could not combine these two data sets, since many words presented in both sets have different morphemic segmentation. Thus, we separately built two classifiers, and their accuracy was evaluated both on fragments of the own and alien corpora – the results are presented in Table 3.

Table 3. Accuracy of morphemic classifiers

Data		Precision				
Training	Validation	Whole word	Prefix	Root	Suffix	Ending
CrossLexica	CrossLexica	**74.2**	**86.13**	**75.10**	**77.13**	97.95
CrossLexica	Wiki	35.91	66.14	56.75	38.57	57.35
Wiki	CrossLexica	46.64	78.11	66.38	50.43	70.20
Wiki	Wiki	65.87	70.92	65.47	71.84	**98.31**

One can see that cross validation on the alien corpus gives a significant loss of accuracy. CrossLexica's data yields the best scores for the most morphemic classes (unlike Wiki dictionary, the data were created by a single human expert, so are more homogeneous). For this reason, corresponding classifier was incorporated into our processor. The accuracy of the incorporated classifier (74,2%) is better than the best result 70% obtained for Turkish in [14].

[12] https://ru.wiktionary.org/wiki/.

7 Conclusions and Future Work

In this paper we have compared functional properties of several popular freely available morphological processors for Russian texts, thus explaining the reasons to develop yet another processor for Russian. The developed open source morphologic processor CrossMorphy has the distinguishing combination of properties that meets our requirements. Evaluation of its functionality has showed sufficiently accurate processing of Russian texts. Across main functions, our processor is competitive with known freely available parsers, and at the same time its functionality is extended by morphemic segmentation.

On the way towards a high-quality morphological processor, the further improvements of CrossMorphy are needed:

– more exhaustive testing and providing convenient documentation;
– providing tools for connecting user dictionaries;
– incorporating additional rules for classifying non-dictionary words based on information about thier morphemic segmentation;
– elaborating a more accurate model of morphological disambiguation;
– providing linguistically correct convertors between different systems of Russian morphological tags; creation of suitable universal system of morphological tags for Russian is a more challenging task.

Acknowledgements. We would like to thank the reviewers of our paper for their helpful comments.

References

1. Bernhard, D.: Simple morpheme labelling in unsupervised morpheme analysis. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 873–880. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-85760-0_112
2. 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 International Conference "Dialogue" (2011)
3. Bolshakov, I.A.: CrossLexica, the universe of links between Russian words. In: Busyness Informatica, No. 3 (2013)
4. Bolshakova, E., Efremova, N., Noskov, A.: LSPL-patterns as a tool for information extraction from natural language texts. In: New Trends in Classification and Data Mining. Markov, K., et al. (eds.) ITHEA, Sofia, pp. 110–118 (2010)
5. Daciuk, J., Mihov, S., Watson, B., Watson, R.: Incremental construction of minimal acyclic finite state automata. Comput. Linguist. **26**(1), 3–16 (2000)
6. Harris, Z.S.: Morpheme boundaries within words: report on a computer test. In: Transformations and Discourse Analysis Papers, vol. 73, pp. 68–77 (1970)
7. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) AIST 2015. CCIS, vol. 542, pp. 320–332. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_31

8. Kuzmenko, E.: Morphological analysis for Russian: integration and comparison of taggers. In: Ignatov, D.I., Khachay, M.Y., Labunets, V.G., Loukachevitch, N., Nikolenko, S.I., Panchenko, A., Savchenko, A.V., Vorontsov, K. (eds.) AIST 2016. CCIS, vol. 661, pp. 162–171. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-52920-2_16
9. Ljashevskaya, O., Astaf'eva, I., Bonch-Osmolovskaja, A., Garejshina, A., Grishina, J., D'jachkov, V., Ionov, M., Koroleva, A., Kudrinskij, M., Litjagina, A., Luchina, E., Sidorova, E., Toldova, S.: NLP evaluation: Russian morphological parsers. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", pp. 318–326 (2010)
10. Muzychka, S.A., Romanenko, A.A., Piontkovskaja, I.I.: Conditional random field for morphological disambiguation in Russian. In.: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", pp. 456–465 (2014)
11. Segalovich, I.A.: Fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA, pp. 273–280 (2003)
12. Shen, Q., Clothiaux, D., Tagtow, E., Littell, P., Dyer, C.: The role of context in neural morphological disambiguation. In: COLING 2016, 26th International Conference on Computational Linguistics. Proceedings of the Conference: Technical Papers, Osaka, Japan. ACL, pp. 181–191 (2016)
13. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44–49 (1994)
14. Smit, P., Virpioja, S., Gronroos, S., Kurimo, M.: Morfessor 2.0: toolkit for statistical morphological segmentation. In: Proceedings of the Demonstrations at the Conference of the European Chapter of the ACL, pp. 21–24 (2014)
15. Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, V., Alexeeva, S., Droganova, K., Fenogenova, A.: MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In: Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialogue 2017, Moscow (2017)
16. Zaliznjak, A.A.: Grammatical Dictionary of Russian: Inflection. Russkij Jazyk Publisher, Moscow (1977)