Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

GRAMEVAL 2020 SHARED TASK: RUSSIAN FULL MORPHOLOGY AND UNIVERSAL DEPENDENCIES PARSING[1]

Lyashevskaya O. N. (olesar@yandex.ru)

National Research University Higher School of Economics;
V. V. Vinogradov Russian Language Institute of RAS,
Moscow, Russia

Shavrina T. O. (rybolos@gmail.com)

National Research University Higher School of Economics;
Sberbank, Moscow, Russia

Trofimov I. V. (itrofimov@gmail.com),
Vlasova N. A. (nathalie.vlassova@gmail.com)

A. K. Ailamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia

The paper presents the results of GramEval 2020, a shared task on Russian morphological and syntactic processing. The objective is to process Russian texts starting from provided tokens to parts of speech (pos), grammatical features, lemmas, and labeled dependency trees. To encourage the multi-domain processing, five genres of Modern Russian are selected as test data: news, social media and electronic communication, wiki-texts, fiction, poetry; Middle Russian texts are used as the sixth test set. The data annotation follows the Universal Dependencies scheme. Unlike in many similar tasks, the collection of existing resources, the annotation of which is not perfectly harmonized, is provided for training, so the variability in annotations is a further source of difficulties. The main metric is the average accuracy of pos, features, and lemma tagging, and LAS.

In this report, the organizers of GramEval 2020 overview the task, training and test data, evaluation methodology, submission routine, and participating systems. The approaches proposed by the participating systems and their results are reported and analyzed.

Key words: morphological tagging, dependency parsing, lemmatization, NLP evaluation, GramEval shared task, Russian

DOI: 10.28995/2075-7182-2020-19-553-569

[1] The publication was partly prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2020 and within the framework of the Russian Academic Excellence Project «5-100».

GRAMEVAL 2020: ДОРОЖКА ПО АВТОМАТИЧЕСКОМУ МОРФОЛОГИЧЕСКОМУ И СИНТАКСИЧЕСКОМУ АНАЛИЗУ РУССКИХ ТЕКСТОВ

Ляшевская О. Н. (olesar@yandex.ru)

НИУ Высшая Школа Экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Шаврина Т. О. (rybolos@gmail.com)

НИУ Высшая Школа Экономики; Сбербанк, Москва, Россия

Трофимов И. В. (itrofimov@gmail.com),
Власова Н. А. (nathalie.vlassova@gmail.com)

Институт программных систем им. А. К. Айламазяна РАН, Переславль-Залесский, Россия

GramEval 2020 — дорожка по оценке методов и технических решений для полного морфологического и синтаксического анализа текстов на русском языке. В 2020 году доминантой была выбрана жанровая репрезентативность текстового материала. Для оценки подходов к автоматическому анализу текста был подготовлен тестовый набор данных, охватывающий пять жанров современного языка: новости, сообщения из социальных сетей и электронную коммуникацую, энциклопедические статьи, художественную литературу, поэзию, а также исторические тексты 17 века.

Текстовый материал для обучения и тестирования предоставлялся в формате Универсальных Зависимостей (Universal Dependencies) версии 2.5. Входной формат содержал информацию о границах предложений и токенов. Задачей систем-участников было определить часть речи, грамматические признаки и лемму каждого токена, а также построить дерево зависимостей каждого предложения с типизацией синтаксических отношений.

В ходе мероприятия участники имели возможность получать оценки качества своих решений благодаря платформе CodaLab. Автоматически предоставлялась детализация оценок по уровням разметки и текстовым регистрам, информация о частотных ошибках. Окончательный рейтинг систем составлялся на основе четырёх показателей: качества определения части речи, грамматических признаков, леммы и построения дерева зависимостей (LAS).

В данной статье организаторы GramEval 2020 рассматривают основные вопросы, связанные с организацией дорожки, а также полученные участниками результаты. Затрагиваются темы методологии оценки, подготовки обучающих и тестовых данных. Приводится краткое описание подходов участников и анализ допущенных ошибок.

Ключевые слова: морфологический анализ, синтаксический парсинг, парсинг зависимостей, лемматизация, оценка систем автоматической обработки текста, дорожка GramEval, русский язык

1. Introduction

Russian grammar has a rich history of theoretical and applied modelling. Starting with the work of A.Zaliznyak [12], the grammatical description has reached a new level, making it possible to build automatic systems of morphological analysis.

Since then, technologies in Russian NLP have made significant advances thanks to data from search engines [7], [1], as well as to shared tasks based on texts from various sources. Since 2010, automatic morphological tagging has become a traditional task for Russian and international researchers.

In 2010, for the first time, a shared task was held for automatic Russian part-of-speech tagging, lemmatization, and morphological analysis, including the subtask of annotating the rare words [5]. The participants achieved 98% accuracy on lemmatization and 97.3% accuracy on the part-of-speech tags.

At the MorphoRuEval 2017 shared task [8], a 97.11% accuracy in all morphological features and 96.91% accuracy in lemmatization were achieved on a balanced set of data from various sources (news, social networks, fiction, etc.).

From 2016 to 2019, morphology also became the main focus of the multilingual competition SIGMORPHON, where for the Russian language [4] a leading result of 94.4% accuracy on word inflexion in context was obtained.

Syntactic parsing was the focus of the Ru-Eval 2012 shared task [11]. The organizers conducted a survey of existing automatic approaches and resources and provided data in a conditional dependency format. In 2017, with the advent the Universal Dependencies (UD) initiative [6], shared tasks on multilingual parsing, including Russian, became possible, combining academic and industrial development systems under a common track. CoNLL shared tasks 2017–2018 [13], [14] has set the task of complete grammatical annotation, from row text to syntax: for the Russian language, the quality of 92.48% accuracy LAS (labelled attachment score) on the materials of UD-SynTagRus and 72.24% accuracy LAS on the materials of social media was achieved. It became apparent that the quality of annotation should be evaluated on balanced datasets representing various styles and registers of writing.

In the above works, morphology and syntax are considered as independent tasks and evaluated separately; in most cases, systems that solve these problems are designed in such a way that they mark data independently at 3 levels—1) morphology 2) lemmatization 3) syntax, or at 2 independent levels—1) morphology and lemmatization and 2) syntax. Meanwhile, the relation among all three levels of the grammatical annotation is obvious: for example, an error in determining part of speech can lead to a lemmatization error and/or to an incorrect identification of syntactic relation.

We believe that the moment has come when the simultaneous intersection of the following factors allows us to create benchmark competitions in the general grammatical annotation of Russian texts, in which the overall level of annotation would be simultaneously assessed by various sources of Russian texts in all their diversity:

1. The development of deep learning, including universal language models, capable, according to some studies, of independently learning ideas about the semantic, lexical, and syntactic levels of the language [3],
2. Accumulation of big data from various sources,
3. The presence of a standardized format for morphological and syntactic annotation—UD 2.0.

The results of this initiative are presented in this article. Continuing the tradition of the previous independent shared tasks for Russian, we propose the new format of the NLP competitions—evaluating the joint models by their generalizing ability on the whole variety of language data of differing periods and sources.

2. Data

The data was provided in the UD 2.0, CONLL-U format, with respect to some variability in various data sources and their annotation methods, which will be described later.

The task of the GramEval 2020 organizers was to provide the most diverse training and test samples, taking into account the benefits of parsing quality improvements for industry, NLP research, digital humanties and theoretical linguistic research. For this reason, the main training sample with manual annotation included the most normative segments—news and fiction—as well as texts of social media, wiki, poetry, and texts of the 17th. century. Poetry was considered, since lexicon, morphemic and syntactic patterns, and word order is considered more variable in verses than in prosaic texts. As for the 17th. century data, the native speakers of Russian have almost no difficulties understanding such texts. Since it is assumed that modern processing systems are a closer match to human performance, it was interesting to take a diachronic look at the Russian NLP evaluation. An equally important factor was the availability of materials for all six registers in the UD format for training.

2.1. Training data

As training data, existing open datasets were collected from various sources: UD repository, MorphoRuEval, and RNC historical corpora collections.

- UD **SynTagRus** v2.5 (1.1M tokens, fiction, news, wiki, nonfiction). Annotation: automatic (ETAP3), human correction in native SynTagRus, then re-tokenized and converted automatically to UD 2.x. Enhanced dependencies removed. Since the treebank was not fully valid for the UD v2.5 scheme, a version with semi-manual corrections was also provided.
- UD Russian **GSD** v2.5 (96K tokens, wiki). Annotation: automatic (Google Stanford Dependencies) converted and manually checked.
- UD Russian **Taiga**: samples extracted from the Taiga Corpus and MorphoRuEval-2017 text collections (mostly social media and poetry, 39K tokens). Annotation: manual.

- **MorphoRuEval** test 2017: news (Lenta.ru, 5K tokens), fiction (magazines.gorky.media, 7K tokens) and social media (VK, 5K tokens). Morphological annotation done during the previous shared task annotations was manually changed to get better agreement with the current UD standards; syntax was annotated manually from scratch.
- **RNC 17th c.**: texts from the Middle Russian corpus (business & law, letters, Church Slavic, hybrid texts, 39K tokens). Annotation: manual, no lemmatization. In addition, 4M tokens were provided with manual morphological and automatic syntactic annotation.

2.2. Supporting data

Additional data were provided 'as is' with fully automatic annotation:

- Twitter: UDPipe pipeline (tokenization, morphology, syntax). Corpus of Russian tweets with sentiment annotation from http://study.mokoron.com.
- Wikipedia: UDPipe pipeline (tokenization, morphology, syntax). The actual dump of Russian Wikipedia, first 100,000 articles
- Comments from Russian Youtube Trends, April 2019. UDPipe pipeline (tokenization, morphology, syntax).
- Lenta.ru news: symbol unification + UDPipe pipeline (tokenization, morphology, syntax). Lenta Ru news, up to 2018.
- Stihi.ru poetry: symbol unification + UDPipe pipeline (tokenization, morphology, syntax).
- Proza.ru fiction: symbol unification + UDPipe pipeline (tokenization, morphology, syntax).
- Fiction Magazines (Taiga): UDPipe pipeline

2.3. Development and test data

The shared task included two stages: public and private test. At each stage, gold data was used, prepared specifically for the shared task. All in all, 7 annotators took part in data labeling at different linguistic levels. After that each sentence was verified by two supervisors. One of them, a contributor to Russian treebanks in the UD repository, checked through all data sets, for better consitency of the annotations. The size of the development and test sets is given in **Table 1**.

Table 1: The number of tokens and sources of the development and test sets

Register	Dev set	Dev source	Test set	Test source
news	1K	MorphoRuEval2017	1K	MorphoRuEval2017
social	1K	MorphoRuEval2017	1K	MorphoRuEval2017 + Taiga
wiki	1K	Russian GSD	1K	Wikipedia
fiction	1K	SynTagRus	1K	Taiga + RNC
poetry	1K	Taiga	1K	Taiga + RNC
17 cent	1K	Middle Russian-RNC	1K	Middle Russian-RNC

During the public test, the participants downloaded files for each of the 6 text registers, in a vertical format. The participants processed the input data with their systems and submitted the results to the leaderboard, obtaining detailed results for each source, error statistics, and could compare their result with the gold data.

During the private test phase, the participants were asked to download one large input file which included 10% test data for all six registers. The gold annotation was kept unavailable to the participants. The private test included news from UD MorphoRuEval2017, social media from UD MorphoRuEval2017 and Taiga, wiki from Wikipedia, fiction from Taiga and RNC, poetry from Taiga and RNC, 17th c. from UD MidRussian RNC.

3. Evaluation metrics

Evaluation procedure is based on the calculation of quality measures in the tasks of pos-tagging (qPos), morphological features tagging (qFeat), lemmatization (qLemma), and dependency parsing (qLas). The arithmetic mean of these values was used as participant's score (1) on a test set.

$$Score = mean(qPos, qFeat, qLemma, qLas) \tag{1}$$

Since separate test sets were created for each register, the composite participant's score (2) was calculated as an arithmetic mean for all registers.

$$Overall\,Score = mean(Score_{news}, Score_{wiki}, Score_{social}, Score_{fiction}, Score_{poetry}, Score_{17\,c.}) \tag{2}$$

$$Overall\,Score = mean(news\ score, wiki\ score, social\ score, fiction\ score, poetry\ score, 17\ c.\ score) \tag{3}$$

3.1. Pos-tagging, morphological features, lemmatization and syntax

Four main metrics—pos accuracy, other morphological features recall (macro-average over tokens), lemmatization accuracy and labeled attachment score (LAS) are measured the same way:

- Metrics are measured for each text source (news, poetry, etc), comparing participant submission results and gold markup:
  - Each predicted token annotation is being compared to the gold one:
    - Whether pos is the same as the gold one or not (POS: 1 or 0)
    - Sum all the matching features is divided by the number of the gold features (FEAT: continuous from 1 to 0)
    - Whether the lemma is the same as the gold one or not (LEMMA: 1 or 0)
    - Whether the syntactic head is the same, and if yes, is the relation correct (LAS: 1 or 0)
  - Sums of POS, FEAT, LEMMA, LAS points are being divided by the number of tokens in the text source—we get qPos, qFeat, qLemma, qLas quality
- All the quality on each source is being averaged (summed and divided by number of sources) to get overall quality.

Besides, when comparing lemmatization, letter capitalization and *e/ë* choice is not considered different. When evaluating LAS, full dependency relation (with tags after ":") was considered.

3.2. Additional metrics

In order to achieve compatibility with the universal standard and the experience of international community, additional metrics not included in the leaderboard were calculated: F1 metrics for pos, features, and dependency relations, lemmatization, as well as UAS, MLAS, BLEX metrics according to the CoNLL method[2]. These metrics were included to avoid situations in which systems could get high accuracy due to the rule-based evaluation hacking, for example, excessive addition of extra tags, etc., as well as for comparison with the results for the Russian language obtained in the previous shared tasks [13], [14].

3.3. Token alignment evaluation

For convenience of participants, the token alignment score was computed. It allows them to control whether the tokenization is corrupted or not. Every sentence from submission was compared to the corresponding gold one. Each token in a sentence was compared with the gold one. If the tokens were considered equal (see above), token alignment sum was incremented; sentence alignment score was a token alignment sum divided by the number of tokens. The participants were given the final alignment score of mean scores of every sentence.

During the private test phase, all systems had their alignment score of 100%.

4. System submission platform and routine

The competition was held on the CodaLab platform[3], which allowed participants to choose the best parameters of their systems and analyze their performance on various text sources that were known in advance.

Each participant could make up to 100 submissions per day.

Participants were also allowed to use any external data for training their systems, including non-open sources. But the resulting system itself and the models should be open source and published on the Github.

Starting from 18 initial teams, 4 systems have reached the final test phase, one of which have provided two final submissions. The authors of the final systems represent different countries (Russia, France), tech companies (Yandex, ABBYY, MTS) and universities (CEA-LIST: Laboratory for Integration of Systems and Technology, Moscow State University, Moscow Institute of Physics and Technology).

---

[2] https://universaldependencies.org/conll18/evaluation.html

[3] https://competitions.codalab.org/competitions/22902

5. Baseline system

5.1. rnnmorph + UDPipe

The starting point for the competition was a hybrid system assembled from the RNNMorph morphological analyzer [2] and the Parsito syntax module [10] (from UDPipe [9]). As of 2017, RNNMorph was a top-notch solution for the morphological analysis of the Russian language [8]. The choice of UDPipe is due to the popularity of this system and the positive experience of using it as a baseline system at CoNLL competitions [13], [14].

GramEval participants were given access to the source codes of the hybrid system with its default settings (without pretraining on the competition data). Besides, at each phase of the competition, the results of the baseline assessment were published.

5.2. Other milestones

In addition to the baseline system, we have trained and evaluated several well-known systems for morphological analysis and dependency parsing. We believe this is of help to the participants to better understand where their solutions fit in with other top ranked systems. This will also show how the quality of analysis has changed over recent years.

We selected MaltParser, SyntaxNet, UDPipe, StanfordNLP, TurkuNLP and rnnmorph for this evaluation. All of them were trained on the same training set. We followed the default settings, where possible, assuming that the developers had determined them in a rational way. Detailed information about the training setup can be found on the GramEval-2020 website[4]. Note that MaltParser, SyntaxNet and rnnmorph generate only part of the markup; these systems were evaluated within their competence. Note also that MaltParser itself cannot generate morphological features that it needs for parsing. Therefore, the morphological layers for MaltParser were generated by UDPipe.

6. Results

Table 2 presents the official leaderboard and tables 3–6 detail the quality of morphological analysis, lemmatization, and parsing for each register, respectively. Besides that, the latter tables show the results of additional baseline systems (indicated by italics).

Table 2: GramEval official leaderboard—Overall score

System	Overall Score
qbic	0.91609
ADVance	0.90762
lima	0.87870
vocative	0.85198
baseline	0.80377

[4] Default models and parameters for each module.

Table 3: Scores: parts of speech

	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.966	0.969	0.947	0.927	0.963
ADVance	0.980	0.965	0.960	0.937	0.921	0.960
lima	0.976	0.971	0.957	0.937	0.925	0.935
vocative	0.975	0.965	0.929	0.917	0.909	0.870
Turku	*0.970*	*0.964*	*0.951*	*0.926*	*0.902*	*0.870*
Stanford	*0.974*	*0.964*	*0.944*	*0.913*	*0.924*	*0.896*
UDPipe	*0.975*	*0.967*	*0.927*	*0.916*	*0.906*	*0.868*
SyntaxNet	*0.953*	*0.952*	*0.906*	*0.884*	*0.904*	*0.866*
rnnmorph	*0.970*	*0.949*	*0.946*	*0.928*	*0.922*	*0.894*

Table 4: Scores: grammatical features

	fiction	news	poetry	social	wiki	17 cent
qbic	0.987	0.981	0.967	0.947	0.944	0.929
ADVance	0.986	0.981	0.960	0.959	0.928	0.929
lima	0.979	0.966	0.956	0.953	0.967	0.896
vocative	0.948	0.944	0.898	0.900	0.904	0.793
Turku	*0.952*	*0.962*	*0.921*	*0.918*	*0.921*	*0.831*
Stanford	*0.949*	*0.957*	*0.914*	*0.904*	*0.923*	*0.841*
UDPipe	*0.946*	*0.946*	*0.899*	*0.899*	*0.902*	*0.791*
SyntaxNet	*0.934*	*0.926*	*0.886*	*0.887*	*0.872*	*0.801*
rnnmorph	*0.878*	*0.858*	*0.857*	*0.852*	*0.838*	*0.825*

Table 5: Scores: lemmatization

	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.982	0.953	0.960	0.936	0.783
ADVance	0.977	0.981	0.952	0.954	0.922	0.797
lima	0.937	0.950	0.913	0.953	0.923	0.610
vocative	0.961	0.955	0.939	0.955	0.915	0.582
Turku	*0.974*	*0.976*	*0.949*	*0.956*	*0.928*	*0.584*
Stanford	*0.973*	*0.959*	*0.926*	*0.952*	*0.922*	*0.571*
UDPipe	*0.963*	*0.957*	*0.912*	*0.941*	*0.934*	*0.579*
rnnmorph	*0.950*	*0.907*	*0.918*	*0.928*	*0.904*	*0.588*
rnnmorph	*0.878*	*0.858*	*0.857*	*0.852*	*0.838*	*0.825*

Table 6: Scores: LAS

	fiction	news	poetry	social	wiki	17 cent
qbic	0.896	0.912	0.814	0.807	0.781	0.665
ADVance	0.869	0.911	0.780	0.784	0.760	0.618
lima	0.850	0.843	0.725	0.713	0.697	0.546
vocative	0.826	0.834	0.660	0.659	0.694	0.500
Turku	*0.859*	*0.877*	*0.731*	*0.733*	*0.711*	*0.502*
Stanford	*0.854*	*0.873*	*0.709*	*0.706*	*0.703*	*0.509*
UDPipe	*0.811*	*0.817*	*0.666*	*0.644*	*0.668*	*0.462*
SyntaxNet	*0.808*	*0.802*	*0.6*	*0.614*	*0.645*	*0.446*
MaltParser	*0.599*	*0.553*	*0.404*	*0.476*	*0.436*	*0.340*

All final submissions outperformed the baseline approach in morphology, lemmatization, and parsing. As for the systems' performance by register, three groups can be distinguished:

- fiction and news are the easiest to process (in general, > 95% in pos, features, and lemmas, > 83% in syntax);
- social media, poetry, and wiki texts are more challenging to process (in general, −2% to −7% drop at the lexico-grammatical levels and more significant drop at the level of syntax, see below);
- the performance on the diachronic 17th c. data is low, especially at the level of syntax. Note, however, that the top-2 systems achieve the 96% quality in pos-tagging and the 93% quality in feature tagging.

As expected, the performance drop is more pronounced in syntax, features, and lemma processing.

In the task of parsing, for all systems there is a significant (> 8%) difference in the quality of analysis on fiction + news, on the one hand, and poetry + social + wiki, on the other hand, see Table 6 and Figure 1. However, only two participants managed to surpass the best of additional baseline systems in all registers.

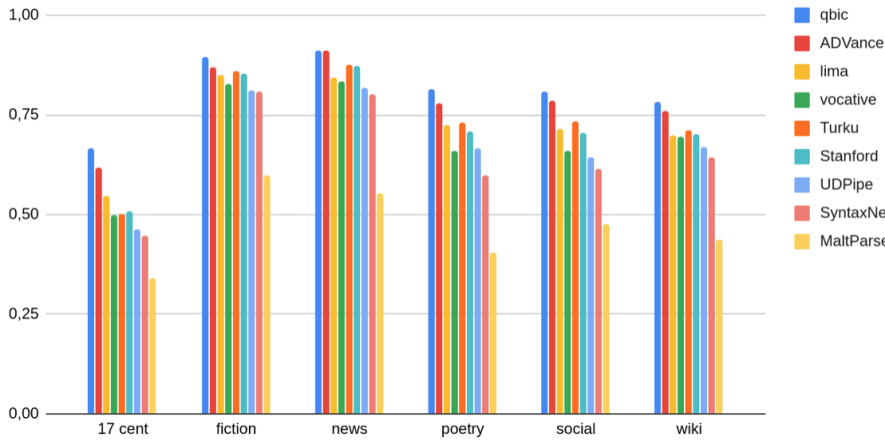

Figure 1: LAS across registers

7. Approaches adopted by the systems

The participants' approaches represent a fairly wide variety of modern neural network approaches—universal BERT transformers, recurrent neural architectures—LSTM and CRF-LSTM, feedforward layers, word and char embedding sources—BERT, word2vec and fasttext—therefore, we can say that they are quite indicative from the point of view of the current level of technology [15]–[18].

Despite the general statement of the problem, some efforts were also spent on fitting the scores of systems on specific data—two participants use rule-based

approaches to adapt the outputs of the systems for 17th c. data, and also use classifiers to detect the most outlying training sources—social media, poetry and historical data. The resulting architectures are described in Table 7.

It is noteworthy that the highest quality was shown by a system (qbic) annotating morphology, syntax, and lemmatization independently. Qualitative differences in the systems' performance are discussed in the next section.

Table 7: Table 7: Architectures of the GramEval 2020 participating systems

Team	Data	Architecture	Embeddings
qbic (1)	All GramEval data except SynTagRus	End-to-End parser: features, lemmas, and dependencies are predicted by joint BERT model with independent modules. Encoder is a single-layer LSTM, decoders are simple feedforward models for predicting lemmas and features, as well as a biaffine attention model for dependencies and their labels	Pretrained RuBERT
ADVance (2)	All GramEval data + poetry Taiga corpus for embedding training	Classifier of 4 main data sources—normative fiction, 17 c., poetry, social. + Morphotagger and parser on BERT, pretrained on SynTagRus 2.5 + 17 c. lemmer on rules	4 separately trained BERTs on GramEval data
lima (3)	All GramEval data	Original implementation of Dozat & Manning: embedding layer + LSTM layer + feedforward layer. Differs from the original models in that morphology and syntax are trained simultaneously in multitask learning mode	Pretrained FastText
vocative (4)	GramEval2020 data with rule-based parser validation for extracting good training samples for pos-tagging and parsing. + clean GramEval data for UDPipe training + own treebank data for pos tagging training	Ensemble model: 1) dictionary-based lemmatizer 2) LSTM-CRF pos tagger, considering the context and features + pure CRF pos tagger for sentences longer than 30 words + Russian UDpipe for pos and features 3) parser: UDPipe trained on GramEval data 4) Rule-based correction for 17 c. data	Pretrained word2vec wordchar2vector

8. Analysis of submitted annotations

Table 8 and 9 outline the systems' agreement in full morphology and dependency markup, respectively. In Table 8, Accuracy / Cohen's kappa for the combination of pos

and feature tags are shown. To calculate Cohen's kappa, the list of categories was determined on the basis of all the observed responses. In Table 9, Accuracy is shown calculated for the combination of syntactic head and dependency relation. Both relation types and subtypes (labels before and after the colon) were considered.

Table 8: Agreement in pos and feature markup

	qbic	ADVance	lima	vocative
gold	0.874 / 0.867	0.803 / 0.791	0.811 / 0.801	0.765 / 0.752
qbic		0.844 / 0.835	0.830 / 0.820	0.781 / 0.769
ADVance			0.767 / 0.754	0.722 / 0.706
lima				0.784 / 0.771

Table 9: Agreement in the syntactic head and dependency relation markup

	qbic	ADVance	lima	vocative
GOLD	0.813	0.788	0.729	0.697
qbic		0.826	0.769	0.720
ADVance			0.765	0.711
lima				0.706

One can note a greater cross-system agreement than that between systems and the gold markup. Indirectly, this suggests that systems make similar mistakes.

The output data of the competing systems show that the errors in morphological analysis are mostly the same as in previous competitions for the Russian language. The errors in lemmatization, pos-tagging and morphological features most often correlate.

In quantitative terms, most errors are associated with uppercase uses and non-standard spellings. Erroneous pos-tagging and morphological features arise in all the outputs at the beginning of the sentence, at the beginning of the line in the poetry, in proper names that share ambiguity with common nouns (*Наука*, *Тигр*). Furthermore, the competing systems encounter difficulties while analyzing words with spelling errors, author spelling, hashtags (typical for social networks), abbreviations and acronyms, for example, in Wikipedia references.

In lemmatization, systems found it difficult to analyze words with a rare inflectional model (*распростертый*, *огороженный*, *объемлет*, *ищет*, *горю* etc.) and pluralia tantum nouns or plural homonyms (*ножницы*, *окова-оковы*, *мозг-мозги*). Difficulties in resolving homonymy remain in pos-tagging, cf. *быть* VERB vs AUX, *что* PRON vs. SCONJ vs. PART vs. ADV, *и* CCONJ vs PART, ADV vs ADJ, DET vs PRON, uses of words like *типа* (NOUN vs ADP vs PART), *походу* (NOUN vs ADV), *смотря*, *значит* (VERB vs ADV). Such errors may also be triggered by low quality markup in training sets since frequent homonyms are difficult to be spot-checked manually. In addition, from the point of view of linguistic theory and existing corpus practices, there are well known cases that can be approached differently and thus tagged inconsistently in various training data. These include:

- participles vs. verbal adjectives (*волнующий*, *образованный*, *греющий*, *начитан*, *обязан*, etc.),
- words such as *нельзя*, *надо*, *пора* tagged as VERB vs ADV vs NOUN,
- inconsistency in lemmatizing the nouns ending with *-ие/-ье* (e.g. *безумие—безумье*) and adjectives ending with *-ой*, *-ый* (*грунтовый—грунтовой*).

In general, we observe that the systems do well with the morphological feature ambiguity. There are relatively few errors due to the paradigm syncretism (eg. *события* Case=Acc Number=Plus vs. *события* Case=Gen Number=Sing). The most common errors in morphological features are as follows:

- **`animacy`** in adjectives, pronouns and numerals (systems add **`animacy`** not only for those word forms where difficulties in analysis are possible); this feature is often mistakenly identified if the word is uppercase and / or comes first in the sentence;
- features of the verb *быть* (all systems add **`aspect`**);
- **`gender`**: the competing systems attribute **`gender`** to adjectives in the plural, make mistakes in determining the gender of proper names, there are also errors with the gender of common nouns in indirect cases (even though the lemma is defined correctly);
- **`case`**: some systems systematically add case to short adjectives and participles in the predicative position;
- **`aspect`**. Errors arise in biaspectual verbs (*подвизается*, *минует*, etc.);
- **`voice`**: all systems mark finite verb forms ended with *-ся* as passive (Pass) rather than middle (Mid);
- **`degree`** in adjectives and adverbs: the participants often do not tag superlative (Sup) and comparative (Cmp) degree.

Table 10 presents top-20 mismatches in the dependency relation labeling[5], with occurrences (N) calculated over all systems. For the most part, these are mismatches between flat syntactic relations, clause and phrase joining relations, verb-argument relations, and modifier relations mixed with either other modifiers or argument relations. The common source of errors is register-specific tokens and constructions such as '==', '*', '***', '&quot' punctuation marks in wiki and social media texts, attachment of interjunctions (**`discourse`**) incorrectly predicted as **`parataxis`**, the **`list`** relations common to the wiki biographies also predicted as **`parataxis`**. It can be seen that the best system is accurate in predicting the **`punct`** relations, which is problematic to the other three systems. At the same time, it is low-sensitive to the **`fixed`**, **`discourse`**, **`parataxis`**, **`vocative`**, **`compound`**, and **`list`** relations.

All systems tend to mix indirect objects (**`iobj`**) in the instrumental case with oblique (**`obl`**). This and a large number of other core argument relation mismatches can probably be attributed to argument relations incorrectly represented in UD-Syntagrus, the main training dataset for Russian parsers. The alluvial confusion plot (**Figure 2**) demonstrates that in many cases, the systems make errors in argument relations

[5] For the definition of relations see the UD site https://universaldependencies.org/u/dep/index.html.

differently. However, we find that, relative to the 2012 shared task, the core argument relation labeling has improved significantly and such errors are rare in all systems.

The 17th c. dataset, on which all systems demonstrated low scores, require separate attention. In these texts, archaic endings and orthography are main factors that noticeably affect the system's performance in pos and lemma analysis. No more than one system was able to analyse words such as *такова*, *всяково*, *топерь*, *акроме*, *окияно*, *козною* for pos and *двесте*, *детеи*, *земнаго*, *плаваючя*, *звер* , *итить*, *шездесят* for lemmas correctly. Register-specific training efforts were justified when analysing the close-class words such as *аз*, and *который*. Unlike in other registers, nominal phrases were challenging for parsing due to time-specific syntactic patterns in named entities (e.g. mixed genitive-possessive construction in *на Романове отца их помесье*, compound in *Иль мурзою*) and the long chains of genitive groups with inverse word order.

Table 10: Most frequent mismatches in dependency relations

N	gold	predicted	N	gold	predicted
74	punct	discourse	19	obj	nsubj
49	parataxis	appos	18	amod	nummod
42	iobj	obl	18	punct	parataxis
40	list	parataxis	17	obj	obl
38	parataxis	conj	16	conj	parataxis
32	discourse	parataxis	15	appos	nmod
24	amod	appos	15	discourse	advmod
24	nmod	appos	15	mark	advmod
23	obl	nmod	15	xcomp	obl
20	nsubj	obj	14	appos	parataxis

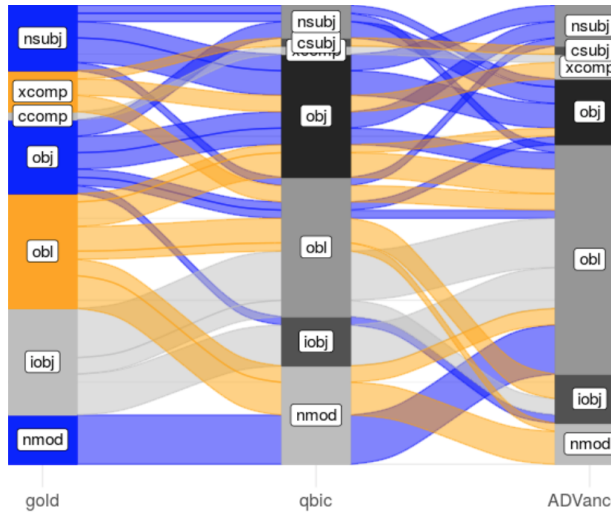

Figure 2: Core argument and nominal modifier relations incorrectly predicted by either of top-2 systems

9. Conclusion

During GramEval shared task we have introduced a new approach to full morphology and dependency parsing evaluation for Russian:

- testing and training procedures were carried out on greater variability of text sources—considering temporary, stylistic and genre variation
- the public and private test phases were organized on an open platform, expanding the capabilities of participants and allowing them to become more familiar with the overall performance of their systems on different data;
- new training data was prepared, both with automatic annotation and with both automatic and expert assessment of the data;
- the competition guidelines provide compatibility with the UD standard, as well as at the level of additional metrics—compatibility with the CoNLL competitions;
- as the result of the competition, a comparison of different parsing strategies was obtained, and a new state-of-the-art method for full Russian morphological parsing of Russian.

The competition leaderboard is now permanent at the CodaLab, and we welcome researchers and developers to submit their systems to the leaderboard and compare their results with other approaches.

All materials of GramEval 2020 including supplementary tables and figures for this paper are available at the shared task repository[6]. As the collection represents the vast variety of genres, registers, corpora, annotation practices, with new development and test data checked manually, we hope that the output GramEval 2020 will stay practical and relevant for the NLP community.

10. Acknowledgements

The authors would like to thank the participants of GramEval 2020 for helpful comments and valuable suggestions. We are grateful to all members of the community who contributed to the data preparation and evaluation, scripts, and setting up the GramEval 2020 CodaLab page, and especially to Vera Davydova, Aleksey Dorkin, Maria Ermolova, Aleksandra Konovalova, Kristina Litvintseva, Elizaveta Nosova, Anna Safaryan, Dmitry Sichinava, and Elena Suleymanova.

[6] https://github.com/dialogue-evaluation/GramEval2020

References

1. *Anastasyev D.* (2020) Exploring pretrained models for joint morpho-syntactic parsing for Russian. In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020, Vol. 19.
2. *Anastasyev D., Gusev I., Indenbom E.* (2018) Improving part-of-speech tagging via multi-task learning and character-level word representations. In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2018, Vol. 17, pp. 14–27.
3. *Bocharov V. V., de Chalendar G.* (2020) The Russian language pipeline in the LIMA multilingual analyzer. In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020, Vol. 19.
4. *Clark K., Khandelwal U., Levy O., and Manning C. D.* (2019) What Does BERT Look At? An Analysis of BERT's Attention. In: arXiv:1906.04341.
5. *Cotterell R., Kirov Ch., Sylak-Glassman J., et al.* (2018) The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In Proceedings of CoNLL–SIGMORPHON 2018.
6. *Dereza O. V., Kayutenko D. A., Fenogenova A. S.* (2016) Automatic morphological analysis for Russian: A comparative study. In Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. Student session (online publication). Retrieved from: http://www.dialog-21.ru/media/3473/dereza.pdf.
7. *Grashchenkov P. V., Koziev I.* (2020) POS-tagger Dataset Augmentation by Ensemble Parsers. Unpublished Ms. Moscow.
8. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A., et al.* (2010) NLP evaluation: Russian morphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2010, Vol. 9 (16), pp. 318–326.
9. *Nivre J., De Marneffe M. C., Ginter F., et al.* (2016) Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1659–1666.
10. *Segalovich I.* (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, In Proceedings of MLMTA-2003, Las Vegas, Nevada, USA.
11. *Sokirko A. V.* (2001) Semantic Dictionaries in the Natural Language Processing: Based on the DIALING system [Semanticheskie slovari v avtomaticheskoj obrabotke teksta: Po materialam sistemy DIALING]. Cand. Tech. Sc. Dissertation. Moscow.
12. *Sorokin A., Shavrina T., Lyashevskaya O., et al.* (2017) MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian. In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2017. Vol. 16–1, pp. 297–313.
13. *Sorokin A. A., Smurov I., Kirjanov D.* (2020) Tagging and parsing of multidomain collections. In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020, Vol. 19.

14. *Straka, M., Hajič, J., Straková, J., and Hajič jr., J.* (2015) Parsing universal dependency treebanks using neural networks and search-based oracle. In Proceedings of 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015), Warszawa, Poland.
15. *Straka M., Hajič J., Straková J.* (2016) UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.
16. *Toldova S., Sokolova E., Astafiyeva I., et al.* (2012) NLP evaluation 2011–2012: Russian syntactic parsers [Ocenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka]. In Computational linguistics and intellectual technologies. Proceedings of International Conference Dialogue 2012. Vol. 11 (18), pp. 797–809.
17. *Zaliznyak, A. A.* (1977) Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo yazyka]. Moscow.
18. *Zeman D., Hajič J., Popel M., et al.* (2018) CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–21.
19. *Zeman D., Popel M., Straka M., et al.* (2017) Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–19, Vancouver, Canada.

Appendix

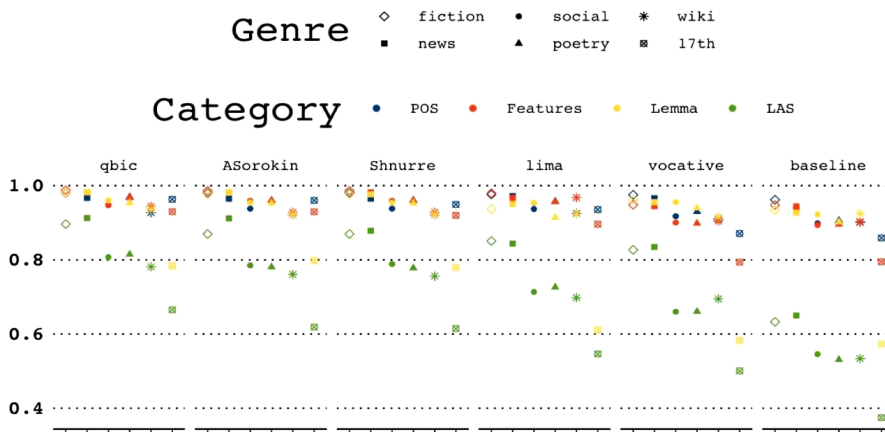

Figure 3: Systems' scores by register