NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Alexey Starchenko, Lev Kazakevich,
Olga Lyashevskaya

APPLYING PROBABILISTIC TAGGING TO RUSSIAN POETRY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 76/LNG/2018

SERIES: *LINGUISTICS*

Alexey Starchenko[1], Lev Kazakevich[2], Olga Lyashevskaya[3]

APPLYING PROBABILISTIC TAGGING TO RUSSIAN POETRY[4]

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 one 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 discuss the 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. These two annotation layers are key to compiling the corpus-based dictionaries, which we consider a long-term goal of our project.

JEL Classification: Z.
Keywords: natural language processing, full morphology tagging, NLP evaluation, Russian language, Russian poetry

1. Introduction

The poetic texts are usually processed with the help of the standard NLP tools which have been originally developed for and tested on prose. The Corpus of the Russian Poetry (a part of the Russian National Corpus, RNC) is currently processed using Mystem (Segalovich 2003), a tagger based on the grammatical dictionary and provided with the statistical module predicting the labels of out-of-vocabulary words. However, the distributional probabilities are different in the prosaic and poetic varieties. Table 1 demonstrates the differences in distribution of the part-of-speech (POS) tags based on the data of two RNC corpora, and the dissimilarities in lexical probabilities are expected to be even more noticeable, as the authors

[1] National Research University Higher School of Economics, Moscow, Russia; E-mail: aleksey-starchenko@mail.ru
[2] National Research University Higher School of Economics, Moscow, Russia; E-mail: lvkazakevich@edu.hse.ru
[3] National Research University Higher School of Economics, Moscow, Russia; Vinogradov Institute of the Russian Language RAS, Moscow, Russia; E-mail: olesar@yandex.ru
[4] The research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2018 (grant \#18-05-0047) and by the Russian Academic Excellence Project «5-100».

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, be 'creative' in the broadest sense of that term. Besides that, the rhythmic structure of poetry also affects syntactic patterns, word order, and the choice of lexical units. All these factors may challenge the cross-genre tagging and bias the prediction of the POS tags, grammatical features, and lemmas, i.e. three important constituents of the full morphological tagging.

Part of speech	Poetry	Prose
noun	30.3%	28.5%
verb	14.9%	17.0%
adjective	13.1%	12.8%
preposition	9.2%	10.5%
conjunction	7.7%	7.9%
adverb	5.5%	6.3%
pronoun	5.5%	7.9%
particle	3.5%	4.5%
interjection	0.2%	0.1%

Table 1. POS tags in the Poetry Corpus (11 MW) and the RNC Standard (prose, 6 MW).

Yet, developing a tool designed exclusively for poetry harbors its own risks. Enchanced lexicon, chargrams and syntax are associated with the sparsity of language models, and using the (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 a 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 (Section 4-6).

2. Distinctiveness datasets

The accuracy of the full morphology tagging applied to the modern languages is as high as 92-95% (Sorokin et al. 2017). The best accuracy of POS-tagging reported for languages like English and German is close to 97%-98% (Horsmann et al. 2015). With such

high scores in assessment, the difference in the taggers' performance can not be seen clearly. The idea behind the use of distinctiveness datasets (e.g. Rare Words dataset, Luong et al., 2013) is to provide the basis for more conservative, lower scores, taking only most challenging data.

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

Dataset A (750 words) is a sample drawn from the RNC Corpus of the Russian Poetry (Grishina 2013). It contains sentences with the high proportion of the out-of-vocabulary (OOV) words. Note that the notion of OOV words is different in the 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 the OOV word 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 would be unlikely present both in a dictionary and in a training collection.

In order to compile the Dataset A, we processed the word list of the Corpus of the 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:

• syllable dilation and contraction: *Zeves*, *poln*

• otrhographic distortion and variation: *što* (instead of *čto*), *šopot* (instead of *šepot*), *ra- // zjaschee* (*tvorchestvo skul'ptora*) (the word is divided by the line boundary)

• archaic and archaic-like words: *drugi*, *oblak* (m.)

• names, named entities: *Io, Eol, Sal'vaterre*

• (quasi-)loan words: *mus'je*

• paradigm extension, non-standard grammatical forms: *mysliju* (noun, Instrumental singular, cf. *mysl'ju*), *uš* (noun, Genitive plural, cf. *ušej*), *ostavja* (gerundive, Perferctive, cf.

ostaviv), *okazalasja* (reflexive verb, Past feminine, cf. *okazalasja*), *mjauchat* (verb, Present 3rd person plural, cf. *mjaukajut*)

• words with *pol-*: *poldorogi*, (na) *poldoroge*

As a next step, we inspected and ranked the OOV words as being easy / 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* and *okazalasja* 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 word were retrieved using the frequency database of the Corpus of the Russian Poetry (Lyashevskaya et al. 2018). 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 prižavšisja*** *k perstu, || V ten', nedostupnuju tumanam i vetram*.

Dataset B (850 words) is sample of syntactically complex and non-standard sentences. We use several syntactic templates which 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 čužoj* in(2);

• nouns in the genitive construction where the genitive form is placed before its head, cf. *kisti kiparisy* in(2);

• pre-position of the direct and indirect object, adverbial modifier; post-position of the subject with regard to the verb, cf. *Sveču* (object) *predpočitaem*, *sverkan'ju* (oblique) *predpočitaem* in (3);

• verb phrases, noun phrases with one or more clause or parenthetic construction inserted inside, see (4).

(2) ***Kisti***.Gen ***čužoj***.Adj ***kiparisy*** *i rozy || Prosalili belyj kak vosk amvon.*

(3) ***Sveču***.*Acc* ***sverkan'ju***.Dat *ljustr* ***predpočitaem***.

(4) *Čto khorošo by, vdrug* ***otoropev****,*
Kak ***vozčik*** *tot, otoropevšij,*
Poljubovat'sja uvjadan'jem dev,
Brjuškom prijatel'skim i pleš'ju.

The data were retrieved using the aforementioned frequency database.

Dataset C (1750 words) is an excerpt drown from the open-source manually annotated UD_Russian-Taiga treebank (Droganova et al. 2018). Among other genres, this corpus includes folk poetry published in social media. The Dataset C was meant to represents the 'average' level of complexity of poetic texts, even though the length of the sentences occurred to be larger in the Dataset C than in the Corpus of the Russian Poetry in general.

3. Taggers

To the date, a number of taggers have been tested on Russian (prosaic) data, both language-specific tools (Mystem, AOT (Sokirko 2004), PyMorphy (Korobov 2015), NLTK4RUSSIAN (Panicheva et al. 2015), UDAR (Reynolds 2015)) and general models (TreeTagger (Schmid, 1994), TnT (Brants 2000), MarMoT/Lemming (Müller et al. 2015), UDpipe (Straka et al. 2016)) trained on Russian data. Evaluation of taggers on the Russian prose data has been done within the framework of RU-EVAL 2010, MorphoRuEval 2017, SIGMORPHON 2016, CONLL 2018 shared tasks (Lyashevskaya et al. 2010, Sorokin 2017, Cotterell 2016, Lyashevskaya et al. 2017), see also evaluation experiments reported by (Kuzmenko 2016, Dereza et al. 2016).

In our study, we applied the following taggers to the material of the Russian poetry:

• **Mystem-RuSyntax**, an implementation of Mystem model currently used in the annotation of the Main RNC corpus (prose texts), with the addition of context rules for POS disambiguation (Droganova, Medyankin 2016);

• **Mystem 3.1**, a standard implementation of Mystem provided by Mystem+ (Dereza 2016);

• **TreeTagger** (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, Schmid, 1994), a tagger using automatic derivation of decision trees

• **Hunpos** (https://code.google.com/archive/p/hunpos/, Halácsy et al. 2007), a reimplementation of TnT tagger (Brants 2000) using a trigram based HMM model;

• **MarMoT** (http://cistern.cis.lmu.de/marmot/, Müller et al. 2015), a higher-order conditional random field (CRF) tagger;

• **Lemming** (http://cistern.cis.lmu.de/lemming/, Müller et al. 2015), 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/, Straka et al. 2016), a rich feature averaged perceptron tagger, a baseline for CONLL 2018 shared task;

• **Stanford** POS tagger (http://nlp.stanford.edu/software/, Toutanova et al. 2003), 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 collection (Zobnin, Nosyrev 2015), the grammatical dictionary is an extended version of (Zaliznyak 2003). Mystem-RuSyntax uses the model adopted to the RNC annotation guidelines (Lyashevskaya et al. 2005): 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 on and the type of output they provide. TreeTagger, Hunpos, and MarMoT were trained on the 6MW corpus of the Modern Russian prose (RNC Standard) in the framework of the Mystem+[5] project (Dereza 2016), 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 (Droganova et al. 2018). The Lemming model was trained by us on a 0.4 MW subcorpus of OpenCorpora prosaic texts (Bocharov et al. 2013).

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 (it adds them 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.

[5] http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/

4. Experiment setup

Gold labels. All datasets were labeled by 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 Figure 1. We followed the conversion rules of MorphoRuEval 2017 (Sorokin et al. 2017, Lyashevskaya et al. 2017) with some adjustments. Animacy and aspect are let in evaluation, and the participle and gerundive forms are treated as the 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 labeled systematically different in different frameworks (e. g. *kotoryj*) were marked off evaluation.

1 1 Мяучат МЯУЧОНОК NOUN Animacy=Anim|Case=Gen|Gender=Masc|Number=Plur
1 2 кошки КОШКА NOUN Animacy=Anim|Case=Gen|Gender=Masc|Number=Plur

Fig. 1. Annotations converted into UD-CONLLU format. The values in the first column indicate if the token is under evaluation.

It should be noted that Mystem 3.1 does not disambiguate among possible grammatical annotations available for the identified lemma and POS and provide them all in the 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 other, but nevertheless we can roughly compare the results of Mystem 3.1 applied to different Datasets (A, B, C).

Hypotheses. According to our assumption, 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 Database A, in which a large number of non-vocabulary words.

In that follows, we will analyse the results of the experiment and check if our assumptions hold.

5. POS tagging

Table 1 shows the accuracy of the POS tagging when applied to the Datasets A, B, C. The last row reports the results obtained on the prosaic texts in (Dereza 2016). Overall, the accuracy of the best systems ranges from 91.9% to 95.2% for the POS tags and from 82.4% to 92.6% for the feature tags on the poetic texts.

Surprisingly, none of the taggers is an absolute winner: Hunpos is the best on the Dataset A (OOV words), Stanford - on the Dataset B (complicated syntax), POS tags, and MarMoT - on the 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 the 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.

	MarMoT		Hunpos		TreeTagger		Stanford	Mystem 3.1	
	POS	Features	POS	Features	POS	Features	POS	POS	Features
Dataset A	93.1%	78.6%	**94.3%**	**82.4%**	87.4%	72.2%	94.1%	91.7%	67.7%
Dataset B	87.8%	**82.6%**	87.8%	79.9%	82.8%	70.6%	**91.9%**	88.5%	71.4%
Dataset C	**95.2%**	**85.5%**	94.3%	83.3%	90.9%	77.1%	93.9%	91.3%	65.8%
Mystem+	96%	—[6]	96.41%	89.29%	**96.94%**	**92.56%**	95.82%	96.43%	—[7]

Table 1. POS and feature tagging.

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 the Dataset B are certainly lower than the scores obtained on the general Dataset C. They are also lower than scores obtained on the Dataset A (in both POS and feature identification tasks, the only exception is MarMoT on feature tagging).

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 the Datasets C and A). Unlike MarMoT and TreeTagger, Hunpos and Stanford demonstrate approximately the same or slightly higher results on the Dataset A. Yet, the accuracy of the Marmot and

[6] The value was not reported in (Dereza 2016) since there was no enough memory to train the taggers.
[7] The value was not reported in (Dereza 2016) since there was no enough memory to train the taggers.

TreeTagger's features decreases considerably as we go from the Dataset C to the Dataset A, as expected.

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

Comparing the accuracy of processing the 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 the greater bias to the type of text than the other taggers.

Table 2 summarizes the correspondence of the gold POS tags (columns) and those predicted by Lemming (lines). Since its accuracy is lower compared to the taggers described in the previous section, we can get enough error data in order to analyze them in more detail.

The table shows the number of words corresponding to each combination of the goal and predicted tags. In each cell, a number of occurrences is given, below thich the row percentages and the column percentages are given. In other words, 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. It can be seen that the words that constitute the small closed classes — conjunctions and prepositions — are most accurately identified. On the opposite, the accuracy of the processing the adverbs is low, almost close to chance. Such a small accuracy can be explained by a large syntactic freedom of adverbs: many adverbs can appear anywhere in the sentence. In addition, a number of errors are caused by the annotation practice in the corpus on which Lemming was trained. Thus, there is a category of 'Praedic' corresponding to the predicatives. This group includes words of different types: adjectival predicates ending on *-o* / *-e* (*khorosho*, *blizko*), predicative nouns (*pora*, *len'*), modal predicates (*dolžen*, *možno*), the negative word *net*. If such a category is not present in the corpus tagset, the predicative words are distributed among other

Gold ↓	Lemming												
	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%

Table 2. Confusion matrix: POS tags.

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

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

6. Lemmatization

In this section we focus on lemmatization. We analyse 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 such a difference, Table 3 also summarizes data on the accuracy of the POS tagging.

	Lemming		Mystem	
	Lemma	POS	Lemma	POS
DatasetA	85.0%	87.7%	87.7%	91.7%
DatasetB	87.7%	87.3%	86.4%	88.5%
DatasetC	87.9%	88.4%	91.4%	91.3%

Table 3. Lemmatization.

Is can be seen that the quality of lemmatization by Lemming и Mystem varies weekly depending on dataset; we can only point out that for the rule lemmatizer, both the datasets with the complex syntactic constructions (B) and the dataset with the out-of-vocabulary words (A) are problematic.

Interestingly, although Lemming learned on a small data set, its accuracy is close to the accuracy of the rule lemmatizer.

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'* 'no'.

Expectedly, 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. *banka* and *banka*. 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 the taggers of different types in the task of the full morphological annotation of the poetic texts. As expected, the poetry in general turns out to be difficult for processing by the taggers designed and trained on prose, the non-standard 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%-92.6%), which can be explained by the complexity of the classification task itself and by some agreements of data evaluation which we follow.

In order to analyze in more detail in which cases the error is more probable and why, a confusion matrix was compiled. The adverbs are most difficult to parse, the least complex are prepositions and conjunctions.

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 the tagger. However, in order to make final conclusions about the accuracy of lemmatization, it is necessary to check these results on other models trained on larger data sets.

The complexity of the structures in the poetic texts and the small amount of the test data may explain the mixed results achieved with the method of distinctive datasets. We did not control for syntactic complexity while mining the dataset for OOV words and vise versa.

None of the parameters was controlled while randomly sampling the Dataset C. The more promising approach would be to annotate according to multiple parameters the word entries 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 non-standard word order, words which have counterparts with overlapping paradigms, and other parameters of the test data profiling.

References

Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V. (2013). Crowdsourcing morphological annotation. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013". Moscow.

Brants T. (2000). TnT - a statistical part-of-speech tagger. In: Proceedings of 6th Applied Natural Language Processing Conference, Seattle, 224–231.

Cotterell R., Kirov Ch., Sylak-Glassman J., Yarowsky D., Eisner J., Hulden M. (2016). The SIGMORPHON 2016 shared task—morphological reinflection. In: Proceedings of the 2016 Meeting of SIGMORPHON.

Dereza O. V., Kayutenko D. A., Fenogenova A. S. (2016) Automatic morphological. analysis for Russian: A comparative study. In Proceedings of the International. Conference Dialogue 2016.

Droganova K., Lyashevskaya O., Zeman D. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks, in: Proceedings of TLT 2018 International Workshop on Treebanks and Linguistic Theories, 13-14 November 2018, Oslo, Norway. NEALT Proceedings Series. Linköping University Electronic Press, 2018.

Droganova K. A., Medyankin N. S. (2016). NLP pipeline for Russian: an easy-to-use web application for morphological and syntactic annotation. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". Moscow.

Grishina E., Korchagin K., Plungian V., Sichinava D. Poeticheskij korpus v ramkakh Nalsional'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]. 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.

Halácsy P., Kornai A., Oravecz Cs. (2007). Hunpos: An open source trigram tagger. InProceedings of the 45th Annual Meeting of the ACLon Interactive Poster and Demonstration Sessions,ACL '07, pages 209–212, Stroudsburg, PA, USA.

Horsmann T., Erbs N., Zesch T. (2015). Fast or accurate? – a comparative evaluation of pos tagging models. In Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology.

Korobov M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In: Proceedings of AIST-2015, International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, pp. 320-332.

Kuleva A. S. Istorija usechennykh prilagatel'nykh v jazyke russkoj poezii [In Russian: Clippped adjectives in the history of the Russian poetry]. Moscow, St-Petersburg: Nestor-Istorija: 2017.

Kuzmenko E. (2016). Morphological analysis for Russian: integration and comparison of taggers. In: Proceedings of AIST-2016, International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 162-171.

Luong T., Socher R., Manning C. (2013). Better word representations with recursive neural networks for morphology. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 104-113.

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. (2010). Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskie parsery russkogo jazyka. In: Proceedings of Dialogue 2010, International Conference on Computational Linguistics and Intellectual Technologies. Moscow, 318–326.

Lyashevskaya O., Bocharov V., Sorokin A., Shavrina T., Granovsky D., Alexeeva S. (2017). Text collections for evaluation of Russian morphological taggers. Jazykovedný casopis, 68(2), 258-267.

Lyashevskaya O., Plunguan V., Sichinava D. (2005). 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, Moscow, 111-135.

Lyashevskaya O., Vlasova E., Litvintseva K., Starchenko A. (2018). A Data Analysis Tool for the Corpus of Russian Poetry. WP BRP Linguistics.

Müller T., Cotterell R., Fraser A., Schütze H. (2015). Joint lemmatization and morphological tagging with lemming. In: Proceedings of EMNLP-2015, Conference on Empirical Methods in Natural Language Processing, 2268-2274.

Panicheva P., Protopopova E., Mitrofanova O., Mirzagitova A. (2015). 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: Proceedings of CORPORA-2015, Saint-Petersburg.

Reynolds, R. (2016). Russian natural language processing for computer-assisted language learning: Capturing the benefits of deep morphological analysis in real-life applications. PhD diss. University of Tromso.

Segalovich I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: Proceedings of MLMTA-2003, Las Vegas.

Sokirko, A. (2004). Morphologicheskie moduli na sajte www.aot.ru [Morphological tools on the website www.aot.ru]. In: Proceedings of Dialog'04, Moscow.

Sorokin A., Shavrina T., Lyashevskaya O., Bocharov B., Alexeeva S., Droganova K., Granovsky D. (2017). MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In: Proceedings of Dialog-2017, International Conference on Computational Linguistics and Intellectual Technologies. Moscow.

Straka M., Hajic J., Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: Proceedings of LREC-2016.

Sharoff S., Nivre J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Sharoff S, Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Proceedings of Dialogue 2011, International Conference on Computational Linguistics and Intellectual Technologies. Moscow. https://pdfs.semanticscholar.org/36df/5fbe04f425e9b089437e979581d1f5375a94.pdf

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, Manchester.

Toutanova K., Klein D., Manning Ch. D., Singer Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, 173–180.

Zaliznyak A. A. (2003). Grammaticheskij slovar' russkogo jazyka [The grammatical dictionary of Russian]. Moscow.

Zobnin A. I., Nosyrev G. V. (2015). Morfologicheskij analizator Mystem 3.0 [A morphological analyzer Mystem 3.0]. In: Trudy Instituta russkogo jazyka im. V.V.Vinogradova [Works of Vinogradov Institute of the Russian Language RAS], (6), 300-310.

Contact details:

Olga Lyashevskaya
National Research University Higher School of Economics (Moscow, Russia), School of linguistics, Professor;
E-mail: olesar@yandex.ru