Morphological Analysis for Russian: Integration and Comparison of Taggers

Elizaveta Kuzmenko

National Research University Higher School of Economics
Moscow, Russia
eakuzmenko_2@hse.edu.ru

Abstract In this paper we present a comparison of three morphological taggers for Russian with regard to the quality of morphological disambiguation performed by these taggers. We test the quality of the analysis in three different ways: lemmatization, POS-tagging and assigning full morphological tags. We analyze the mistakes made by the taggers, outline their strengths and weaknesses, and present a possible way to improve the quality of morphological analysis for Russian.

Keywords: morphological analysis, Russian, POS-tagging, gold standard

1 Introduction

In this paper we present the results of testing different morphological taggers for the Russian language. Russian is a highly inflective and morphologically rich language, and developing high-quality morphological tools for Russian presents a serious problem even for advanced researchers.

A considerable number of taggers provide morphological disambiguation while performing POS-tagging for Russian, but all of them are erroneous in some way. However, this disadvantage can be beneficial since the taggers make errors in different issues: when one analyzer fails, another may guess the correct tag. Therefore, it could be very useful to inspect the performance of each tagger and reveal the specificity of the mistakes it makes. These findings can then help to build an improved tagger for Russian that will combine in itself all the forces of other taggers. The near future of morphological analysis of Russian, as we see it, is meta-learning, in which all the cases where taggers guess tags correctly are taken and all the cases where the taggers make errors are omitted.

The question is then: do the cases where taggers make errors overlap or not? We answer this question in our paper via the experiment in which we build a gold standard corpus and compare the tags found in this corpus to those that are output by our taggers. In case of discrepancy, we analyze the cause of an error.

The structure of this paper is as follows: in Section 2 we describe the previous work in this field: how the standards for morphology annotation were defined and what the specific morphology problems for the Russian language are. We also give an overview of the instruments developed for Russian: taggers *Freeling*, *Pymorphy*, *MyStem* and *TreeTagger*, and describe previous attempts to compare their performance. In Section 3 we present an experiment in comparing the taggers: analyze the differences in the tagsets and define the rules for unification of morphological tags. In Section 4 we present the results of our experiment, and in Section 6 we discuss these results and propose the way towards organizing meta-learning of the taggers.

2 Background

The Russian language presents certain problems with regard to morphology annotation, because it is a highly inflectional language with many grammatical categories. There is no standard even for part-of-speech annotation, let alone subtle grammatical categories such as (im)perfectiveness and animacy. Theoretical disputes concerning Russian morphology lead to variety of solutions for morphology annotation – from positional tags following the MULTEXT-East guidelines [1] to combinations of tags employed in RNC[1]. An additional problem arises from the fact that tags in Russian can be combined and simplified in different ways. Some systems do not account for one or another grammatical category (for example, transitivity in *TreeTagger*), whereas other systems define some value of a category as the default one. Thus, the active voice of verbs is not marked in *pymorphy*, which, in its turn, follows the *OpenCorpora* guidelines [2].

The comparison of taggers for Russian is also complicated by the fact that there are different theoretical traditions for the lemmatization process. For example, some taggers count for verbs as lemmas of participles, and other taggers lemmatize participles as adjectives.

In addition to not having unified rules for morphology annotation in Russian, until recently there were no standard golden corpus of any kind. Presently, there are two corpora that could serve as models for annotation tasks: a disambiguated subcorpus of *RNC* and *Opencorpora* [2]. Moreover, there has been organized the *RU-EVAL* shared task [3], in which the participants proposed unification rules for the output of different morphological taggers and created a gold standard corpus consisting of 3 thousand word tokens.

The taggers used in our experiment are the following ones:

- *MyStem*[2] [4] is a morphological analyzer with disambiguation developed for the Russian language by Ilya Segalovich and Vitaliy Titov at “Yandex”. In the core of the software lies a dictionary that helps generate morphological hypotheses for both known and unknown words.
- *Pymorphy2*[3] [5] is a morphological analyzer developed for the Russian language by Mikhail Korobov on the basis of OpenCorpora dictionaries. PyMorphy2 is written fully in the Python programming language and is able

[1] http://ruscorpora.ru/

[2] https://tech.yandex.ru/mystem/

[3] https://pymorphy2.readthedocs.org/en/latest/

to normalize, decline and conjugate words, provide analyses or give predictions for unknown words.

- *Freeling* [6] is a set of open source linguistic analyzers for several languages. It features tokenizing, sentence splitting, morphology analyzers with disambiguation, syntax parsing, named entity recognition, etc. In this research, we use only morphological analyzer for Russian.
- *TreeTagger* [7][8] is a language independent part-of-speech tagger developed by Helmut Schmid. TreeTagger is based on decision trees and should be trained on a lexicon and a manually tagged training corpus. The program can annotate texts with part-of-speech and lemma information.

  These are not all existing morphological analyzers for Russian. The choice of taggers for the comparison was motivated by their availability. For example, the *TnT* tagger, which has trained models for Russian [9], is not freely available, and we faced some problems when obtaining it from the developers. However, our work still tests all major analyzers for Russian.

  Within the chosen set of analyzers, there are several issues connected to their comparability. Apart from different guidelines for lemmatization and assigning morphological categories, the taggers also feature various algorithmic designs. Thus, *pymorphy* analyzes tokens separately, without taking the context into account, whereas other analyzers determine the word characteristics from its neighboring words. However, we do not judge from the developer's point of views and do not evaluate the efficiency of various POS-tagging techniques. We take each tagger as a final product and estimate their efficiency from the user's point of view.

3 Experiment design

In this work we evaluate the taggers' performance on two gold standard sets. The first set is the disambiguated subcorpus of the RNC, and the focus of evaluation is on the strict correspondence between taggers' output and the RNC data. The second set is taken from the *RU-EVAL* competition [3]. In this case we do not strictly follow the RNC guidelines and do not count the absence of some categories in the output as an error (for example, the absence of the active voice for verbs in the *pymorphy* analysis is not taken into account), and the resulting figures can be considered more objective.

The disambiguated subcorpus of Russian National Corpus contains 5.9 million tokens, annotated morphologically with the help of *MyStem* and further disambiguated and refined by hand. All tokens have only one morphological analysis, and the tagset in this corpus generally complies to the one developed for *MyStem*. An example of an annotated sentence can be found below.

```
<se>
<w><ana lex="береза" gr="S,f,inan=sg,nom"/>Берёза</w>
<w><ana lex="ждать" gr="V,ipf,tran,act=sg,praes,3p,indic"/>ждёт</w>
<w><ana lex="мороз" gr="S,m,inan=sg,gen"/>мор‘оза</w>!"
</se>
```

However, if we choose only RNC as the gold standard, this leads to some limitations. First, Tretagger was trained on the disambiguated subcorpus of RNC, so it has some advantage compared to other taggers. Second, RNC has a very balanced and detailed tagset, but it is sensible to exclude some grammatical categories from the analysis, as they are highly important only for purely linguistic tasks. Thus, we also use the second gold standard set from the *RU-EVAL* task. This set contains 3300 tokens, annotated by hand. An example of an annotated sentence can be found below.

```
как как CONJ
казалось казаться V n,past,sg
раньше раньше ADV
```

One of the problems in our experiment is that all analyzers have different notations for parts of speech and morphological categories. The discrepancies between the tagsets can be of different kinds:

- **Some morphological category is present in the tagset of the gold standard but absent in the tagset of another morphological analyzer**: for example, *Mystem* distinguishes between animacy and inanimateness as it has specific dictionaries where these characteristics are defined for every word. *TreeTagger*, however, does not consider this feature to be important and does not include it in the analysis.
- **Morphological analyzers have different standards concerning part of speech identification**: for example, *Freeling* identifies participles as a separate part of speech, whereas other morphological analyzers identify participles as verbal forms.
- Consequently, **alongside with different standards towards part of speech identification, parsers assign different lemmas to tokens problematic in this aspect**: therefore, the lemma for the word '*сделанной*' would be '*сделанный*' in *Freeling* and '*сделать*' in *Mystem*.
- **If the part of speech is identified uniformly by the taggers, there still can be problems with lemmatization**: for example, *TreeTagger* assigns one and the same lemma to Russian verbs in different aspects, and so does *Freeling*. For example, the verbs '*выплывать*' and '*выплыть*' will be assigned one and the same lemma '*выплывать*', even if the aspect of a given word instance is reflected in its analysis. At the same time, other tagsets do not require the aspect to be changed in the process of lemmatization.

Due to these problems, we need to define conventions that will allow to make comparison of the taggers possible and more correct. As our gold standard is annotated by *MyStem*, we decided to convert all our tags into *MyStem* tags. The rules of conversion are presented in Table 1.

The rules for conversion into the *RU-EVAL* tagset were the same. In addition, we excluded from the analysis the following cases:

- absence of voice and mood for verbs;

Table 1. Rules for conversion of the tagset into the tagset defined for *RNC*

Gold standard tag	**Tag counted as correct**
A-NUM (numeric adj.)	NUM (numeral)
PARENTH(parenthesis)	ADV(adverb)
ADV-PRO (adv.-pronoun)	PRO (pronoun)
A-PRO (adj.-pronoun)	PRO (pronoun)
m-f (common gender)	both are correct
anim (animacy)	not important
inan (inanimateness)	not important
dat2 (the 2nd dative)	dat (dative)
gen2 (the 2nd genitive)	gen (genitive)
acc2 (the 2nd accusative)	acc (accusative)
loc2 (the 2nd locative)	loc (locative)
adnum (count form)	NUM (numeral)
intr (intransitiveness)	not important
tran (transitiveness)	not important

– confusion between predicates and other parts of speech;
– verbs which end with '*ся*';
– numerals;
– distinction between full and shortened forms for adjectives and participles.

In general, these rules mean that we accept as the right output less specific tags, for example, *dat* (the dative case) instead of *dat2* (the second dative). This leads to loss of some linguistic information, but accounts for the tagsets with less strict linguistic background. The rules for the RU-EVAL tagset in addition eliminate cases when the taggers' results differ because of tagging guidelines.

However, these are the rules only for the least problematic cases. The most problematic cases include, as it was mentioned earlier, lemmatization of participles and perfective verbs. These issues we solve by assigning lemmas given by the analyzer and taking the tag itself from another analyzer. In addition, we do not consider identifying patronyms, zoonyms and other lexical classes to be of importance for the task of morphological analysis and exclude them from our experiment.

The experimental procedure itself was as follows:

1. take the text files from the gold standard corpus and extract the tokens and their morphological characteristics;
2. analyze the tokens by the taggers in question;
3. convert the output into the RNC tagset;
4. compare token by token the output from the tagger to the morphological analysis found in the gold standard corpus.

4 Evaluation

For each word we compared the analyses of the three taggers and the analysis given in the gold standard corpus. In particular, we checked whether the part of speech was the same and if the set of grammatical categories contained in the tag was identical to the gold standard. There were three modes of evaluation:

1. checking the correspondence between assigned lemmas;
2. checking the correspondence between assigned parts of speech;
3. checking the correspondence between assigned morphological tags in the whole.

If the lemma, the part of speech or the tag output by the tagger agreed with the gold standard, the answer of the tagger was counted as correct for the corresponding evaluation mode. Thus, the performance of the taggers was evaluated using the accuracy metric, roughly, the proportion of correct answers given by a tagger. Table 2 presents the results for all our taggers in three modes and two sets.

Table 2. Evaluation of the taggers' performance

Tagger	**Mode**	**Accuracy**	
		RNC	**RU-EVAL**
Freeling	lemma	0.822	0.816
	POS	0.907	0.911
	full tag	0.833	0.851
Pymorphy	lemma	0.882	0.871
	POS	0.915	0.904
	full tag	0.647	0.742
TreeTagger	lemma	0.970	0.869
	POS	0.952	0.882
	full tag	0.924	0.863

As it can be seen from the Table 2, all the taggers present decent results, but none of them perform without mistakes. *TreeTagger* was trained on the disambiguated subcorpus of RNC, and after we apply it to the *RU-EVAL* gold standard, the quality of its analysis gets worse. Other taggers perform slightly better in the *full tag* mode because of milder error criteria.

5 The analysis of the errors

After evaluating overall taggers' performance, let's have a look at the nature of the errors.

As it was said earlier, the variety of annotation guidelines makes the very notion of error in this task very ambiguous. Should a particular case of discrepancy between two taggers be attributed to the bad performance of one of them or to the differences in their guidelines? For example, if some tagger analyzes the word '*здесь*' as a predicate, and another taggers considers it as an adverb, which answer is the right one? Or, as it was described earlier, if *pymorphy* presupposes the active voice for all verbs and doesn't explicitly mark this, is this the underperformance or a tagger's feature?

There can a lot of reasoning on these grounds, and no resolution can be considered as accurate. For the purpose of our analysis we count all cases of discrepancies between the gold standard and another tagger to be errors. The tagset designed for RNC is very exhaustive and detailed, and any differences which are not taken into account by conversion rules signify either the loss of information or a proper error. Thus, the absence of active voice in the analyses of *pymorphy* is considered to be an error, as well as different representations for parts of speech (for example, analyzing a substantivized adjective '*новое*' as a noun or an adjective).

We do not claim for our definition of an error to be the ground truth. Other conventions for the correspondences between tagsets can lead to alternative figures. However, we take our decision for a balanced one and appealing to the task of morphological analyzing compliant with the RNC tagset.

Table 3 gives the figures for the taggers performance in *POS* and *lemma* modes with regard to the POS tag of a given word as determined by the tagger. The gold standard set in this task was the disambiguated subcorpus of RNC. Thus, for all words analyzed as nouns by *Freeling*, 4% of them proved not to be nouns in the gold standard set, and almost 17% of them did not match the gold standard lemma.

These figures allow to draw several interesting conclusion about the taggers' performance.

- The main parts of speech (such as nouns, verbs, and adjectives) are less prone to errors while tagging. The same is true in most cases for auxiliary parts of speech, as they form a closed subset.
- The heel of Achilles for the taggers are such parts of speech as pronouns of different types and categories on the border between two parts of speech. In these cases it is likely that taggers would have different tagging guidelines.
- purely erroneous tagging of a particular part of speech indicates that either this category is absent in the tagset or it is tagged as another POS. Probably, such cases should be eliminated from the analysis and evaluation.
- striking difference between error rate for POS and for lemma implies that there is a conflict between lemmatizing standards. Thus, 42% of wrong answers for verb lemmas in the *Freeling* data can be attributed mostly to the change of aspect. This is probably should be eliminated from the analysis as well, or the lemmas should be defined uniformly with the help of a dictionary of aspectual pairs.

Table 3. Proportion of wrong answers given in **POS** and **lemma** modes depending on the POS of a token

POS	Freeling		Pymorphy		TreeTagger	
	POS	lemma	POS	lemma	POS	lemma
S	0.039	0.166	0.043	0.080	0.027	0.092
S-PRO	0.144	0.090	0.075	0.077	1	0.114
V	0.018	0.422	0.022	0.029	0.024	0.377
ADJ	0.175	0.233	0.115	0.126	0.094	0.197
ADJ-PRO	0.085	0.085	0.195	0.181	1	0.117
ADJ-NUM	1	0.015	0.026	0.995	0.238	0.006
ADV	0.377	0.098	0.426	0.055	0.226	0.009
ADV-PRO	0.059	0.001	0.762	0.558	0.009	0.002
PR	0.004	0.001	0.008	0.034	0.002	0.001
CONJ	0.055	0.008	0.233	0.033	0.060	0.008
NUM	0.054	0.112	0.005	0.005	0.042	0.098
PART	0.061	0.004	0.191	0.014	0.008	0.001
INTJ	0.279	0.177	0.728	0.516	0.118	0.113

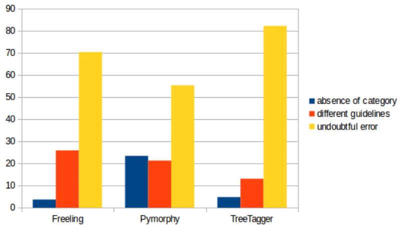

Figure 1. Causes of errors among the taggers

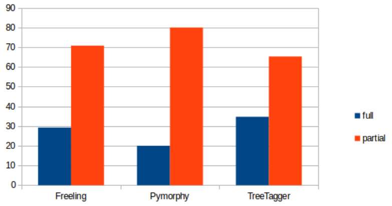

Figure 2. The percentage of full or partial errors

All in all, it can be seen from Table 3 that the errors produced by taggers do not overlap in most cases. Trusting *pymorphy* on its output for interjections (INTJ) is not the best option, whereas TreeTagger shows high results in this case. On the contrary, *pymorphy* has the lowest percentage of errors for numerals (both lemma and POS) while other taggers stand down for this part of speech. This makes possible the meta-learning technique we described in the beginning of the present paper.

As for the errors made in the *full tag* mode, they are less dependent on POS. Besides, the main parts of speech (nouns, verbs, adjective) have more grammatical characteristics and thus give more space for errors. We performed the analysis of error causes in 500 erroneous cases for every tagger by hand. These results can also be of interest, though they are not formalized.

Figure 1 depicts the percentage of different error causes among the taggers. **Absence of category** accounts for cases when tags did not match because some category is not present in an analyzer's tagset. This is the case, for example, for the absence of the active voice in *pymorphy*. **Different guidelines** refers to the cases when two taggers treat the word differently because of diverse approaches to the issue. For example, it accounts for the confusion between *predicate* and *adverb* parts of speech.

Figure 2 demonstrates the percentage of 'full' and 'partial' errors in tags. Full errors are mostly represented by confusion between POS tags. If a word is tagged as a noun by TreeTagger, and in the gold standard it is an adjective, that would be the 'full' error. 'Partial' errors concern one or two categories that do not match the gold standard tag. This is the case with mismatch between assigned cases or gender.

6 Conclusion

In this paper we presented an analysis of the performance of three taggers for Russian. The comparison procedure was performed in three modes: assigning the POS tag, assigning lemma and assigning the full tag. Apart from evaluating the accuracy of each tagger, we analyzed the errors made by the taggers. The proportion of errors connected to different parts of speech shows that the errors produced by the taggers do not overlap. For almost every POS tag there is an analyzer that has high accuracy and an analyzer that performs significantly worse. At the same time, all the taggers show decent performance, so there is no tagger that would lose all the modes of comparison.

The received results are of interest to anyone engaged in morphological analysis of Russian. As a future step we plan to build a meta-learning system based on several taggers. Such system will take as input the morphological analyses from several taggers, identify which tagger provides the best guess for each particular case, and give as output the combination of correct variants. We expect this system to be highly accurate.

As the future work, apart from building an analyzer with meta-learning, we plan to investigate more thoroughly in which cases the taggers are more prone to errors, and what are the exact causes of these errors for every analyzer.

7 Acknowledgments

I would like to thank Elmira Mustakimova, Svetlana Toldova and Timofey Arkhangelskiy for their participation in the project. I am also grateful to Mikhail Korobov for his valuable remarks on *pymorphy* performance and explanations of error causes (and for developing *pymorphy*, of course).

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE).

References

1. Erjavec, T.: Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In: LREC. (2004)
2. Bocharov, V., Bichineva, S., Granovsky, D., Ostapuk, N., Stepanova, M.: Quality assurance tools in the opencorpora project. In: Computational Linguistics and Intelligent Technology: Proceeding of the International Conference «Dialog. (2011) 10–17
3. Astaf'eva, I., Bonch-Osmolovskaya, A., Garejshina, A., Grishina, J., D'jachkov, V., Ionov, M., Koroleva, A., Kudrinsky, M., Lityagina, A., Luchina, E., et al.: Nlp evaluation: Russian morphological parsers. In: Proceedings of Dialog Conference, Moscow, Russia. (2010)
4. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA, Citeseer (2003) 273–280
5. 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.: Analysis of Images, Social Networks and Texts. Volume 542 of Communications in Computer and Information Science. Springer International Publishing (2015) 320–332
6. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC2012. (2012)
7. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: In Proceedings of the ACL SIGDAT-Workshop, Citeseer (1995)
8. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing. Volume 12., Citeseer (1994) 44–49
9. Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., Divjak, D.: Designing and evaluating a russian tagset. In: LREC. (2008)