Morphological annotation of social media corpora with reference to its reliability for linguistic research

Mariia Michurina
ABBYY Lab, MIPT,
Dolgoprudny, Russia
RSUH, Moscow, Russia
marimitchurina@gmail.com

Alexandra Ivoylova
ABBYY Lab, MIPT,
Dolgoprudny, Russia
RSUH, Moscow, Russia
aleksandra.ivoilova@abbyy.com

Nikolay Kopylov
ABBYY Lab, MIPT,
Dolgoprudny, Russia
nikolay.kopylov@abbyy.com

Daniil Selegey
ABBYY Lab, MIPT,
Dolgoprudny, Russia
daniil.selegey@abbyy.com

Abstract

This paper presents the results of the study devoted to the applicability of SOTA methods for morphological corpus annotation (based on GramEval2020) for analytical sociolinguistic research. The study shows that statistically successful technologies of morphosyntactic annotation for such purposes create a number of problems for researchers if they are used purely i.e. without any linguistic knowledge. In this paper, methods for improving the morphological annotation, successfully implemented in GICR, from the point of view of its reliability are presented.

Keywords: automatic morphotagging, morphosyntactic annotation, lemmatization, NLP evaluation, morpho-parsers for Russian, language of social media

DOI: 10.28995/2075-7182-2021-20-492-504

Морфоразметка корпуса текстов из социальных сетей с точки зрения надежности лингвистических исследований

Мария Мичурина
ABBYY Lab, МФТИ,
Долгопрудный, Россия
РГГУ, Москва, Россия
marimitchurina@gmail.com

Александра Ивойлова
ABBYY Lab, МФТИ,
Долгопрудный, Россия
РГГУ, Москва, Россия
aleksandra.ivoilova@abbyy.com

Николай Копылов
ABBYY Lab, МФТИ,
Долгопрудный, Россия
nikolay.kopylov@abbyy.com

Даниил Селегей
ABBYY Lab, МФТИ,
Долгопрудный, Россия
daniil.selegey@abbyy.com

Аннотация

В работе приводятся результаты проведенного исследования по применимости SOTA-методов морфоразметки русскоязычных корпусов (по данным GramEval2020) для аналитических социолингвистических исследований. Показано, что механическое применение статистически успешных технологий разметки для таких целей порождает ряд проблем для исследователя - теоретического лингвиста. Приводятся методы улучшения разметки с точки зрения надежности получаемых результатов, успешно примененные при создании новой версии ГИКРЯ.

Ключевые слова: автоматическая морфоразметка, морфосинтаксический анализ, лемматизация, оценка систем автоматической обработки текста, морфопарсеры для русского языка, язык социальных медиа

1 Introduction

In modern linguistic research, the so-called mega corpora [3], or extra-large corpora [5], created according to the Web as Corpora (WAC) technology and containing billions of words, are widely used. It is quite obvious that manual annotation in such corpora is an unbearable task for a linguist. Thus, the only option is automatic annotation.

This paper examines the quality of automatic morphosyntactic annotation of mega corpora for sociolinguistic studies of Russian, processed by SOTA methods, which participated in the GramEval2020 competition [10]. The integral morphosyntactic parser for Russian[1] [1] was selected for our evaluation as it had achieved best results in the evaluation (hereinafter referred to as IMParser). The research was carried out within the framework of the new version of the General Internet Corpus of Russian (GICR) [2]. The GICR is one of the four existing mega corpora of Russian (the other three are ruTenTen [8], Aranea [4] and Taiga [14]. Unlike ruTenTen and Aranea, GICR is a differentiated corpus, i.e., divided into segments depending on the source of the texts. From our point of view, it is the advantage of GICR as it allows us to test the IMParser on texts from different segments of Russian social networks that may be hard for parsers. Taiga, on the other hand, is a corpus designed for computational linguists and NLP-specialists, not linguistic researchers [14].

Thus, the work evaluates the progress in the field of automatic corpus annotation over the past few years: the TnT parser [6], [13], used in the first version of the GICR, is a typical representative of statistical automatic parsers (for example, Aranea was annotated with the Tree Tagger [16], and ruTenTen with Tree Tagger and RFTagger [17]; both mentioned parsers, just like TnT, use hidden Markov models and, therefore, the quality of their annotation does not differ much. IMParser is, in a sense, a typical representative of the new generation of parsers. Consequently, on the one hand, they are the standard representatives of the parsers of their generation and allow us to assess the progress of text processing methods in general. On the other hand, both have been used in the GICR, and it is important for us to evaluate the improvement in the annotation quality.

There are a number of morphological parsers that use different formats and quite a few of them have their own tagsets (e.g., SynTagRus, OpenCorpora, RNC, MSD-GICR, MULTEXT-East, etc.). However, there is a Universal Dependencies (UD) project [11], which annotation guidelines unite more and more languages and corpora. Its use seems quite promising to us. It is the UD annotation scheme that IMParser uses.

According to the purpose of this study, we were faced with the following tasks:

- Evaluate the work of the parser in relation to various phenomena that should be of interest to users of such corpora as GICR; determine the benefits of integrating morphosyntactic annotation;
- Propose a new pipeline for corpus annotation, which gives a satisfactory final result from the point of view of a linguistic researcher, including adjusting the work of the parser;
- Assess the applicability of UD as a corpus annotation scheme for linguistic and sociolinguistic studies of the Russian language.

The GICR is intended primarily for theoretical linguistic research. Therefore, the quality of lemmatization, PoS-labelling, and disambiguation is important here. In this regard, in our work we carried out not only a numerical assessment, comparing the percentage of parsing accuracy, but also a manual quality assessment. Moreover, not only the quality of data annotation processing is important, but also its speed (the standard sizes of mega corpora force their developers to pay attention to it). Tests on GramEval data have shown that solutions based on pretrained BERT models are slightly better than fine-tuned ELMo ones, but they are much slower.

Below in this article, the results of solving the listed problems will be considered in detail: in the second paragraph, we will talk about the work of IMParser on the GICR data, in the third, the adaptation of the UD scheme for tagging the GICR will be discussed.

[1] https://github.com/DanAnastasyev/GramEval2020

2 The IMParser and its evaluation

The overall accuracy for five genres of the modern Russian language ("news, social media and electronic communication, wiki-texts, fiction, poetry; Middle Russian texts are used as the sixth test set" [10]) of the IMParser is the following: 0.916 versus the baseline accuracy of 0.804 (rnnmorph for lemmatization and morphology and UDPipe for syntax).

The IMParser is a combination of three interacting elements. Firstly, this is a fairly simple classifier that predicts the lemmatization rule for a word form, secondly, it is a morphological parser based on embeddings, and thirdly, a dependency parser. For morphological analysis, the model can use different versions of embeddings (character-level embeddings as well as two variants of contextual embeddings that have already proven themselves in NLP: BERT and ELMo); moreover, the parser uses grammeme embeddings that contain information about the grammatical meanings of a word and give the parser information about the interaction of this word with others. Models with contextual embeddings, especially BERT model, have shown the best quality. Dependency parser uses Edmonds' algorithm for finding minimum spanning trees on directed graphs for decoding; it produces syntactic parsing within the UD guidelines. All the three elements interact with each other: "The latter model uses shared representations between the morphological parser, the lemmatizer and the dependency parser" [1]. Thus, the parser simultaneously processes lemmatization, morphological tagging and syntactic parsing. A similar approach to automatic data annotation has already been used before, for example, in ETAP-4 [7], and the GramEval2020 rules were based on the decision that morphology and syntax should be analyzed simultaneously and be related.

2.1 Quality evaluation methods of automatic annotation

One of the main tasks of this study is manually evaluating the quality of automatic annotation of IMParser, which is of particular interest because of its multitask approach.

During the GramEval2020 competition, quality testing was carried out. It was aimed at cross-system comparison: the organizers automatically compared the manual annotation (gold set, inaccessible to the participants of the competition) and parser annotations, and published the average score, paying special attention to errors common to all systems, which directly follows from the objectives of the competition. We are interested in meaningful analysis, including analysis of particular ambiguous units. We want to understand to what extent the integral parsing methods are applicable for labelling corpora, which are intended not for NLP tasks, but for studies of the language. General quality metrics are important, but some types of errors may be unacceptable for language research.

As part of our research, we manually compared the quality of the IMParser to the end-to-end morphological analysis of the TnT parser for the Russian language. The main concern is both lemmatization and PoS-tagging. A special quality evaluation of IMParser's verb and noun lemmatization was carried out due to cases with a complex paradigm in these parts of speech. For this experiment, 10,000 tokens of random sentences were taken from VKontakte segment of the GICR. We also focused on the quality of lemmatization of out-of-vocabulary (OOV) words: lexemes that are absent from both standard dictionaries and training data. Social network texts usually possess newly created lexemes [15] and therefore a morphosyntactic parser has to cope well enough with such things. Finally, to test the claim that integrated morphosyntactic parsing improves the quality of disambiguation, some experiments with full and PoS homonyms were carried out.

2.2 Quality evaluation results

An experiment comparing the annotation of the TnT-parser and the IMParser gave the following results: the TnT-parser is not good enough in disambiguation, like other parsers in its category. However, IMParser has serious problems with lemmatization of non-homonymous word forms, namely verbs (see Fig. 1). This, apparently, is due to the fact that the parser does not use a dictionary for lemmatization. Non-dictionary approach should give an advantage that is important for processing texts of social networks: for the parser there is no fundamental difference between dictionary and OOV words. However, the presence of hallucinations, a serious negative consequence, was discovered (by hallucinations we mean cases when the parser generates lemmas that do not exist in the language). For this analysis, the ELMo model (trainable ELMo LSTM) was used as a compromise in terms of the quality and the parsing speed.

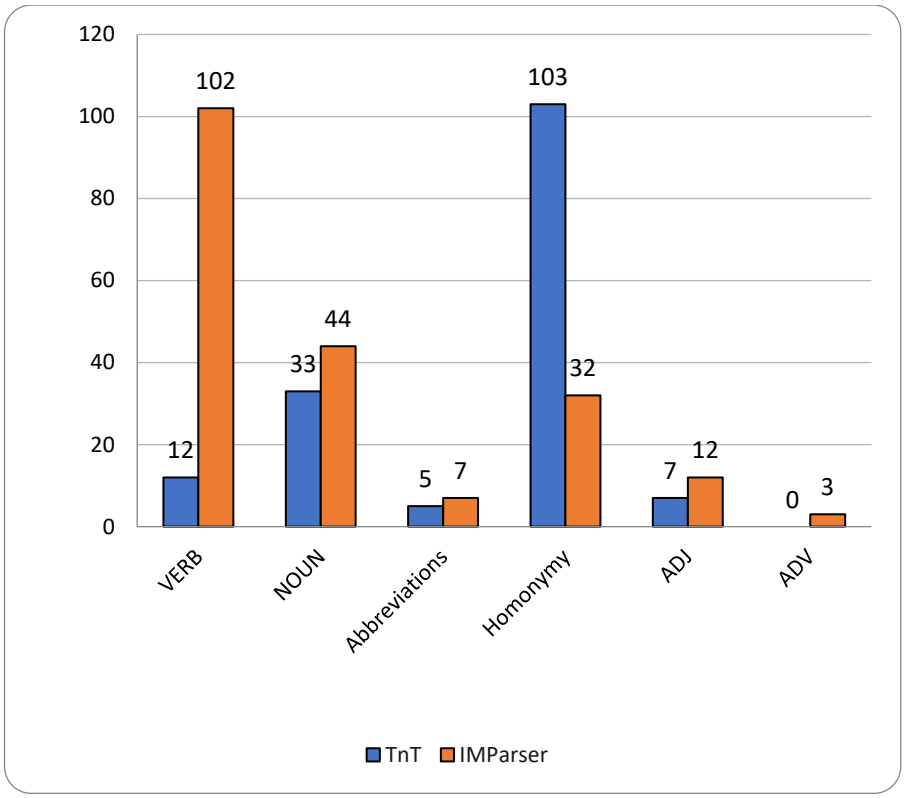

Fig. 1. Comparison of lemmatization errors per 10,000 tokens

In contrast to lemmatization, the IMParser copes with PoS-tagging better than the TnT parser (see Fig. 2). This is especially noticeable in nouns and adverbs tagging.

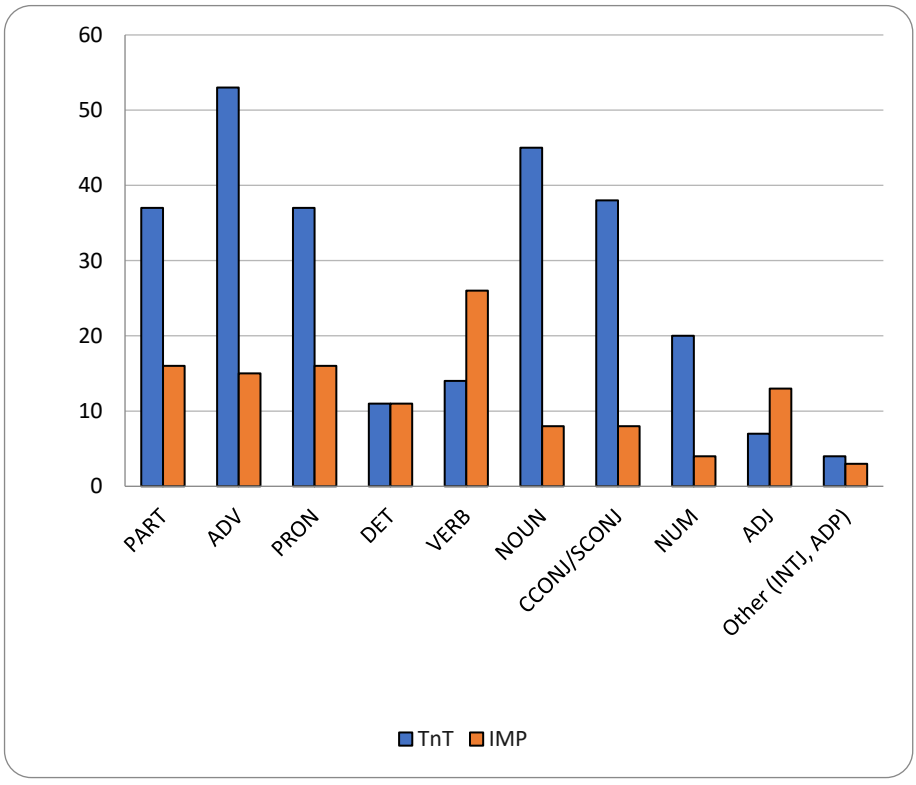

Fig. 2. PoS-tagging errors (tags show PoS that should be given instead of the wrong ones)

The quality of the IMParser for most of the other parameters (including parsing speed, PoS and feature tagging) turned out to be good, and this made it possible to use the parser for automatic parsing of the GICR.

2.3 Lemmatization problem

Since the quality of lemmatization shown by the IMParser was noticeably worse than the one performed by the TnT-parser, it was decided to concentrate on lemmatization tests, including the "problematic" parts of speech, i.e. verbs and nouns. To assess the quality of lemmatization, a dataset of 10 000 tokens was parsed, using BERT and ELMo models. As far as the data source, VKontakte as the "dirtiest" and the most difficult segment to parse was taken. All word forms with errors (typos, spelling errors) were excluded since GramEval did not presuppose spelling corrections. Lemmatization of nouns and verbs was assessed manually. For the results, see table 1.

	Verbs	Nouns
BERT	5.1	3.4
ELMo	7	5

Table 1. Lemmatization error rate, verbs and nouns, per 10,000 tokens

The results of quality evaluation of the IMParser revealed a serious problem with lemmatization: even when using the BERT model, which showed the best results in the competition, hallucination errors were found. Hallucination errors (see table 2) play a particularly important role because from a human point of view they are difficult to explain and predict comparing to disambiguation errors. Therefore, such errors may lead to users' mistrust in corpus annotation or incorrect data and statistics in studies if a linguist gives the annotation too much credence. We have also found that both models systematically miscalculate lemmas for word forms in uppercase ("ПЯТЫЙ" is defined as "пяты", and "РЕБЯТА" as "ребят"), although the parser was trained, among other things, on social network data where the uppercase is quite common. However, there are only a few errors in disambiguation.

Wordform	**Right lemma**	**BERT**	**ELMo**
потерь	потеря	потеь	потерьа
подсел	подсесть	подйти	подсеть
льдах	лед	льер	льд
прилечу	прилететь	прилестить	прилечуть
пою	петь	повать	поть
берите	брать	беыть	берить
бегите	бежать	бяться	бегять
шипящими	шипеть	шипить	шипть
стань	стать	стантъ	стть
зажгли	зажечь	зжечь	зажгть

Table 2. Examples of hallucination errors

If we compare the quality figures from the competition, parser results seem quite good (dev / test sets in lemmatization: 98.3% / 95.8% ELMo and 98.5% / 96.4% BERT, respectively). Nonetheless, the evaluation of the two most important parts of speech showed that it is impossible to use the results of the lemmatization in the corpus without a dictionary check. Perhaps a difference between the competition numbers and our noun-verb experiment arises from the fact that GramEval evaluation script took into account unchangeable word forms when lemmas always correspond to the original word (e.g. conjunctions, particles and prepositions), therefore, improving the lemmatization quality rate.

The IMParser doesn't use dictionary-based lemmatization as a matter of principle, moreover, it is pointed out that its approach to lemmatization, i.e. the compilation of rules for modifying word forms according to the training corpus (less than 1,000 classes of rules in total) and the application of these rules for test data lemmatization "is less likely to hallucinate an invalid lemma than in the sequence-to-sequence approach" [1]. Table 3 presents statistics on lemmas with hallucination errors for the same sample of nouns and verbs in 10,000 random tokens.

	Verbs	Nouns
BERT	49.3	61.67
ELMo	44.16	87.06

Table 3. Hallucination error rate to overall lemmatization error rate, verbs and nouns, per 10,000 tokens

2.4 Difficult disambiguation based on syntax

In order to assess how joint processing of morphology and syntax affects the quality of parsing (in particular, lemmatization), an experiment was carried out with full homonyms "пла́чу" and "плачу́", "сто́ит" and "стои́т".

In addition to these pairs, there was an attempt to experiment with the word form "лечу" (lemma "лететь"), but it turned out that there was no such word form in the training set at all, and the parser gave either "лечить" or hallucination errors in all the cases. This is a serious problem because the error found affects the language core (the verb "лететь" is part of the basic vocabulary and is more frequent than "лечить") and can provoke users' distrust of the corpus, while we strive to raise the level of confidence of linguistic users in the web corpus. Such errors should be excluded, for example, through the use of a dictionary. We think that an experiment with a large number of frequently used Russian verbs is needed in order to objectively assess the scale of the problem.

The form and grammatical features of "плачу́" and "пла́чу" verbs match fully, but "пла́чу" is an intransitive verb and cannot have a direct object, whereas the verb "плачу́" is a transitive one. Moreover, supposedly only the verb "плачу́" can have an argument with the preposition "за" (e.g., "плачу за обучение"). We assume that if the analysis of morphological and syntactic characteristics as well as lemmatization is processed simultaneously, the verb "плачу́" with a direct object or with a noun phrase with the preposition "за" is more likely to be lemmatized as "платить" than cases of "плачу́" without a direct object.

For the experiment, 221 sentences were selected from VKontakte and LiveJournal segments with the lemma "платить", 120 of them contain the word form "плачу́" with a direct object (DOBJ), and the remaining 101 without a direct object. In addition, in 98 sentences out of total 221 the verb "плачу́" has an argument with the preposition "за". The analysis was carried out both with BERT and ELMo models. The results are presented in table 4.

	BERT	**ELMO**
Average score (платить)	25.3	26.7
Платить + DOBJ	34.2	35
Платить – DOBJ	15	17
Платить + “за”	24.5	25.5
Платить – “за”	26	27.6

Table 4. Percentage of right lemmas of the verb “плачу́”

Having analyzed the statistics obtained, we can conclude that syntax in cases with direct object truly contributes to correct disambiguation. However, the preposition “за” does not affect the parsing. Moreover, as the anonymous reviewer rightly pointed out, the preposition “за” can actually occur in combination with the verb “плакать”:

*Каждый вечер **плачу за** тобой. Вернись быстрее ты домой.*
*Каждый вечер слушаю эту музыку и **плачу за** ним!*
*мне больно, и я **плачу за** тех кто живёт на донбассе.*

It was found that the parser is much more likely to lemmatize the verb as “плакать”, possibly due to a skew in the training dataset, but the statistics on the training set is as follows:

Lemma	**Total amount in train data**	**Word form**	**Amount of word forms in train data**
плакать	65	плачу (лемма: плакать)	4
платить	151	плачу (лемма: платить)	2

That is, although the number of the “плачу” option with the “плакать” lemma is formally twice as large, the absolute numbers are too small.

A similar case is represented by the words “сто́ит” and “стои́т”: in this form they differ only in stress, but for the first variant the lemma will be “стоить”, and for the second “стоять”. Also, the verb “стоить” is transitive, but the verb “стоять” is not. The following experiment was based on this difference.

For the experiment, 213 sentences were selected from the VKontakte segment with the “стоить” lemma, in 113 of them the word form “stand” with a direct object (DOBJ) occurs, in the remaining 100 the verb goes without a direct object. The results are presented in table 5.

	BERT	**ELMO**
Average score (стоить)	83.5	66.6
Стоить + DOBJ	89.3	78.7
Стоить – DOBJ	77	53

Table 5. Percentage of right lemmas of the verb “сто́ит”

In this case, it is obvious that the number of correct lemmas is higher for the direct object verb.

To sum up, we can say that syntax improves the quality of disambiguation, and it is worth noting that the cases selected for experiments are quite rare and complex.

2.5 PoS and grammatical disambiguation based on syntax

Homonymy, including PoS and grammatical one, often causes errors in automatic morphological parsing. It makes researchers pay special attention to this issue in studies related to automatic morphological labeling. According to the developers of GICR 1.0, the quality of the disambiguation of 'complicated' cases was 90% for adjectives and 68% for nominalized adjectives (nouns), as well as one of the worst indicators – 66% for accusative of animate nouns [12].

The purpose of this experiment was to test how well two models we are considering will distinguish between adjectives and substantives, as well as the coinciding forms of nouns in the nominative, genitive and accusative.

For the first experiment, the most frequent nouns derived from adjectives with no morphological transformation were extracted from "A New Frequency Dictionary of Russian" [9] (some words denoting abstract concepts were excluded, e.g., "основное", "главное", "целое"). In total, four words were selected:

- прошлое
- ученый
- русский
- больной

We used VKontakte and LiveJournal segments as a data source. 200 sentences were selected for each pair of words (100 sentences for a noun, 100 sentences for an adjective). These sentences were labelled manually according to the experiment task in such a way that if a word does not have a nominal head, then the word form gets the "noun" tag. Cases with a paired structure, where the ellipsis of the nominal head is obvious, were annotated as adjectives. Complicated cases with homonymy were tagged according to the semantics of a construction. The results of this experiment are presented in table 6.

	BERT	**ELMo**
Прошлое (NOUN)	95	99
Прошлый (ADJ)	97	95
Ученый (NOUN)	99	99
Ученый (ADJ)	88	80
Русский (NOUN)	85	78
Русский (ADJ)	100	96
Больной (NOUN)	97	94
Больной (ADJ)	77	63
Mean NOUN	94	92.5
Mean ADJ	90.5	83.5

Table 6. Adjectives and nominalized adjectives (nouns)

The quality of disambiguation in these cases are obviously higher than that of the TnT-parser (the accuracy of the latter is 68% for substantives on average, while IMParser shows 83% accuracy). Thus, we may expect the improvement in quality of such cases.

For the second experiment the following words were selected from the frequency dictionary with coinciding word forms in nominative and accusative:

- время
- дело
- жизнь
- слово
- место

The volume of the selected data, as in the previous experiment, was 100 sentences for each noun in the nominative and 100 sentences for each noun in the accusative; all data was reviewed and tagged manually. See the results in table 7.

Lexeme	**BERT**	**ELMo**
дело, Nom	100	100
дело, Acc	96	95
время, Nom	99	97
время, Acc	99	99
место, Nom	96	94
место, Acc	99	97
слово, Nom	95	97
слово, Acc	96	89
жизнь, Nom	98	96
жизнь, Acc	97	99
Mean, Nom	97.6	96.8
Mean, Acc	97.4	95.8

Table 7. Percentage of correct features, Nom and Acc

Finally, lexemes with matching accusative and genitive forms were selected:

- бог
- ребенок
- человек
- друг
- отец

The results of the experiment with the same volume of data are shown in table 8.

Lexeme	BERT	ELMo
бог, Acc	89	86
бог, Gen	97	98
ребенок, Acc	97	95
ребенок, Gen	99	100
человек, Acc	97	91
человек, Gen	100	99
друг, Acc	95	97
друг, Gen	98	98
отец, Acc	94	93
отец, Gen	100	100
Mean, Acc	94.4	92.4
Mean, Gen	98.8	99

Table 8. Percentage of correct features, Acc and Gen

Thus, we see that the correct scores for IMParser do not fall below 86 for a particular lexeme.

2.6 Lemmatization and PoS-tagging quality of out-of-vocabulary words

The aim of the next experiment was to check how well the models selected for the study deal with lemmatization and PoS-tagging of OOV words (neologisms, nonce words, slang, borrowings, etc.). For the experiment, 468 random lexemes with 639 occurrences were selected that were found neither in the Compreno dictionary nor in the training data, word forms with typos were removed. The results are presented in table 9.

	BERT	ELMo
Lemmatization	85.26	80.34
PoS-tagging	91.88	89.1
Different lemmas of the same lexemes (based on lemmatization of lexemes with several occurrences)	17.58	18.68

Table 9. Percentage of right lemmas of out-of-vocabulary occurrences

BERT and ELMo commit up to 15% and 20% of lemmatization errors respectively; but we should bear in mind that these are “complex”, non-dictionary words that are not found in the training corpus. It is difficult to establish correct lemmas for some of them (e.g., is the lexeme “друзьяшки” plurale tantum?). Unfortunately, there is no way to correct the lemmatization of neologisms and slang, so this percentage of accuracy is final.

2.7 Lemma corrections and grammatical system adaptation. Correction results

Since we came to the conclusion that the lemmatization quality of the selected models is not high enough, it was decided to improve the results. The possible way of correction is to use a dictionary. This hybrid method is not innovative, as well as not the only one available, but it has its advantages: firstly, it is a high speed of work, and secondly, predictability and verifiability of results.

Thus, we decided to use Compreno dictionary that contains more than 200,000 lemmas and more than 6.6 million word forms with PoS-tags and grammatical features. This dictionary was used to correct the morphology of GICR 1.0 (with the TnT parser) and helped to significantly improve the accuracy of disambiguation (by 30% for some categories of nouns).

We converted the dictionary to UD format; then word forms, PoS and grammatical features were checked: if all the three parameters matched but the lemma was different, then the lemma was replaced with the correct one taken from the dictionary.

	Verbs	**Nouns**
BERT	5.1	3.4
BERT (with Compreno)	0.7	2
ELMo	7	5
ELMo (with Compreno)	1.1	2.5

Table 10. Lemmatization error rate, verbs and nouns,
before and after using Compreno dictionary, per 10,000 tokens

This decision significantly improved the quality of lemmatization. It shows that a high-quality corpus should still be based on a dictionary with the inflection model; however, certain problems are still there:

- Homonyms with completely identical grammatical features cannot be resolved (e.g., “честный” and “честной” in oblique cases, “небо” and “нёбо” because of the letter “ё” as “е” is often replaced by “ё”);
- If the parser gave a false PoS or grammatical tag, then the lemma either will not be corrected, or it may be changed to a wrong one.

Although certain results have already been achieved in correcting lemmatization, some work still needs to be done. Moreover, there are no simple solutions to the above-mentioned problems. The importance of correct lemmas is obvious: corpus user should not think how to avoid errors of automatic lemmatization and compose corpus queries with disjunction of word forms but can safely use the lemma search.

3 GICR annotation within UD guidelines

The UD framework is an actively developing project with one of the best annotation formats. There were a lot of discussions around it and changes to it continue to be made today. We do not claim to change the UD Russian tagset as a whole, but we would like to slightly adjust the tagset that will be used in the GICR to simplify it for theoretical linguists who, unlike computational linguists who are actively using UD treebanks, have no experience with this tagset. At present, the situation with the UD format for the GICR is as follows.

The following changes have already been made:

- The PROPN tag, which denotes a proper name, has been replaced with NOUN (to avoid ambiguities in words like “Президент”, “Дед” and other non-proper names written with a capital letter);
- the particle “бы/б” would not be labelled as AUX, because it will probably seem strange to linguists dealing with the Russian language: the tag for this particle would be PART

To do:

- Although according to the terms of GramEval 2020 and the UD guidelines “Pass” tag (passive voice) should only be related to participles, 13% of “-ся/сь” verbs in standard training UD corpora of Russian possess the “Pass” tag (and not the “Middle” tag). It is due to the large amount of training data that is very difficult to verify manually. Because of this, the parser gets additional errors.
- A single tag is required for foreign words. At the moment, the annotation of foreign words is carried out in an ambiguous way: foreign words representing the names of large companies, cities, etc., are labelled as PROPN or NOUN, and all other foreign words are marked as X. The difference between NOUN and X is too subjective and therefore it is better to unify it in some way.

Furthermore, we have identified some features that may cause difficulties for linguists working with the Russian language. In particular, the following ones:

- transitivity / intransitivity of verbs in not labelled;
- there is no separate tag for “предикатив”;
- no tags for Plurale / Singulare Tantum in the category “Number”;
- Plurale Tantum nouns may have a genus, but the source of such information is unknown. (It is not discussed within Russian UD guidelines, but the training corpora contain this annotation. Thus, the parser also assigns gender to Plurale Tantum words).

The UD format has many advantages, including universality and readability. It covers more and more languages and treebanks. We tried to simplify its use for theoretical linguists a little in the GICR corpus without changing the general concept.

4 Conclusion

As a result of this study, the following conclusions were obtained:

- The importance of analyzing automatic markup errors from the standpoint of theoretical linguistics has been shown, including analysis of the annotation scheme;
- A number of problems have been identified that prevent the use of automatic markup for the needs of linguistic researchers without adjustments;
- The need for vocabulary support at some stage of the pipeline has been proven.

We believe that not only the national corpus, but also the web corpus should be treated as a serious source, so it should include:

A) a clear, human-readable annotation format, which the UD format successfully handles;
B) vocabulary support to ensure correct analysis of at least the language core;
C) a thorough analysis of the entire pipeline, taking into account the impact of segmentation and tokenization, which possess special features of social media texts, on the final annotation quality.

In the near future, based on the results of the research, the Compreno dictionary with UD format annotation and a sample of the GICR corpus annotation (silver standard) will be made available to the public.

Acknowledgements

This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, and by RFBR, grant № 17-29-09163.

References

[1] *Anastasyev D. G. (2020)*, Exploring Pretrained Models for Joint Morpho-Syntactic Parser of Russian, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020", Moscow, June 17–20, 2020.

[2] *Belikov V., Kopylov N., Piperski A., Selegey V., and Sharoff S. (2013)*, Corpus as language: from scalability to register variation, in Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.

[3] *Belikov V., Selegey V. and Sharoff S. (2014)*, Preliminary considerations towards developing the General Internet Corpus of Russian, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2012", Moscow, pp. 37-49.

[4] *Benko V. (2014)*, Aranea: Yet Another Family of (Comparable) Web Corpora, in Proceedings of Text, Speech and Dialogue, 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, 2014, pp. 257–264.

[5] *Benko V. and Zakharov V. P. (2016)*, Very Large Russian Corpora: New Opportunities and New Challenges, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow, June 1–4, 2016.

[6] *Brants T. (2000)*, TnT – A Statistical Part-of-Speech Tagger, in Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, April 29 – May 3, 2000, Seattle, WA.

[7] *Inshakova E. S, Iomdin L. L., Mitushin L. G., Sizov V. G., Frolova T. I., and Zinman L. L. (2019)*, SinTagRus segodnya (SinTagRus today), in Trudy Instituta Russkogo Yazyka im. V.V. Vinogradova, t. 21, Moscow, 2019, pp. 14-41.

[8] *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., and Suchomel V. (2013)*, The TenTen Corpus Family, in Proceedings of the 7th International Corpus Linguistics Conference, Lancaster, 2013, pp. 125–127.

[9] *Lyashevskaya O. N. and Sharoff S. A. (2009)*, A Frequency Dictionary of Russian. Access mode: http://dict.ruslang.ru/freq.php.

[10] *Lyashevskaya O. N., Shavrina T. O., Trofimov I. V., and Vlasova N. A. (2020)*, GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2020", Moscow.

[11] *de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006)*, Generating typed dependency parses from phrase structure parses, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).

[12] *Selegey D., Shavrina T., Selegey V., and Sharoff S. (2016)*, Automatic morphological tagging of Russian social media corpora: training and testing, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow.

[13] *Sharoff S. and Nivre J. (2011)*, The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2011", Bekasovo, pp. 591–605.

[14] *Shavrina T., Shapovalova O. (2017)*, To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser, in proc. of "CORPORA 2017", international conference, Saint-Petersbourg, 2017. P. 78-84

[15] *Shavrina T. O. (2017)*, Metody obnaruzhenia i ispravlenia opechatok: istoricheskiy obzor (Methods of mistypes detection and correction: a historical review), in Voprosy Yazykoznania, 2017, №4, pp. 115–134.

[16] *Helmut Schmid (1994)*, Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

[17] *Helmut Schmid and Florian Laws (2008)*, Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging, in COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK.