A cross-genre morphological tagging and lemmatization of the Russian poetry: distinctive test sets and evaluation*

Aleksey Starchenko[1] and Olga Lyashevskaya[1,2][0000−0001−8374−423X]

[1] National Research University Higher School of Economics
amstarchenko@edu.hse.ru,olyashevskaya@hse.ru
https://www.hse.ru/org/persons/208526218, https://www.hse.ru/en/staff/olesar
[2] Vinogradov Institute of the Russian Language RAS, Moscow, Russia

Abstract. The poetic texts pose a challenge to full morphological tagging and lemmatization since the authors seek to extend the vocabulary, employ morphologically and semantically deficient forms, go beyond standard syntactic templates, use non-projective constructions and non-standard word order, among other techniques of the creative language game. In this paper we evaluate a number of probabilistic taggers based on decision trees, CRF and neural network algorithms as well as a state-of-the-art dictionary-based tagger. The taggers were trained on prosaic texts and tested on three poetic samples of different complexity. Firstly, we suggest a method to compile the gold standard datasets for the Russian poetry. Secondly, we focus on the taggers' performance in the identification of the part of speech tags and lemmas. We reveal what kind of POS classes, paradigm classes and syntactic patterns mostly affect the quality of processing.

Keywords: NLP evaluation · Full morphology tagging · POS-tagging · Lemmatization · Russian language · Russian poetry.

1 Introduction

Poetic texts are usually processed with the help of the standard NLP tools that have been originally developed for and tested on prose. Ringger et al. [21] report a 8% drop in tagging accuracy on BNC poetry date while using a tagger trained on prosaic data. The Corpus of Russian Poetry (a part of the Russian National Corpus, RNC) is currently processed using Mystem [22], in which a statistical module is trained on web texts and prosaic RNC texts.

However, the distributional probabilities are different in the prosaic and poetic varieties. There are more nouns (30.3%) and adjectives (13.1%) and less

* The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2018 — 2019 (grant № 18-05-0047) and within the framework of the Russian Academic Excellence Project «5-100».

verbs (14.9%) in the RNC Poetry Corpus than in the RNC Standard (prose) corpus (28.5%, 12.8%, 17.0%, respectively). The dissimilarities in lexical probabilities are even more noticeable, as the authors of poetry strive to enrich the lexicon, pick up rare gourmet rhymes, play with lacunae in grammar, be innovative in word derivation, etc., that is, to be 'creative' in the broadest sense of the term. Besides that, the rhythmic structure of poetry also affects syntactic patterns, word order, and the choice of lexical units and collocations. All these factors may challenge the cross-genre tagging and bias the prediction of the POS tags, grammatical features, and lemmas: three important constituents of the full morphological tagging.

Yet, developing a system designed specifically for poetry carries its own risks. Enhanced lexicon and more variable features such as the character and word ngrams are associated with the sparsity of language models, and using a (presumably) smaller genre-specific annotated corpus to train the new tagger is not always the best remedy in such cases. The aim of this papers is twofold. On the one hand, we discuss possible ways to compile poetic datasets as material for tagger evaluation (Section 2) and describe the taggers we used (Section 3). On the other hand, we report a preliminary experiment on the evaluation of the standard well proven tools developed for prose as a baseline for future comparison of existing and new genre-specific models (Sections 4-6).

2 Distinctive test sets

The accuracy of full morphological tagging applied to modern languages is as high as 92-95% [24]. The best accuracy of POS-tagging reported for languages like English and German is close to 97%-98% [10]. With such high scores in assessment, the difference in the taggers' performance cannot be seen clearly. The idea behind the use of distinctiveness datasets (e.g. Rare Words dataset, [13]) is to provide the basis for more conservative, lower scores, taking into account only the most challenging data.

Since the low probability of a word or a word sequence is known as a bottleneck in text processing, three data sets were created: the first (Dataset A) is compiled so that it has a large percentage of out-of-vocabulary words, the second (Dataset B) includes complicated, in particular non-projective, syntactic constructions, and the third (Dataset C) contains a random poetic text as a 'general' sample.

Dataset A (750 words) is a sample drawn from the RNC Corpus of Russian Poetry [8]. It contains sentences with the high proportion of irregular forms and out-of-vocabulary words (OOV). Note that the notion of OOV words is different in dictionary-based and probabilistic tagging. If the words are not attested in the dictionary, they cannot be labeled by the dictionary-based tagger, and if the words have not been seen in the training set, they are harder to be correctly labeled by the probabilistic tagger than words which have been seen in the training data. Thus, the inventory of OOV words depends on a particular dictionary used by the tagger (cf. the grammatical dictionaries of Mystem and OpenCorpora)

and on a particular training corpus and its size (cf. the RNC Standard, 6 MW, and SynTagRus, 1 MW). Still, we assume that the 'rare' words are unlikely to be present both in a dictionary and in a training collection and use the term OOV for both.

In order to compile Dataset A, we processed the word list of the Corpus of Russian Poetry by Mystem 3.1, which has an option to label the OOV words. Among the words which have been obtained, the following types are characteristic of the poetry texts:

- words with orthographic distortion and variation: *što* 'that' (cf. *čto*), *šopot* 'whisper' (cf. *šepot*), *ra-* || *zjaščee* 'smashing' (the word is divided by the line boundary);
- syllable dilation and contraction: *Zeves* 'Zeus' (cf. *Zevs*), *poln* 'full' (cf. *polon*);
- non-native names: *Io, Eol, Sal'vaterre*;
- archaic and archaic-like words: *drugi* 'friends', *oblak* 'cloud' (masculine);
- (quasi-)loan words: *mus'je* 'monsieur';
- non-standard grammatical forms: *mysliju* 'thought' (Instr. sing. noun, cf. *mysl'ju*), *uš* 'ears' (Gen. plural noun, cf. *ušej*), *ostavja* 'leaving' (Perfective gerundive, cf. *ostaviv*), *okazalasja* 'occur' (reflexive Past feminine verb, cf. *okazalas'*), *mjaučat* 'mew' (Present 3rd person plural verb, cf. *mjaukajut*).

As a next step, we inspected and ranked the OOV words from easy to difficult in terms of (a) POS identification, (b) inflectional form identification, and (c) lemma identification. For example, the short (2-3 character) words are difficult in all three aspects whereas words such as oblak are assumed to be classified correctly in terms of POS but misclassified in terms of gender labeling and lemmatization. Finally, a sample of sentences which contain at least two 'difficult' OOV words were retrieved using the frequency database of the Corpus of Russian Poetry [17]. As an instance, there are two non-standard grammatical forms in (1), and the fact that they are placed side-by-side, makes the sentence more difficult to be processed correctly.

(1) [Lanitoju] [pryžavšisja] k perstu, || V ten', nedostupnuju tumanam i vetram.
lit. 'With a cheek pressing to the finger, || To the shadow, inaccessible to mists and winds'

Dataset B (850 words) is a sample of syntactically complex and non-standard sentences. We use several syntactic templates that we consider to be typical of the Russian poetry to retrieve the sentences for Dataset B:

- adjectives in the attributive position placed after their head, cf. *kisti.Noun.Gen čužoj.*Adj 'brush of someone else' in (2);
- nouns in the genitive construction where the genitive form is placed before its head, cf. *kisti.Gen kiparisy* 'cypress from one's brush' in (2);
- pre-position of the direct and indirect object, cf. (3); post-position of the subject with regard to the verb;

– verb phrases, noun phrases with one or more clause or parenthetic construction inserted inside.

(2) Kisti.Gen čužoj.Adj kiparisy i rozy. || Prosalili belyj kak vosk amvon.
lit. 'From the brush of someone else, cypress and roses || Saturated a wax white ambo.'
(3) Sveču.Acc sverkan'ju.Dat ljustr predpočitaem.
lit. '(It is) candle (that) we prefer to sparkling chandeliers.'

Dataset C (1750 words) is an excerpt drawn from the open source manually annotated UD_Russian-Taiga treebank [6]. Among other genres, this corpus includes folk poetry published on social media. Dataset C was meant to represent the 'average' level of complexity of poetic texts, although the length of the sentences appeared to be longer in Dataset C than in the Corpus of Russian Poetry in general.

3 Taggers

To date, a number of taggers have been tested on Russian (prosaic) data, both language-specific tools (Mystem, AOT [23], PyMorphy [11], NLTK4RUSSIAN [19], UDAR [20]), and general models trained on Russian data (TreeTagger [26], TnT [3], MarMoT/Lemming [18], UDpipe [25], various versions of BiLSTM taggers [1]). Evaluation of taggers on Russian prose data has been carried out within the framework of RU-EVAL 2010, MorphoRuEval 2017, SIGMORPHON 2016, CONLL 2018 shared tasks [14,24,4,29], see also evaluation experiments reported by [12,5]. In our study, we applied the following taggers to the Russian poetic material:

– Mystem-RuSyntax, an implementation of the Mystem model currently used in the annotation of the Main RNC corpus (consisting of prose texts), with the addition of context rules for POS disambiguation [7];
– Mystem 3.1, a standard implementation of Mystem provided by Mystem+ [5];
– TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, [26]), a tagger using automatic derivation of decision trees
– Hunpos (https://code.google.com/archive/p/hunpos/, [9]), a reimplementation of TnT tagger [3] using a trigram based HMM model;
– MarMoT (http://cistern.cis.lmu.de/marmot/, [18], a higher-order conditional random field (CRF) tagger;
– Lemming (http://cistern.cis.lmu.de/lemming/, [18], a modular log-linear tool based on the principles of a deterministic pre-extraction of edit trees, which jointly models lemmatization and tagging, an add-on to MarMoT;
– UDpipe (http://ufal.mff.cuni.cz/udpipe/, [25]), a rich feature averaged perceptron tagger, a baseline for CONLL 2018 shared task;
– Stanford POS tagger (http://nlp.stanford.edu/software/, [27]), a maximum entropy POS tagger (a bidirectional option) provided as a part of the Stanford CoreNLP Natural Language Processing Toolkit.

We use two versions of Mystem as a dictionary-based, rule-based baseline. The hypothesis builder for the OOV words in MyStem was trained on a big Yandex web collection [30], and the grammatical dictionary used is an extended version of [28]. Mystem-RuSyntax uses a model adopted to the RNC annotation guidelines [16]: unlike Mystem 3.1, it assigns separate lemmas to the perfective and the imperfective verbs and makes use of the stop list of annotations never attested in the RNC. The other taggers are probabilistic and differ in the size and type of the corpus on which the model was trained and the type of output they provide. TreeTagger, Hunpos, and MarMoT were trained on the 6MW corpus of Modern Russian prose (RNC Standard) in the framework of the Mystem+ project [5], therefore comparing their results achieved on the testing sets allows one to compare exactly the performance of the models, and not the quality of the training sample. When compared with Mystem, it should not be forgotten that the results of the comparison may change when the training sample is changed. UDpipe was trained on a 1 MW SynTagRus collection converted into UD format [6]. The Lemming model was trained by us on a 0.4 MW subcorpus of OpenCorpora prosaic texts [2]. The taggers learn from the following annotation types and therefore provide them in the output:

- Stanford POS tagger - only POS tags;
- TreeTagger, Hunpos, MarMoT - POS, grammatical features;
- Lemming - lemmas (adds lemmas to the output of MarMoT);
- UDpipe - POS, grammatical features, lemmas.

Thus, we can compare POS tagging across all models, lemmas - in Mystem, Lemming, and UDpipe, and grammatical features - across all models except Stanford and Lemming.

4 Experimental setup

Gold labels. All datasets were labeled with POS tags, grammatical feature tags, and lemmas. Each dataset was corrected manually by one annotator, and a small number of errors were also corrected post-hoc during evaluation stage. **Predicted labels**. The processed data were converted into the Universal Dependencies v. 2.0 standard, see Fig. 1. We followed the conversion rules of MorphoRuEval 2017 [24,15], with some adjustments. Animacy and aspect are let in evaluation, and the participle and gerundive forms are treated as forms of the verb. The predicted data were matched token by token to the gold collection. Punctuation marks, which are not returned by some taggers, and a number of frequent words known to be systematically labeled differently in different frameworks (e. g. *kotoryj* 'which') were marked off evaluation.

It should be noted that Mystem 3.1 does not disambiguate among possible grammatical annotations available for the identified lemma and POS and provides them all in alphabetical order. Technically, we assigned the first grammatical annotation to the token in evaluation. As a result, we cannot compare the accuracy of this tagger with the accuracy of the others, but nevertheless we

```
1 елей   ель     NOUN _ Animacy=Inan|Case=Gen|Gender=Fem|... 3 obl    _ _
2 ночь   ночь    NOUN _ Animacy=Inan|Case=Nom|Gender=Fem... 3 nsubj  _ _
3 стоит  стоять  VERB _ Aspect=Imp|Mood=Ind|Number=Sing|...  0 root   _ _
4 густая густой  ADJ  _ Case=Nom|Degree=Pos|Gender=Fem|Nu... 3 obl    _ _
```

Fig. 1. Annotations converted into UD-CONLLU format. Glossing of the clause: *елей* 'fur-NOUN.Gen.pl' *ночь* 'night-NOUN.Nom.sg' *стоит* 'stand-VERB.present.3sg' *густая* 'thick-ADJ.Nom.f.sg' 'The night of furs is thick'.

can roughly compare the results of Mystem 3.1 applied to different Datasets (A, B, C).

We hypothesize that when processing Dataset B, taggers using probabilistic learning should show less stable results compared to their performance on Dataset A and C, since these taggers rely on word co-occurrence and syntax. The dictionary-based tagger Mystem should show a higher percentage of errors while parsing Dataset A, which contains a large number of non-vocabulary words. In what follows, we will analyse the results of the experiment and check if our assumptions hold.

5 POS-tagging

Table 1 shows the accuracy of POS-tagging when applied to Datasets A, B, C. Here and below, the accuracy metrics are calculated in %, with punctuation not taken into account. To compare with, the last row reports by default the results obtained on the prosaic texts in [5]. Overall, the accuracy of the best systems on the poetic texts ranges from 91.9% to 95.2% for the POS tags and from 82.4% to 92.6% for the feature tags.

Surprisingly, none of the taggers is an absolute winner: Hunpos is the best on Dataset A (OOV words), Stanford – on Dataset B (complicated syntax), POS tags, and MarMoT – on Dataset C (general). Even more surprisingly, TreeTagger, which performed best on the prosaic texts, occurs to be the least accurate on the poetic texts. The accuracy of the identification of grammatical labels does not exceed 86% (more than 10% less than the POS accuracy in winning systems) and, since Stanford does not provide this type of data, MarMoT wins the race on both Datasets B and C.

If we compare the results across datasets, we see that our assumption that the text with a complex syntactic structure is problematic for machine-based taggers has been confirmed: the scores obtained on Dataset B are certainly lower than the scores obtained on the general Dataset C. They are also lower than scores obtained on Dataset A (in both POS and feature identification tasks, the only exception is MarMoT on feature tagging).

[1] Values not reported in [5].

[2] Models evaluated on the UD 2.3 test dataset of Russian-Syntagrus (without punctuation, 96k words).

Table 1. POS and feature tagging, accuracy in %.

	MarMoT		Hunpos		TreeTagger		Stanford	UDpipe		Mystem 3.1	
Dataset	**POS**	**Feat**	**POS**	**Feat**	**POS**	**Feat**	**POS**	**POS**	**Feat**	**POS**	**Feat**
A(oov)	93.1	78.6	**94.3**	**82.4**	87.4	72.2	94.1	88.9	74.3	91.7	67.7
B(order)	87.8	**82.6**	87.8	79.9	82.8	70.6	**91.9**	91.5	78.4	88.5	71.4
C(web)	**95.2**	**85.5**	94.3	83.3	90.9	77.1	93.9	92.5	78.1	91.3	65.8
Prose	96.0	—[1]	96.4	89.3	**96.9**	**92.6**	95.8[2]	98.2[2]	92.5[2]	96.4	—[1]

The other hypothesis, that the accuracy will noticeably decrease with the increase in the number of non-vocabulary words, is not confirmed (compare the scores for Datasets C and A). Unlike MarMoT and TreeTagger, Hunpos and Stanford demonstrate approximately the same or slightly higher results on Dataset A. Yet, the accuracy of Marmot and TreeTagger's features decreases considerably as we move from Dataset C to Dataset A, as was expected.

Finally, Mystem, a dictionary-based tagger, shows generally uncommon results: it processes Dataset A with higher accuracy than Dataset C, even though the ratio of OOV words is higher in Dataset A. We can suggest that the tagging quality is affected by other factors which were not taken into account when we constructed the test sets. For example, there is an uneven proportion of nouns in Datasets A, B, and C: 34.2%, 30.2%, and 65.3%, respectively. As nouns usually show a greater tendency toward the grammatical ambiguity of forms, the method to get rid of homonymy we chose can lead to a greater number of errors in the case of words with ambiguous forms.

Comparing the accuracy of processing poetry vs. prose, we see that the scores are expectedly higher in the latter case, although the difference in POS tagging is not particularly noticeable. Interestingly, TreeTagger, which showed the best results in the tagging of prose, fails on poetry, demonstrating a greater bias to the type of text than the other taggers.

Table 2 summarizes the correspondence of the gold POS tags (lines) and those predicted by MarMoT/Lemming trained on a smaller 0.4 MW corpus (columns), see Section 6. Since its accuracy is lower compared to the accuracy of taggers described above, we can get enough error data to analyze them in more detail.

The analysis shows that words that constitute small, closed classes — i.e. conjunctions and prepositions — are most accurately identified. On the opposite side, the accuracy of processing for adverbs is low, almost close to chance. Such a low accuracy can be explained by the relative syntactic freedom of adverbs: many adverbs can appear anywhere in the sentence. In addition, a number of errors are caused by the mismatches between annotation practice in our gold data and the corpus on which Lemming was trained. Let us consider the case of predicatives. This group includes words of different types: adjectival predicates ending in *-o/* textit-e (*khorosho* 'good', *blizko* 'close to'), predicative nouns (*pora* 'it's time', *len'* 'lazy'), modal predicates (*dolžen* 'have to', *možno* 'possible'), the negative word *net* 'there is no'. If such a category is not present in the corpus

Table 2. Confusion matrix for MarMoT: POS tags (based on Dataset A). In each cell, a number of occurrences is given; below them, the first percentage shows the ratio of the gold labels classified by the tagger as a particular POS; the second percentage is a relative frequency of the class in all cases predicted as a particular POS.

POS	**adj**	**adp**	**adv**	**conj**	**det**	**intj**	**noun**	**num**	**part**	**pron**	**verb**	**x**	**Total**
ADJ	67 85% 87%		2 3% 9%		1 1% 5%		5 6% 2%				2 3% 2%	2 3% 6%	79 100% 11%
ADP		86 97% 99%					1 1% 0%					2 2% 6%	89 100% 12%
ADV	2 7% 3%		16 53% 70%	3 10% 5%			4 13% 2%	1 3% 25%				4 13% 11%	30 100% 4%
CONJ				51 94% 91%					1 2% 6%	1 2% 2%		1 2% 3%	54 100% 7%
DET					20 83% 95%		1 4% 0%			3 13% 7%			24 100% 3%
INTJ						3 75% 100%						1 25% 3%	4 100% 1%
NOUN	4 2% 5%	1 0% 1%	2 1% 9%				230 91% 89%				2 1% 2%	15 6% 43%	254 100% 34%
NUM								3 75% 75%				1 25% 3%	4 100% 1%
PART							2 11% 1%		16 84% 89%		1 5% 1%		19 100% 3%
PRON	1 2% 1%			2 4% 4%			3 6% 1%		1 2% 6%	42 86% 91%			49 100% 7%
VERB	3 2% 4%		3 2% 13%				12 8% 5%				116 81% 95%	9 6% 26%	143 100% 19%
X							1 50% 0%				1 50% 1%		2 100% 0%
Total	77 10% 100%	87 12% 100%	23 3% 100%	56 7% 100%	21 3% 100%	3 0% 100%	259 34% 100%	4 1% 100%	18 2% 100%	46 6% 100%	122 16% 100%	35 5% 100%	751 100% 100%

tag set, the predicative words are distributed among other POS classes: adverbs, nouns, verbs. When we compared the two sets of tags, a technical decision was made (according to the practice adopted in the UD corpus from which Dataset C was taken) to label as adverbs all predicatives but the word *net*, which is considered a verb. As a result, a few predicative nouns are not labeled correctly.

The identification of some parts of speech is expectedly asymmetric. Thus, on the one hand, 95% of all verbs in the dataset are correctly identified by Lemming, which is a good result. On the other hand, Lemming also assigns the label "verb" to a number of words belonging to other parts of speech, so that its accuracy is not very high - only 80%.

6 Lemmatization

This section focuses on lemmatization. We analyze the accuracy of lemma labeling and consider a number of challenging cases. Table 3 presents the accuracy of lemmatization predicted by two lemmatizers: Lemming (probabilistic) and Mystem (hybrid, dictionary-based). Since the size of the corpus on which Lemming was trained is small (0.4 MW), the accuracy of POS and feature labels predicted by Lemming is lower than that predicted by the taggers presented above. In order to display this difference, Table 3 also summarizes data on the accuracy of the POS tagging.

Table 3. Lemmatization, accuracy in %.

	Lemming/MarMoT		**Mystem**	
Dataset	**Lemma**	**POS**	**Lemma**	**POS**
A	85.0	87.7	87.7	91.7
B	87.7	87.3	86.4	88.5
C	87.9	88.4	91.4	91.3

It can be seen that the quality of lemmatization by Lemming и Mystem varies weakly depending the dataset tested; we can only point out that for the dictionary-based lemmatizer, both the dataset with complex syntactic constructions (B) and the dataset with the out-of-vocabulary words (A) are problematic. At the same time, although Lemming was trained on a small data set, its accuracy is close to the accuracy of Mystem.

As for difficult cases, there is a number of OOV words with non-standard endings (*nest'* 'carry', *prinest'* 'bring', *unest'* 'carry out', instead of *nesti*, *prinesti*, *unesti*). Since no rules implemented in Mystem to support orthographic variation these infinitives are incorrectly tagged as predicatives because of their similarity with the word *nest'* 'there is no'.

Not surprisingly, Lemming often makes mistakes when applied to the cases in which the part of speech tags were incorrectly identified. In particular, when

the part-tag tag cannot be chosen (that is, the tag X is selected), lemmatization is not performed: the word form is chosen as the lemma.

One more frequent type of errors is a wrong choice of the ending in the cases in which there are two words in the language with the overlapping paradigm, cf. *bank* 'bank' and *banka* 'jar'. This error is known as misclassification of the type of declension, and usually the nouns of different grammatical gender are mixed. Thus, the lemma *kos* is assigned instead of *kosa* 'braid', *kail* 'Kyle' instead of *kailo* 'pick', *platka* 'patch' instead of *platok* 'handkerchief'. This error sometimes occurs even if the morphological gender is correctly defined. The choice between two possible allomorphs can also be incorrect, cf. *khudyj* 'thin' instead of *khudoj* 'thin', *dysat'* instead of *dyshat'* 'breathing'.

7 Conclusions

We compared taggers of different types in a full morphological annotation task for poetic texts. As expected, poetry in general turns out to be difficult for processing by taggers which were designed and trained on prose, the nonstandard syntactic patterns being the most challenging. The accuracy of POS tags ranges from 91.9% to 95.2%. The drop in accuracy is more significant in the feature tagging (82.4%-85.5%), which can be explained by the complexity of the classification task itself and by some conventions of data evaluation which we followed. The case study shows that adverbs are most difficult to pars and the label verb is frequently assigned to the words of other parts of speech. As for lemmatization, it turned out that its accuracy weakly depends on the type of text and - for the selected taggers - on the type of tagger. There is no doubt that these results have to be verified in further experiments with models trained on larger data.

Overall, the distinctive test sets help to identify the gap for improvement and make it linguistically interpretable. However, since our collections are small, they do not provide enough statistics to make definitive conclusions on taggers' performance. The complexity of the structures in the poetic texts and the small amount of test data may explain the mixed results achieved with the distinctive datasets. We did not control for syntactic complexity while mining the dataset for OOV words and vice versa. None of the parameters was controlled while randomly sampling Dataset C. A more promising approach would be to annotate the word entries according to multiple parameters within one large test collection. After that, a set of additional individual metrics will be obtained by choosing a subset of the test data such as words positioned in nonstandard word order, words which have counterparts with overlapping paradigms, and other parameters of the test data profiling.

8 Acknowledgements

We would like to thank Lev Kazakevich who assisted with data collection for this study, and three anonymous reviewers for insightful comments on the first version of this paper.

References

1. Anastasiev, D. G., Gusev, I. O., Indenbom, E. M.: Improving part-of-speech tagging via multi-task learning and character-level word representations. In: Dialogue'2018, Int. Conf. on Computational Linguistics and Intellectual Technologies. Moscow (2018)
2. Bocharov, V. V., Alexeeva, S. V., Granovsky, D. V., Protopopova, E. V., Stepanova, M. E., Surikov, A. V.: Crowdsourcing morphological annotation. In: Proceedings of Dialogue 2013. Moscow (2013)
3. Brants, T.: TnT – a statistical part-of-speech tagger. In: Proceedings of the 6th Applied Natural Language Processing Conference, pp. 224–231. Seattle (2000)
4. Cotterell, R., Kirov, Ch., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M.: The SIGMORPHON 2016 shared task – morphological reinflection. In: Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 10–22. ACL (2016)
5. Dereza, O. V., Kayutenko, D. A., Fenogenova, A. S.: Automatic morphological analysis for Russian: A comparative study. In: Proceedings of Dialogue 2016. Moscow (2016)
6. Droganova, K., Lyashevskaya, O., Zeman, D.: Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. In: Proceedings of TLT 2018. Oslo (2018).
7. Droganova, K. A., Medyankin, N. S.: NLP pipeline for Russian: an easy-to-use web application for morphological and syntactic annotation. In: Proceedings of Dialogue 2016. Moscow (2016).
8. Grishina, E., Korchagin, K., Plungian, V., Sichinava, D.: Poeticheskij korpus v ramkakh Nasional'nogo korpusa russkogo yazyka: obshchaya struktura i perspektivy ispol'zovaniya [The corpus of poetry within the Russian National Corpus: a general outline and perspectives of use]. In: Natsional'nyj korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The Russian National Corpus: 2006–2008. New results and prospects]. St. Petersburg: Nestor–Istoriya Publ. (2009)
9. Halácsy, P., Kornai, A., Oravecz, Cs.: Hunpos: An open source trigram tagger. In: ACL'07, Interactive Poster and Demonstration Sessions, pp. 209–212. Stroudsburg, PA (2007)
10. Horsmann, T., Erbs, N., Zesch, T.: Fast or accurate? — a comparative evaluation of pos tagging models. In: GSCL 2015, pp. 22–30. Duisburg-Essen, Germany (2015)
11. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: AIST–2015, pp. 320–332. Springer, Cham (2015)
12. Kuzmenko, E.: Morphological analysis for Russian: integration and comparison of taggers. In: AIST–2016, pp. 162–171. Springer, Cham (2016)
13. Luong, T., Socher, R., Manning, C.: Better word representations with recursive neural networks for morphology. In: CONLL 2013, pp. 104–113. Sofia, Bulgaria (2013)
14. Lyashevskaya, O., Astaf'eva, I., Bonch-Osmolovskaya, A., Gareyshina, A., Grishina, Ju., D'yachkov, V., Ionov, M., Koroleva, A., Kudrinskiy, M., Lityagina, A., Luchina, E., Sidorova, E., Toldova, S., Savchuk, S., Koval', S.: Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskie parsery russkogo jazyka. In: Proceedings of Dialogue 2010, pp. 318-–326. Moscow (2010)
15. Lyashevskaya, O., Bocharov, V., Sorokin, A., Shavrina, T., Granovsky, D., Alexeeva, S.: Text collections for evaluation of Russian morphological taggers. Jazykovedný casopis 68(2), 258–267 (2017)

16. Lyashevskaya, O., Plunguan, V., Sichinava, D.: O morpfologicheskom standarte Natsional'nogo korpusa russkogo jazyka [On the morphological standard of the Russian National Corpus]. In: Natsional'nyj korpus russkogo jazyka 2003–2005, pp. 111–135. Moscow (2005)
17. Lyashevskaya, O., Litvintseva, K., Vlasova, E., Sechina, E.: A Data Analysis Tool for the Corpus of Russian Poetry. WP BRP Linguistics, Moscow, HSE University (2018)
18. Müller, T., Cotterell, R., Fraser, A., Schütze, H.: Joint lemmatization and morphological tagging with lemming. In: EMNLP-2015, pp. 2268–2274. Lisbon, Portugal (2015)
19. Panicheva, P., Protopopova, E., Mitrofanova, O., Mirzagitova, A.: Razrabotka lingvisticheskogo kompleksa dlja morfologicheskogo analiza russkojazychnykh korpusov tekstov na osnove PyMorphy i NLTK [Development of an NLP toolkit for morphological analysis of Russian text corpora based on PyMorphy and NLTK]. In: Int. Conf. CORPORA–2015, Saint-Petersburg (2015)
20. Reynolds, R.: Russian natural language processing for computer-assisted language learning: Capturing the benefits of deep morphological analysis in real-life applications. PhD diss. Tromsø, University of Tromsø (2016)
21. Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., Seppi, K., Lonsdale, D.: Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. In: Linguistic Annotation Workshop. Prague (2007)
22. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA–2003. Las Vegas (2003)
23. Sokirko, A.: Morphologicheskie moduli na sajte www.aot.ru [Morphological tools on the website www.aot.ru]. In: Dialog'04, Int. Conf. on Computational Linguistics and Intellectual Technologies. Moscow (2004)
24. Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, B., Alexeeva, S., Droganova, K., Granovsky, D.: MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In: Proceedings of Dialog-2017. Moscow (2017)
25. Straka, M., Hajič, J., Straková, J.: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: LREC–2016. Portorož, Slovenia (2016)
26. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. Manchester (1994)
27. Toutanova, K., Klein, D., Manning, Ch. D., Singer, Y.:. Feature-rich part-of-speech tagging with a cyclic dependency network. In: NAACL, vol. 1, pp. 173–180 (2003)
28. Zaliznyak, A. A.: Grammaticheskij slovar' russkogo jazyka [The grammatical dictionary of Russian]. Moscow (2003)
29. Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–21. ACL (2018)
30. Zobnin, A. I., Nosyrev, G. V.: Morfologicheskij analizator Mystem 3.0 [A morphological analyzer Mystem 3.0]. Trudy Instituta russkogo jazyka im. V. V. Vinogradova [Works of Vinogradov Institute of the Russian Language RAS] (6), 300–310 (2015)