Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Disambiguation in context in the Russian National Corpus: 20 yeas later

Olga Lyashevskaya
HSE University
Vinogradov Russian Language Institute RAS
Moscow, Russia
olesar@yandex.ru

Ilia Afanasev
HSE University
MTS AI
Moscow, Russia
szrnamerg@gmail.com

Stefan Rebrikov
HSE University
Kurchatov Institute
Moscow, Russia
robstef85@gmail.com

Yana Shishkina
HSE University
Moscow Institute of Physics and Technology
Moscow, Russia
yanaalekseevna2000@mail.ru

Elena Suleymanova **Igor Trofimov** **Natalia Vlasova**
A. K. Ailamazyan Program Systems Institute of RAS
Pereslavl-Zalessky, Russia
yes2helen@gmail.com itrofimov@gmail.com nathalie.vlassova@gmail.com

Abstract

An updated annotation of the Main, Media, and some other corpora of the Russian National Corpus (RNC) features the part-of-speech and other morphological information, lemmas, dependency structures, and constituency types. Transformer-based architectures are used to resolve the homonymy in context according to a schema based on the manually disambiguated subcorpus of the Main corpus (morphology and lexicon) and UD-SynTagRus (syntax). The paper discusses the challenges in applying the models to texts of different registers, orthographies, and time periods, on the one hand, and making the new version convenient for users accustomed to the old search practices, on the other. The re-annotated corpus data form the basis for the enhancement of the RNC tools such as word and n-gram frequency lists, collocations, corpus comparison, and Word at a glance.

Keywords: morphological tagging; dependency parsing; lemmatization; disambiguation; NLP evaluation; Russian National Corpus; Russian

DOI: 10.28995/2075-7182-2023-22-307-318

Разрешение неоднозначности в контексте для Национального корпуса русского языка: 20 лет спустя

О. Н. Ляшевская[1,2], **И. А. Афанасьев**[1,3], **С. А. Ребриков**[1,4], **Я. А. Шишкина**[1,5], **Е. А. Сулейманова**[6], **И. В. Трофимов**[6], **Н. А. Власова**[6]

[1]Национальный исследовательский университет «Высшая школа экономики»
[2]Институт русского языка им. В. В. Виноградова РАН
[3]МТС ИИ
[4]НИЦ «Курчатовский институт»
[5]МФТИ
Москва, Россия
[6]Институт программных систем им. А. К. Айламазяна РАН
г. Переславль-Залесский, Ярославская обл., Россия
olesar@yandex.ru, {szrnamerg, robstef85}@gmail.com, yanaalekseevna2000@mail.ru,
{yes2helen, itrofimov, nathalie.vlassova}@gmail.com

Аннотация

Обновление разметки Основного, Газетного и ряда других корпусов Национального корпуса русского языка (НКРЯ) касается информации о части речи, других морфологических признаках, леммах (словарных формах слов), структурах зависимостей предложения и типах составляющих. Для разрешения лингвистической неоднозначности в контексте используются нейросетевые архитектуры на основе трансформеров. Разметка воспроизводит схему, применяемую в подкорпусе Основного корпуса со снятой вручную грамматической омонимией (морфология и леммы) и UD-SynTagRus (синтаксис). В статье рассматриваются проблемы применения моделей к текстам, написанным в различных функциональных стилях, орфографиях и в разные периоды времени. Поскольку в ряде случаев текстовому фрагменту в заданном контексте можно сопоставить более одного теоретически возможного лингвистического разбора, необходимо принимать во внимание поддержку множественных разборов. Кроме того, обсуждаются вопросы совместимости старой и новой разметки в плане адаптации пользователей к новому поисковому функционалу корпуса. Автоматически дизамбигуированные данные больших корпусов позволили улучшить существующие и разработать новые сервисы поисковой платформы НКРЯ, такие как частотные списки слов и n-грамм, коллокации, сравнение корпусов и портрет слова.

Ключевые слова: автоматическое разрешение лексико-грамматической неоднозначности, морфологическая разметка, синтаксическая разметка, русский язык, Национальный корпус русского языка

1 Introduction

For almost 20 years, the lexico-grammatical annotation of the Russian National Corpus (RNC) existed in three formats. (1) In the Syntactic corpus (SynTagRus, 1.4 MW), each word was provided with one and only one morphological and lemma analysis appropriate in context, and each sentence was analysed as one syntactic dependency tree. (2) In the the manually disambiguated subcorpus of the Main corpus ("Snyatnik", 6 MW) and in the Educational corpus (0,6 MW), only morphology and lemmas were analysed based on a somewhat different tagset and grammatical dictionary compared to SynTagRus. The majority of historical RNC corpora were annotated generally in the same way and oriented on their own markup schemas, tagsets, and dictionaries. (3) Finally, there were no disambiguation in the largest part of the modern Russian texts (more than 1 billion words) and Church Slavonic texts (5,3 MW): each word corresponded to as many analyses as the grammatical dictionary stores, regardless of the context. If the word form of a modern language is not attested in the dictionary, the MyStem hypothesis module assigns a few of the most probable annotations to it (Segalovich, 2003; Zobnin and Nosyrev, 2015).

One of the objectives of the Corpus 2.0 project (2020-2022) was to add syntactic annotations and resolve lexical and morphological ambiguity in modern Russian texts. Firstly, this allows users to constraint the search window by defining syntactic relations between elements or setting up a certain type of clause or phrase within which the elements should occur. Secondly, this makes it possible to significantly reduce the number of irrelevant examples in the search output. Thirdly, other search facilities such as lexical groups-based search, frequency lists, collocations, associated words, etc. definitely benefit from the less noisy annotation input. Fourthly, the use of syntactic n-grams based on dependency parses (Goldberg and Orwant, 2013) in addition to ordinary sequential n-gram opens the way to a new kind of high-quality tools for researchers. All these changes also involve technical improvements in the infrastructure of the corpus search engine such as reducing the size of the search indices and the time spent performing the calculations, extending the amount of annotated data and information conveyed to the user.

2 Related Work

The approaches to the three grammar tasks that form the basic NLP pipeline, namely, part-of-speech/morphological tagging, lemmatisation, and dependency parsing, rapidly developed for the last half a century (Hann, 1974) (Spyns, 1996) (Aduriz et al., 1996) (Branco and Silva, 2003)] (Qi et al., 2020) (Kumar et al., 2022). Currently pipeline models that combine part-of-speech/morphological tagging, lemmatisation , and parsing, dominate the landscape (Straka and Straková, 2017) (Kondratyuk, 2019) (Kanerva et al., 2021). However, despite this pursuit to develop the language-independent tagger for benchmark datasets (Toleu et al., 2022) that provide satisfying for all the included languages, yet moderate for each of them results, there is a growing concern that low-resourced language NLP, and

probably NLP in general, is going to suffer from the trend (Alonso-Alonso et al., 2022). Frw works clearly state the intention to make a universal tagger, which is based upon the multi-lingual training and switching parameters to fine-tune for a single language (Üstün et al., 2020). The models, trained for the particular task-language pair, still seem to deserve attention, as (Dyer, 2022) states for the case of Wolof language.

Automatic morphological tagging systems currently employ the pair of dominating approaches, the single-language rule-based one (Gambäck, 2012), and the machine learning-based one, which can assume both monolingual (Berdičevskis et al., 2016) (Qi et al., 2018) (Qi et al., 2020) (Scherrer, 2021) and multi-lingual (Straka and Straková, 2017) forms. Instead of targeting the multi-lingual level, now morphological tagging shifts into the multi-lect one to be able to deal with the very close (Obeid et al., 2022), yet significantly different lects, as is the cast with Arabic (Inoue et al., 2022) (Fashwan and Alansary, 2022). This also provokes a lot of discussion for morphological tagging of low-resouced languages (Blum, 2022) (Wiemerslage et al., 2022). The discussion about data quality takes place within the common morphology tagging discourse (Muradoglu and Hulden, 2022). New methods are being developed, for instance, graph-based part-of-speech tagging (ImaniGooghari et al., 2022), or using compressed FastText models (Nevěřilová, 2022). Specifically concerning Russian, joined morphological analysis and morpheme segmentation models were proposed recently (Bolshakova and Sapin, 2022).

Lemmatisation follows the same patterns that morphological tagging does. Currently, there is a division between the universal lemmatisation tools (Straka and Straková, 2017) (Bergmanis and Goldwater, 2018) (Kanerva et al., 2021), and language, or domain-specific (Fernández, 2020) The sequence-to-sequence architecture (Sutskever et al., 2014) (Cho et al., 2014) prevails now, and within it the encoder-decoder transformers dominate (Lewis et al., 2020) The ensemble models that enhance lemmatisation efficiency with external resources (Milintsevich and Sirts, 2021) are gaining popularity (de Graaf et al., 2022)

Dependency parsing is probably the most dynamically developing area of the three, as it still presents the highest challenge of the three for the automated corpus tools. New methods are constantly being implemented: the last three years witnessed a combination of the second-order graph-based and headed-span-based projective dependency parsing (Yang and Tu, 2022), the domain adaptation (Li et al., 2022) and the dependency parsing being treated as machine reading comprehension (MRC)-based span-span prediction (Gan et al., 2022) and using structure preserving embeddings for dependency parsing (Kádár et al., 2021) The state-of-the-art method, biaffine parsing, is modified (Xu et al., 2022). The previously under-utilised concepts, such as *nuclei* (semantically independent units consisting of a content word together with its grammatical markers, regardless of whether the latter are realised in dependent words or not (Basirat and Nivre, 2021)), are introduced to the frameworks. The data augmentation techniques are implemented to enhance the performance of the models (Goodwin et al., 2022). (Eggleston and O'Connor, 2022) and (Langedijk et al., 2022) introduce cross-lect dependency parsing, getting in line with papers that consider low-resourced languages (Tian et al., 2022) and zero-shot (de Lhoneux et al., 2022) (Shi et al., 2022) dependency parsing. The issues of the dataset construction that affect evaluation are discussed in (Krasner et al., 2022) Artificial performance inflation is a problem that should be addressed across the pipeline of morphological tagging, lemmatisation and part-of-speech tagging (Goldman et al., 2022).

3 Data for Training and Evaluation

We conducted experiments involving a diverse panel of text samples. A variety of genres, types, domains, time periods of creation, and orthographies were presented in the following datasets for modern Russian (1700-2020s):

- `SynTagRus UD 2.8` - 1,1 M tokens (contemporary fiction, popular science, newspaper and journal articles dated between 1960 and 2016, texts of online news etc.). This portion of the RNC Syntactic Corpus converted to the Universal Dependencies (UD) format was the main training dataset used in the GramEval-2020 shared task.
- `SynTagRus UD 2015` - 400k tokens. An addition to the RNC Syntactic Corpus annotated in 2015-

GramEval-2020 (Taiga)	dev	test	New RNC datasets	dev	test
fiction	1.0k	1.0k	prose-XX	10.4k	20.0 k
news	1.0k	1.0k	newspapers-XXI	7.8k	14.4k
poetry	1.0k	1.0k	prose-XIX	41.7k	80.7k
social	1.0k	1.0k	poetry-XIX	1.4k	1.4k
wiki	1.0k	1.0k	old-orthography	14.8k	14.8k
			old-orthography-XVIII	6.1k	6.1k
			Middle Russian: LEG	16.5k	39.0k
			bezobrazov		519.0k

Table 1: Size of the validation and test sets, tokens.

2020; converted and added to UD v.2.9. New genres: wikipedia.

- Taiga - 200 k tokens. Modern text samples extracted from Taiga Corpus, MorphoRuEval-2017 and GramEval-2020 shared tasks collections. Genres include electronic communication (VK, Twitter and other social media, YouTube comments, questions & answers from otvet.mail.ru, reviews from reviews.yandex.ru); poetry from stihi.ru (naïve poetry) and RNC Corpus of Russian poetry; fiction; news (lenta.ru etc.); wiki (Russian wikipedia). Taiga includes, among others, development and test data of the GramEval-2020 shared task (modern Russian), which was subdivided into the following subsets: fiction, news, poetry, social, wiki.
- newspapers-XXI - 34 k tokens. Samples extracted from the RNC National media and Regional and international media corpora.
- prose-XX - 423 k tokens. Texts of the 20th c. and the beginning of the 21th c. in modern orthography (RNC Main corpus). Fiction includes stories by V. M. Shukshin, I. V. Evdokimov, and M. K. Pervukhin, non-fiction - diaries and memories, journalism covers general news, finance, church news, recipes and tips.
- prose-XIX - 108 k tokens. Texts of the 19th c. in modern orthography (RNC Main corpus). The dataset includes drama by A. V. Sukhovo-Kobylin, A. Pisemsky, M. Gorky, etc., fiction by N. V. Gogol, S. T. Aksakov, E. A. Salias etc., non-fiction on history, hygiene, memories and essays.
- poetry-XIX - 50 k tokens. Samples from the RNC Russian Poetry Corpus written before 1917 and provided in modern orthography.
- old-orthography - 108 k tokens. Texts of the 19th - early 20th cc. in pre-revolutionary orthography (S. T. Aksakov, P. A. Kulish, M. Pogodin, A. Spaso-Kukotsky, N. I. Grech)
- old-orthography-XVIII - 6 k tokens. 18th century texts in old orthography (by Peter the Great, S. Pufendorf, P. I. Pogoretsky, F. A. Emin)

As for historical Russian data (1400-1700s), we used official legal and business writing texts, as the other RNC Middle Russian collections, like vernacular gramotki, were distinctly different in the occurrences of old grammatical forms and constructions, in phonetic features reflected in orthography, and in genre-specific lexical distributions. We split the taken texts into two datasets:

- LEG(acy) texts written in 15th – 17th cc. (ca. 1.1 M tokens), and
- Bezobrazov - recently added to the RNC texts of the latter half of the 17th c. from Bezobrazov's archive (500 k tokens).

Table 1 summarises the size of the development and test data used in experiments. In the experiments reported below, the models were trained on a joined modern Russian training dataset (1700-2020s) or historical Russian data (1400-1700s).

All data are presented in the CONLL-U format and annotated according to the Russian UD-Ext scheme (Lyashevskaya, 2019). This scheme assumes the use of a standard inventory of the UD-Russian dependency relations and common RNC and UD policy for lemmatisation. Enhanced dependency relations are not provided. To make morphological annotations of the RNC Main corpus and Russian UD compatible,

the following features are added to the GramEval2020 and SynTagRus data and used in all new datasets:

- parts of speech: PRED for predicatives (eg. *можно, холодно, жаль*), ADVPRO for pronominal adverbs (eg. *тут*), PREDPRO for pronominal predicatives (eg. *некого*), PARENTH for parentheticals (eg. *конечно*), ANUM for ordinal numerals (eg. *второй*).
- grammatical features: Transit={Tran,Intr} for transitivity, Case={Acc2,Loc2} for secondary cases, Degree=Cmp2 for comparatives with the prefix *по-*, Anom=Yes for anomalous forms.

PoS-tags that are absent from the UD format were added by automatic replacement with the use of wordlists. Some PoS-tags were added manually, e.g. ANUM for numerals written with numbers, PRED for ambiguous words. PoS-tag disambiguation (e.g. *холодно* - ADV vs. ADJ vs. PRED; мало NUM vs. ADVPRO vs. PRED) and corresponding correction of dependency relations were performed manually. Necessary grammatical features were corrected or added using the wordlists and lists of tokens with manual correction. The transitivity feature was manually checked in context with the dependency relations correction.

4 Rubic: a Model for Tagging and Parsing

The study is divided into the following parts. In the first one we examine the previous results of the GramEval-2020 shared task. From this data, we form our expectations for the next suitable model to achieve in morphological tagging, lemmatisation, and dependency parsing. The second stage of the research is the description of the new model, and its results on the GramEval data. In some tasks, the model is challenged by the other models, specifically trained for this task on the particular dataset, to explore the possible enhancements. The third part of the study is dedicated to the analysis of the key errata that the proposed model makes, and whether the other models struggle with the same issues.

The model that we are starting with, our baseline, is the one that has been previously used for the annotation of the RNC corpus data, qbic (Anastasyev, 2020), a winner of the GramEval-2020 shared task. Qbic is a RuBERT encoder accompanied by three classifier decoders performing the part-of-speech classification, lemmatisation, and dependency parsing, respectively. Lemmatisation is conducted in two stages, with the classifier assigning the particular rule to a token, after which the rules themselves are applied. Each lemmatisation rule specifies the number of characters to be cut and a combination of characters to be added, thus comprising a total of 1000 to 2000 rules, depending on the amount of training data (cf. also “less than 1,000 classes of rules in total” in (Michurina et al., 2021)). The rules form in the following manner:

- Training set yields sequences of transformations that are required to transform a token into its lemma (delete postfix/suffix of a certain length > add some sequence of characters to the end > capitalise/decapitalise)
- We take the sequences of transformations that are met more than 3 times (to exclude noise)
- The remaining sequences become rules

Table 2 shows the performance of qbic on the re-annotated GramEval-2020 datasets. A standard CONLL18 script was used to calculate accuracy scores for parts of speech (PoS), morphological features, lemmas, and labeled attachment score for syntactic dependencies (LAS, basic relation inventory, ie. `nummod` and `nummod:gov` are considered the same). The model performed in a satisfactory way in most of the aspects. However, its performance on dependency parsing was below expectations. Non-standard patterns in poetry, social media texts, and wiki presented an especially hard challenge for it. Additionally, qbic was not robust in full morphological tagging and lemmatisation in the case of social media, poetry, diaries, and encyclopedic texts, which contain abbreviations, non-standard punctuation, transcript notes, rare named entities, and especially in the case of the RNC subcorpus of older orthographies (ca. 13M tokens).

To meet this challenge, we present Rubic, a model that utilises the same architecture as qbic, with enhancements, see Figure 1. For an encoder, we use sberbank-ai/ruBert pretrained on 30 GB data. In our model, the lemmatisation module receives additional information from the part-of-speech tagging classifier. Rubic checks lemma candidates against a supplementary dictionary compiled manually. The dictionary is a pair of lemma and part of speech, split by tab, e.g. *автоматизм NOUN*. Besides that,

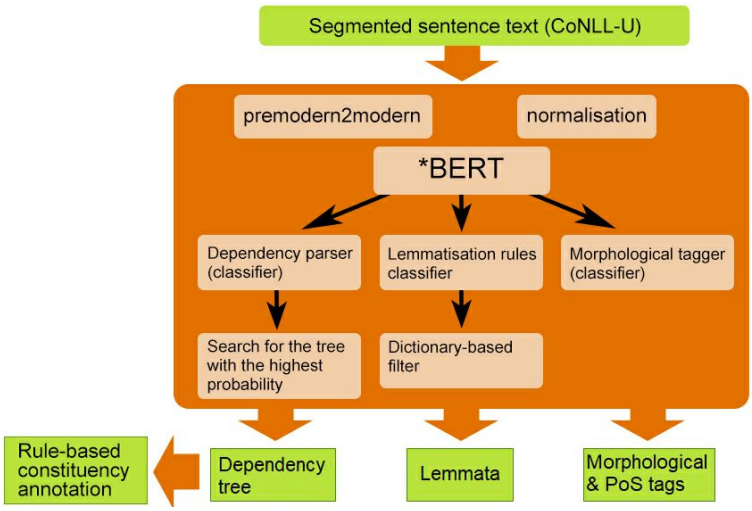

Figure 1: Key principles and architecture of Rubic.

Dataset	fiction	news	poetry	social	wiki
PoS	98.0	96.6	96.9	94.7	92.7
Morph.features	98.7	96.1	96.7	94.7	94.4
Lemmatisation	98.0	98.2	95.3	96.0	93.6
LAS	89.6	91.2	81.4	80.7	78.1

Table 2: Accuracy score of qbic on GramEval-2020 dataset, %

the symbol sequences unlikely to occur in Russian texts are preprocessed. We specifically set up Rubic to process data with non-standard orthography by implementing a graphic premodern2modern heuristic, and mapping the tokens in older orthography to tokens in modern orthography.

We perform data augmentation when training Rubic. We use the calculation of "the lexical usefulness weight" that prioritise the use of rare tokens for the further pipeline of data augmentation. If a sentence contains two, and exactly two quotation marks, we add another sentence to the dataset, that contains guillemets instead (we add 450 sentences via this heuristic). We use the heuristic of jo-fication, transforming e into $\ddot{e}$, in words, where it is possible (we add more than 800 sentences via this heuristic). We use the capitalisation heuristic, when the tokens are randomly capitalised for the purposes of better recognition (we acquire nearly 2000 additional sentences via this heuristic. %; we take only 20% of the sentences, generated by the previous heuristic).

With all these enhancements, the results of the model expectedly grow. We provide the difference between accuracy scores in Table 3. Rubic improves in parsing, and some improvements can be seen in tagging and lemmatisation. It underperforms on the fiction dataset, and wiki morphology presents it with some challenges. All this may also signal about overfitting, so we use the other datasets of the modern Russian language: CONLL18, and IWPT21. The results are presented in Table 4.

We also evaluated Rubic on the RNC test sets prepared specifically for the task of full corpus re-annotation. The results are shown in Table 5. In all datasets, Rubic performs well on major and most frequent part of speech categories such as verbs, nouns, proper nouns, prepositions, and coordinate conjunctions. Noun case accuracy is above 98% in all datasets except poetry and old orthography-XVIII. Mixing adjectives vs. participles, adjectives vs. adverbs is higher in the latter datasets and Taiga. Annotation of predicatives and corresponding syntactic structures is problematic in poetry, fiction and non-fiction written in the 20th c. and earlier, in which a wider variety of constructions and lexical fillers is available. Expectedly, parsing quality drops on longer sentences, and non-standard symbols, non-

Dataset	fiction	news	poetry	social	wiki
PoS	+0.1	+1.4	+1.7	+1.0	+0.5
Morph.features	-0.1	+0.3	+0.1	+0.6	-0.4
Lemmatisation	-0.3	+0.0	+0.2	+0.6	+0.5
LAS	+0.5	+0.8	+1.3	+0.3	+2.8

Table 3: Change in accuracy score for Rubic compared to qbic, %, GramEval-2020 datasets

Dataset	CONLL18	IWPT21
PoS	99.23	99.14
Morph.features	98.27	98.19
Lemmatisation	97.49	97.83
LAS	95.51	95.47

Table 4: Accuracy score of Rubic on standard modern Russian datasets, %

standard place of punctuation marks and other non-letters, and out-of-vocabulary abbreviations misleads the model.

5 Lemmatisation: Further Experiments

Rubic, thus, does not overfit for GramEval-2020 datasets. However, we wanted to see if there is a possibility to enhance its performance. To test this, we picked the lemmatisation task and trained two BART-large-based lemmatiser models (Lewis et al., 2020). This is a sequence-to-sequence state-of-the-art multilingual method that can help to reveal critical points in which Rubic needs enhancement.

The comparison is based on the following data: modern RNC datasets, historical LEG and Bezobrazov datasets. Both Rubic and BART-large were separately fine-tuned for modern and historical data. The results of comparison between BART-large and Rubic are in Table 6.

The news dataset witnesses a better performance of Rubic, by 0.1 per cent: the Rubic heuristics adapt the model for the specific language variety. However, it seems that the texts of the Middle Russian period require much more intricate heuristics, which leads to the striking 12 to 20, depending on data quality, per cent difference between BART-large and Rubic accuracy in favour of the former. Overall, BART-large beats Rubic by a significant margin of 0.4 to 3 per cent. The main challenges are non-standard orthography and syntactic structures of XIX century poetry, which encourage a more generalising approach of BART-large.

The Rubic model, despite implemented heuristics, is challenged by two main classes of words: non-productive verb models (*скорбатъ* instead of *скорбетъ* 'mourn'), and proper names (*Любовя* instead of *Любовъ* 'Lyubov'). The non-standard modern orthography also takes its toll: *наср@ла* is returned instead of *насратъ* 'do not give a damn about smth' likely due to the special symbol that was not normalised. Sometimes model generates empty lemmata, due to the rule-based nature of its lemmatiser module.

BART-large sequence-to-sequence architecture helps to deal with the aforementioned problems. It still overgeneralises, creating the syntagmae, similar to *-жо-* in verbs (*ожотъся* instead of *ожечься* 'get fired by'), or choosing the more general ending, completely confusing the word class, cf. *Стоцка* instead of *Стоцкая* 'Stotskaja'. Generalisation also leads into the model being unable to deal with orthography issues (odd *с* in *естесственный* 'natural'; odd *о* in *-пр-*, cf. *предупорежденне* instead of *предупреждение* 'warning'). Probably, the same factor leads to the appearance of hyphens in lemmas for the words that were transitioned from string to string somewhere in the data, sometimes with character replacing, for instance, in *пеп-льница* instead of *пепельница* 'ashpot'. Compound pronouns, such as *ни о чём* 'about nothing', often lose their negative particle (*ни*) part. The words that contain similar

Dataset	Taiga	newspa-pers-XXI	prose -XX	prose -XIX	poetry -XIX	old ortho-graphy	old ortho-graphy-XVIII
PoS	97.8	99.0	98.9	99.2	97.4	98.9	95.8
Morph.features	94.6	97.3	97.2	97.7	94.2	95.9	90.1
Lemmatisation	97.6	99.1	98.3	98.9	95.9	97.5	93.7
LAS	85.7	95.1	94.1	94.6	85.6	94.0	83.7

Table 5: The accuracy score of Rubic on RNC datasets, %

Dataset	Rubic, accuracy, %	BART-large, accuracy, %
Taiga	97.6	**98.0**
newspapers-XXI	**99.1**	99.0
prose-XX	98.3	**98.7**
prose-XIX	98.9	**99.3**
poetry-XIX	95.9	**98.9**
old orthography	97.4	**98.7**
old orthography-XVIII	93.7	**93.8**
LEG(al) test, 1400-1700	85.4	**98.0**
Bezobrazov	73.8	**85.0 (92.6 with normalisation)**

Table 6: Lemmatisation accuracy scores for Rubic and BART-large models on RNC datasets. The best results are highlighted in bold.

syllables, such as *царица* 'empress', are often reduced to a single syllable, in this case, *ца*: probably, the original BART-large dataset was trained to eliminate reduplication. The model clearly lacks knowledge of how the lemmas in particular language should look, which leads to generating adjective lemmas that after the adjectival affix *-ск-* have *-ив-* instead of *-ий-*. The model often does not pay attention to the morphology tagging (generated verb lemmas with *Aspect=Perf* tag often contain *-ывать*, which is a strong marker of continuous aspect in Russian verbs; prefix *по-* for *Degree=Cmp2* adjectives generated lemmas).

BART-large experiments show that sequence-to-sequence is not a necessarily ideal solution. It appears to be slow when annotating large amount of texts. However, this method reveals room for improvement of models like Rubic, particularly when it concerns the dataset construction, non-standard orthography, and low-productive paradigms, such as proper names and some verb classes. We are going to dedicate further research to these particular issues.

6 Corpus annotation and future development

At the moment, Main corpus, Regional Media, and Educational corpora are annotated by Rubic. In order to make it easier for users to switch from the old version to the new one, two lemma layers – annotations provided by Mystem and Rubic – are searchable. By default, the search is conducted on the layer automatically disambiguated by Rubic only.

We decided to apply three techniques to improve the Rubic outcome. Firstly, although the neural model is set up to produce only one analysis per token, in the case of theoretically plausible equivalent linguistic interpretations (eg. adjective vs. participle, see the practice of the manually disambiguated RNC subcorpus) additional morphological and lexical analyses were provided by rules. Secondly, lemmas that occur 30 times and more in the corpus and are not found in the Mystem dictionary, were checked and corrected manually. Thirdly, a number of heuristics were applied to the dependency annotations to provide search by constituency types and unlabeled tree configurations (eg. search within subordinate clauses; within participial phrases; search words that do not have dependents).

In the future, based on the results of the users' feedback, more disambiguated RNC corpora will be made available, with necessary adjustments in the annotation methods. RNC services such as frequency lists, graphs by year, lemma-based corpus portraits and comparison, collocation tools, Word at a glance sketch tool, and search by lexico-semantic features, depend critically on the quality of data lemmatisation. More work should be done in terms of finding new text classes on which the models underperform and adding relevant excepts to training; balancing the training collection by text types; balancing learning rate for different task. Decoding of abbreviated words is likely to be formulated as a separate since the distribution of such forms in large corpora cannot be modeled in the same way as lemmatisation rules.

The project's repository containing supplementary materials is available at: `https://github.com/olesar/RNC2.0`.

Acknowledgements

This work was carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793: "Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research".

References

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, and Ruben Urizar. 1996. Euslem: A lemmatiser/tagger for basque. // Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, and Catalina Röjder Papmehl, *Proceedings of the 7th EURALEX International Congress*, P 27–35, Göteborg, Sweden, aug. Novum Grafiska AB.

Iago Alonso-Alonso, David Vilares, and Carlos Gómez-Rodríguez. 2022. The fragility of multi-treebank parsing evaluation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5345–5359, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Daniil Anastasyev. 2020. Exploring pretrained models for joint morphosyntactic parsing of Russian. // *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*, volume 19, P 1–12.

Ali Basirat and Joakim Nivre. 2021. Syntactic nuclei in dependency parsing – a multilingual exploration. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1376–1387, Online, April. Association for Computational Linguistics.

Aleksandrs Berdičevskis, Hanna Eckhoff, and Tatiana Gavrilova. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. // *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog»*, P 99–111, Moscow, Russia. RSSU.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1391–1400, New Orleans, Louisiana, June. Association for Computational Linguistics.

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, P 1–9, Dublin, Ireland, May. Association for Computational Linguistics.

Elena I Bolshakova and Alexander S Sapin. 2022. Building a combined morphological model for Russian word forms. // *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 45–55. Springer.

Antònio Branco and Jõao Silva. 2003. Portuguese specific issues in the rapid development of state of the art taggers. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 7–9, Paris. European Language Resources Association.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. // *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. // *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, P 5334–5344, Marseille, France, June. European Language Resources Association.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 578–587, Dublin, Ireland, May. Association for Computational Linguistics.

Bill Dyer. 2022. New syntactic insights for automated Wolof Universal Dependency parsing. // *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, P 5–12, Dublin, Ireland, May. Association for Computational Linguistics.

Chloe Eggleston and Brendan O'Connor. 2022. Cross-dialect social media dependency parsing for social scientific entity attribute analysis. // *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, P 38–50, Gyeongju, Republic of Korea, October. Association for Computational Linguistics.

Amany Fashwan and Sameh Alansary. 2022. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for Egyptian Arabic. // *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, P 142–160, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Laura García Fernández. 2020. A contribution to old english lexicography. *NOWELE / North-Western European Language Evolution*, 73(2):236–251.

Björn Gambäck. 2012. Tagging and verifying an amharic news corpus. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 79–84, Paris. European Language Resources Association.

Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. 2022. Dependency parsing as MRC-based span-span prediction. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2427–2437, Dublin, Ireland, May. Association for Computational Linguistics.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. // *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, P 241–247.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 864–870, Dublin, Ireland, May. Association for Computational Linguistics.

Emily Goodwin, Siva Reddy, Timothy O'Donnell, and Dzmitry Bahdanau. 2022. Compositional generalization in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6482–6493, Dublin, Ireland, May. Association for Computational Linguistics.

Michael Hann. 1974. Principles of automatic lemmatisation. *ITL Review of Applied Linguistics*, 23(1):3–22.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 1577–1589, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for arabic and its dialects. // *Proceedings of the Findings of the Association for Computational Linguistics: ACL2022*, Dublin, Ireland, May. Association for Computational Linguistics.

Ákos Kádár, Lan Xiao, Mete Kemertas, Federico Fancellu, Allan Jepson, and Afsaneh Fazly. 2021. Dependency parsing with structure preserving embeddings. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1684–1697, Online, April. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. // *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 12–18, Florence, Italy, August. Association for Computational Linguistics.

Nathaniel Krasner, Miriam Wanner, and Antonios Anastasopoulos. 2022. Revisiting the effects of leakage on dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 2925–2934, Dublin, Ireland, May. Association for Computational Linguistics.

C S Ayush Kumar, Advaith Maharana, Srinath Murali, Premjith B, and Soman Kp. 2022. BERT-based sequence labelling approach for dependency parsing in Tamil. // *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, P 1–8, Dublin, Ireland, May. Association for Computational Linguistics.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 8503–8520, Dublin, Ireland, May. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 7871–7880, Online, July. Association for Computational Linguistics.

Ying Li, Shuaike Li, and Min Zhang. 2022. Semi-supervised domain adaptation for dependency parsing with dynamic matching network. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1035–1045, Dublin, Ireland, May. Association for Computational Linguistics.

Olga Lyashevskaya. 2019. A reusable tagset for the morphologically rich language in change: A case of Middle Russian. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 422–434.

Mariia Michurina, Alexandra Ivoylova, Nikolay Kopylov, and Daniil Selegey. 2021. Morphological annotation of social media corpora with reference to its reliability for linguistic research. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 492–504.

Kirill Milintsevich and Kairit Sirts. 2021. Enhancing sequence-to-sequence neural lemmatization with external resources. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 3112–3122, Online, April. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 7294–7303, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Zuzana Nevěřilová. 2022. Compressed FastText Models for Czech Tagger. // *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*, P 79–87, Tribun EU. European Language Resources Association.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. // *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 319–326, Abu Dhabi, UAE, December. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. // *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 101–108, Online, July. Association for Computational Linguistics.

Yves Scherrer, 2021. *Adaptation of Morphosyntactic Taggers*, P 138–166. Studies in Natural Language Processing. Cambridge University Press.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // *MLMTA*, P 273–280, 01.

Freda Shi, Kevin Gimpel, and Karen Livescu. 2022. Substructure distribution projection for zero-shot cross-lingual dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6547–6563, Dublin, Ireland, May. Association for Computational Linguistics.

Peter Spyns. 1996. A tagger/lemmatiser for Dutch medical language. // *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, P 1147–1150, USA. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. // *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. // *Advances in neural information processing systems*, P 3104–3112.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Enhancing structure-aware encoder with extremely limited data for graph-based dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5438–5449, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Alymzhan Toleu, Gulmira Tolegen, and Rustam Mussabayev. 2022. Language-independent approach for morphological disambiguation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5288–5297, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 2302–2315, Online, November. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 988–1007. Association for Computational Linguistics.

Ziyao Xu, Houfeng Wang, and Bingdong Wang. 2022. Multi-layer pseudo-Siamese biaffine model for dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5476–5487, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Songlin Yang and Kewei Tu. 2022. Combining (second-order) graph-based and headed-span-based projective dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 1428–1434, Dublin, Ireland, May. Association for Computational Linguistics.

AI Zobnin and GV Nosyrev. 2015. Morfologicheskij analizator MyStem 3.0. *Trudy Instituta russkogo yazyka im. VV Vinogradova*, 6:300–310.