Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2026”

June 24–26, 2026

Beyond Accuracy: Systematic Errors in Russian Morphological Analyzers and the Ensemble Consensus-based Approach

Valery Solovyev
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
Maki.solovyev@mail.ru

Anna Ivleva
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
ivleva.anna.igorevna@yandex.ru

Abstract

While numerous tools for Russian morphological analysis exist, their systematic errors and biases can undermine downstream linguistic applications. This paper presents a comprehensive evaluation of six state-of-the-art Russian morphological analyzers (Natasha, SpaCy, pymystem3, DeepPavlov, Stanza, UDPipe). Moving beyond aggregate accuracy scores, we conduct a fine-grained quantitative and qualitative analysis of systematic errors using the GRAMEVAL 2020 and SynTagRus datasets. Our analysis reveals significant biases and systematic patterns in lemmatization and POS-tagging. To mitigate these individual biases, we propose and validate two consensus-based ensemble methods which achieve statistically significant improvements over the best individual analyzer (from +0.31% to +2.14% gain). Our results demonstrate that while no single tool is universally optimal, their complementary strengths can be leveraged to produce more reliable annotations for linguistically sensitive research. Still, no consensus eliminates errors shared by all or a majority of tools.

Keywords: Russian morphological analysis; lemmatization; part-of-speech tagging; systematic error analysis; ensemble consensus-based methods; Universal Dependencies

DOI: 10.29003/2075-7182-2026-24-606-617

За пределами accuracy: систематические ошибки морфологических анализаторов русского языка и ансамблевый консенсус-метод

Соловьев В. Д.
Лаборатория лингвистики и ИИ,
Институт филологии и межкультурной коммуникации,
Казанский федеральный университет,
г. Казань, Россия
maki.solovyev@mail.ru

Ивлева А. И.
Лаборатория лингвистики и ИИ,
Институт филологии и межкультурной коммуникации,
Казанский федеральный университет,
г. Казань, Россия
ivleva.anna.igorevna@yandex.ru

Аннотация

Несмотря на высокое качество современных инструментов для морфологического анализа русскоязычных текстов, систематические ошибки анализаторов могут снижать качество дальнейшего лингвистического анализа. В данной статье представлена всесторонняя оценка шести современных анализаторов русской морфологии (Natasha, SpaCy, pymystem3, DeepPavlov, Stanza, UDPipe). Кроме общепринятых показателей точности лемматизации и частеречной разметки, мы проводим детальный количественный и качественный анализ систематических ошибок на основе данных GRAMEVAL 2020 и SynTagRus. В ходе анализа выявлены значительные погрешности и систематические закономерности при лемматизации и разметке частей речи. В работе предлагается два ансамблевых метода, основанных на консенсусе, которые позволяют нивелировать некоторые из систематических погрешностей анализаторов. Предложенные методы достигают статистически значимых улучшений по сравнению с наилучшим

анализатором (прирост от +0.31% до +2.14%). Результаты исследования показывают, что ни один отдельный инструмент не является универсально оптимальным, однако взаимодополняющие «сильные стороны» можно использовать для получения более надежной морфологической разметки для чувствительных к точности лингвистических исследований. Тем не менее, ни один консенсус не устраняет общие для всех или большинства инструментов ошибки.

Ключевые слова: морфологический анализ русскоязычных текстов; лемматизация; частеречная разметка; систематические ошибки; ансамблевые методы на основе консенсуса; Universal Dependencies

1 Introduction

High-quality morphological analysis, particularly lemmatization and part-of-speech (POS) tagging, forms the foundation for most NLP pipelines. For morphologically rich languages like Russian, where words take numerous inflected forms, accuracy at this stage is critical for downstream tasks.

While large language models (LLMs) implicitly encode morphological knowledge, they do not guarantee consistency, reproducibility, or controllability of morphological predictions, especially in low-frequency, ambiguous, or domain-specific cases [1, 2].

Explicit morphological analysis remains essential in corpus linguistics as a prerequisite for constructing frequency dictionaries and studying lexical and grammatical distributions in large corpora. Morphological annotation is also widely used in digital humanities, where researchers analyze stylistic variation, authorship, and diachronic language change on large text collections. In addition, interpretable linguistic features, such as lemmas and POS tags, remain important in explainable NLP pipelines, as well as in educational technologies and language learning systems. This makes improving robustness of morphological analyzers a practically relevant task.

Although multiple Russian morphological analyzers report strong benchmark performance, aggregate accuracy may conceal concentrated errors in specific grammatical categories or lexemes. For linguistic applications such as frequency analysis or POS distribution studies, such systematic biases may distort conclusions even when overall scores are high.

Therefore, we evaluate six state-of-the-art analyzers (Natasha, SpaCy, pymystem3, DeepPavlov, Stanza, UDPipe) not only in terms of accuracy but also with respect to systematic error patterns.

We additionally propose and validate consensus-based ensemble methods designed to mitigate complementary biases across tools. While such methods cannot eliminate errors shared by all analyzers, they can reduce sensitivity to individual model biases.

The main contributions of this work are: (1) a comprehensive performance benchmark of six state-of-art Russian morphological analyzers on a multi-genre dataset; (2) qualitative and quantitative analysis of persistent and systematic errors; (3) a consensus-based ensemble method to compensate for complementary biases; (4) rigorous statistical analysis using the Wilcoxon signed-rank test.

Section 2 reviews related comparative studies and provides a brief overview of the six tools. Section 3 describes datasets, tools, and evaluation methodology. Section 4 presents results and discussion, including overall performance comparison, quantitative analysis of systematic biases, disagreement analysis, qualitative error patterns, and statistical validation of consensus methods.

2 General Instructions

2.1 Related comparative studies

Most surveys can be classified as narrow-scope evaluations examining tool performance in particular settings, or broad-coverage benchmarks offering common ground for comparing multiple systems on diverse datasets.

Early comparative evaluations highlighted solution diversity but predated neural models including transformer-based architectures. In [3], they evaluated three Russian morphological taggers (TreeTagger, TnT, Stanford POS Tagger) for lemmatization, POS-tagging, and full morphological tagging. In [4], authors compared a wider range of parsers including MyStem and TreeTagger. [5] is an application-oriented survey comparing TreeTagger, UDPipe, and the neural rnnmorph model, confirming neural superiority for general-purpose texts. Research [6] evaluated tools on Russian social media, demonstrating significant accuracy drops when tools trained on news or literary corpora are applied to informal content.

The GRAMEVAL 2020 shared task [7] provided a unified multi-genre benchmark for Russian morphology and dependency parsing with Universal Dependencies (UD) annotations, allowing direct comparison of diverse systems including Stanza and UDPipe. Politsyna et al. [8] analyzed several morphological tools on Russian National Corpus (RNC) data, underscoring the need for ongoing refinement. Recent work [9] introduced Rubic2, an ensemble model for Russian lemmatization combining a fine-tuned BART model with an existing neural solution. Their results demonstrate that ensemble methods can surpass current state-of-the-art performance. This observation is consistent with prior work on classifier agreement, which shows that different models tend to exhibit complementary error patterns, creating opportunities for consensus-based improvement [10].

Modern evaluation frameworks increasingly prioritize aggregate metrics such as full morphological tagging and dependency accuracy (see, for instance, [11]), often assuming that lower-level tasks like lemmatization and POS-tagging are largely solved. Consequently, most comparative studies focus on overall accuracy while paying limited attention to systematic biases and their potential impact on linguistic analysis. We argue that this assumption is premature: even state-of-the-art analyzers exhibit persistent systematic errors that accumulate in large corpora and may distort linguistic conclusions [10].

The emergence of LLMs has introduced a new paradigm in morphological processing. Unlike traditional analyzers trained explicitly for sequence labeling, LLMs acquire morphological and syntactic regularities implicitly during large-scale pretraining [1, 13, 14]. This enables zero-shot and few-shot morphological tagging, often with competitive performance.

However, several limitations prevent LLMs from fully replacing dedicated morphological analyzers. First, LLM predictions do not guarantee consistency across prompts and runs. Second, they do not ensure adherence to formal annotation schemes such as Universal Dependencies. Third, their errors are difficult to interpret and systematically analyze, which is critical for linguistic research [2].

As a result, recent work increasingly explores hybrid approaches, where LLMs are used to complement rather than replace traditional tools — for instance, in reranking, disambiguation, or ensemble settings [9, 15]. This trend reinforces the importance of understanding systematic error patterns in existing analyzers, as these errors directly affect the behavior of higher-level systems built on top of them.

2.2 Modern Russian morphological analyzers

The landscape of Russian morphological analysis includes tools based on diverse architectural principles. We selected six widely used and actively maintained analyzers representing rule-based, hybrid, and neural approaches.

Natasha [16] employs a neural NewsMorphTagger model based on slovnet morphology. SpaCy [17] is an industrial NLP framework; we use its ru_core_news_lg model. Pymystem3 [18] is a Python wrapper for Yandex MyStem, a hybrid dictionary-based and rule-based system. DeepPavlov [19] provides a BERT-based morpho_ru_syntagrus_bert model fine-tuned for sequence labeling. Stanza [20] offers a fully neural Bi-LSTM-CRF pipeline trained on UD data. UDPipe [21] is a trainable pipeline using perceptron-based models optimized for UD treebanks.

We also include qbic, the top-performing system from the GRAMEVAL 2020 shared task, as a reference point since its results are published in the shared task evaluation.

In Section 4 (Table 1), we present accuracy for the above analyzers. By including these tools, we capture the diversity of modern approaches to Russian morphology.

3 Data and Methods

3.1 Data

For test purposes, we selected development datasets from the GRAMEVAL 2020 shared task [22], combining four distinct sources: Wikipedia-style texts (GSD-wiki), news articles from Lenta.ru, social media posts from VKontakte, and sentences from the SynTagRus corpus of modern Russian literature. This multi-genre combined dataset comprises 266 sentences and 3,272 tokens, challenging analyzers' ability to handle varying lexical and stylistic norms.

In addition, we report results on a larger SynTagRus test dataset (128,358 tokens) and on data derived from the RNC (1 million tokens of manual markdown) to assess the robustness of the observed

patterns. The smaller dataset allows controlled comparison across genres, while the larger corpora provide a more statistically stable estimate of overall performance.

For systematic bias evaluation, we additionally combined SynTagRus development and test datasets, 252,702 tokens in total.

3.2 Tools and experimental setup

The models evaluated are: Natasha (NewsMorphTagger), SpaCy (ru_core_news_lg v3.7.2), pymystem3 (v0.2.0), DeepPavlov (morpho_ru_syntagrus_bert v1.1.1, run in Python 3.9 due to dependencies), Stanza (v1.4.0 with default Russian model), and UDPipe (v2.0 with pre-trained Russian model). All tools except DeepPavlov were executed in Python 3.12 on an AMD Ryzen 7 7700 processor.

POS tag compatibility. Most libraries follow UD standards and directly produce compatible tags. However, pymystem3 uses a proprietary tagset requiring mapping. We map S tags with proper noun indicators (имя, фам, отч, гео, орг, аббр) to PROPN, other S tags to NOUN; SPRO to PRON; APRO to DET; CONJ matches either CCONJ or SCONJ; A to ADJ; V to VERB; ADVPRO to ADV; PR to ADP; ANUM to NUM; NONLEX, INIT, UNKN to X.

However, residual differences in annotation schemes and training data may still influence model predictions. These differences are reflected in systematic error patterns and contribute to the complementary behavior of analyzers observed in this study. This further motivates the use of ensemble methods, which can mitigate inconsistencies arising from heterogeneous training data.

3.3 Token alignment

Different tokenization schemes pose a challenge for comparative evaluation. We used the tokenizations library (https://github.com/explosion/tokenizations) to calculate character-level alignments between gold tokens and each tool's tokens. For one-to-many alignments (a single gold token split into multiple tool tokens), we reconstructed the predicted lemma by concatenating lemmas of all aligned tokens. The POS tag was taken from the first constituent token. All gold tokens except punctuation were included in evaluation.

3.4 Evaluation metrics

General performance was assessed using token-level accuracy for lemmatization and POS-tagging.
To identify systematic biases, we calculated:

1. Error rates by POS.
2. 100% error rate ($ER_{100\%}$): the percentage of errors that are completely systematic. For each (token, gold POS) pair appearing at least 4 times (for statistical reliability and trade-off between coverage and robustness) in the corpus, we identified cases where the analyzer was wrong on every occurrence:

$$ER_{100\%} = \frac{N_{100\% errors}}{N_{errors}} \cdot 100\%$$

3. Disagreement measure: for a pair of classifiers A and B,

$$\text{Disagreement}(A, B) = \frac{(N_{01}+N_{10})}{N},$$

where N_{01} is the number of tokens, where A is wrong and B correct, N_{10} – vice versa. Higher disagreement indicates greater potential for error compensation in ensembles.

We focus our evaluation on lemmatization and POS tagging as the core components of morphological analysis. These tasks are directly used in a wide range of downstream applications, including frequency dictionary construction, corpus linguistics, and information retrieval, where reliable normalization and grammatical categorization are essential.

In addition, lemmatization and POS tagging are the most consistently implemented and comparable outputs across different morphological analyzers. In contrast, full morphological feature annotation (e.g., case, gender, number, etc.) is less standardized and often relies on tool-specific tagsets or partially incompatible representations, which complicates direct comparison and large-scale evaluation.

3.5 Statistical significance testing

We employed the Wilcoxon signed-rank test, appropriate for paired observations without normality assumptions, accounting for both direction and magnitude of differences.

3.6 Consensus methods

The baseline was a simple agreement rule: a prediction is accepted in case of plurality/majority (6-0, 5-1, 4-2, 4-1-1, 3-2-1, etc. splits). For cases without majority (e.g., splits like 3-3, 2-2-2), we tested two weighted consensus strategies:

1. Simple Weighted Consensus (SWC) uses static weights representing each analyzer's general accuracy on an independent development dataset (ru_syntagrus-ud-dev.conllu).
2. POS-dependent Weighted Consensus (PWC) is more nuanced: for cases without majority stage 1 determines a preliminary POS tag using SWC; stage 2 makes the final prediction by maximizing the sum of POS-specific weights pre-calculated on the development set. Both methods provide decisions in all cases.

4 Results and Discussion

4.1 Overall performance

Tables 1 and 2 summarize lemmatization and POS-tagging accuracy on the combined multi-genre dataset (3,272 tokens), SynTagRus test dataset (128,358 tokens) and RNC dataset (1 million tokens).

Tool	GSD-wiki	Lenta-news	Social media	Fiction	Combined dataset	Syn-TagRus	RNC dataset
DeepPavlov	88.53	91.71	84.22	87.07	87.93	85.71	86.26
pymystem3	92.43	89.34	94.42	91.82	92.00	91.39	92.73
Natasha	94.92	97.75	96.48	96.31	96.36	95.21	93.58
SpaCy	91.96	93.60	88.47	89.58	90.95	90.28	88.34
Stanza	93.62	96.68	96.12	**98.15**	96.09	96.44	93.41
UDPipe	92.55	95.85	93.93	97.63	94.93	95.32	91.95
qbic*	93.60	98.20	96.00	98.00	–	–	–
SWC	96.34	97.99	97.57	98.02	97.46	97.34	95.32
PWC	**97.05**	**98.34**	**97.82**	97.89	**97.77**	**97.65**	**95.57**

* GramEval2020 winner

Table 1: Lemmatization accuracy of individual analyzers and consensus methods (percentages).

Tool	GSD-wiki	Lenta-news	Social media	Fiction	Combined dataset (3,272 tokens)	Syn-TagRus test (128,358 tokens)	RNC test (1 million tokens)
DeepPavlov	91.84	97.27	93.93	98.15	95.23	95.26	92.09
pymystem3	87.00	91.82	91.38	92.22	90.56	91.28	**93.48**
Natasha	92.67	97.99	92.11	93.40	94.07	93.64	88.33
SpaCy	94.68	97.75	92.48	95.78	95.17	94.45	90.13
Stanza	93.85	97.75	94.66	97.49	95.91	95.65	91.16
UDPipe	91.13	97.63	93.57	**98.81**	95.20	94.45	90.37
qbic*	92.70	96.60	94.70	98.00	–	–	–
SWC	94.33	98.10	94.78	98.42	96.36	96.30	91.94
PWC	**94.92**	**98.14**	**94.78**	98.42	**96.52**	**96.42**	92.06

* GramEval2020 winner

Table 2: POS-tagging accuracy of individual analyzers and consensus methods (percentages).

Performance varies significantly by genre. On news texts, all models achieve maximum results (POS accuracy up to 98.1%). On social media, lemmatization quality drops sharply: DeepPavlov from 91.71% to 84.22%, SpaCy from 93.6% to 88.47%. Informal texts remain most challenging.
Tools demonstrate different balances of lemmatization vs. POS tagging. Pymystem shows high lemmatization (94.42% on social texts) but lags in POS tagging. DeepPavlov and SpaCy excel at POS tagging across genres. Stanza, Natasha, and UDPipe exhibit more balanced performance. As shown in Tables 1–2, the proposed PWC method achieves the best overall performance on all the genres (except for fiction), on the combined dataset, Syntagrus and RNC data (lemmatization only).

Improvements in POS tagging are less pronounced on RNC. This is due to substantial differences between the RNC annotation scheme and Universal Dependencies (UD). To enable comparison, we introduced an explicit mapping between the two systems. However, many categories remain inherently difficult to align unambiguously. In particular, the distinctions between AUX and VERB, APRO and DET, SPRO and PRON, as well as the treatment of ADVPRO, predicative and parenthetical categories (PRAEDIC, PARENTH, PRAEDICPRO) introduce substantial annotation ambiguity. In addition, RNC does not distinguish between coordinating and subordinating conjunctions, and categories such as ANUM may overlap with both ADJ and NUM interpretations. Pymystem3 expectedly achieves the best POS-tagging results on RNC, which supports the hypothesis that RNC annotation conventions are structurally closer to traditional Russian morphological annotation schemes than to UD-based representations.

Nevertheless, the overall qualitative trends remain consistent, confirming that ensemble approach PWC demonstrates reliability for heterogeneous data.

4.2 Quantitative error analysis

4.2.1 Error rates by POS

Error rates by POS reveal complementarity patterns (see Appendix A). SpaCy and DeepPavlov exhibit highest failure rates in pronoun analysis (19.3% and 18.9% respectively), an order of magnitude higher than other tools (Stanza: 0.5%, UDPipe: 2.0%, Natasha: 5.2%). The same pattern holds for determiners: SpaCy (12.1%) and DeepPavlov (15.7%) versus 2.9-9.3% for others. Verb analysis is more balanced (4.6% for Natasha to 19.8% for pymystem). All tools struggle moderately with proper nouns (8-17% errors). The best-performing tool varies by POS: Stanza leads in pronouns and determiners, Natasha excels with verbs, demonstrating no single analyzer achieves universal superiority.

4.2.2 Error rates by identical tokens

On the large SynTagRus dataset (252,702 tokens), Stanza (96.5% lemma accuracy, 95.6% POS accuracy) made 8,959 lemmatization errors. Of these, 2,429 had incorrect POS also, while 6,530 were lemma-only errors. Additionally, 8,776 POS errors occurred with correct lemmas. High-frequency items account for disproportionate shares of mistakes.

Table 2 presents $ER_{100\%}$ (percentage of token-POS pairs with frequency >=4 where analyzer was wrong on every occurrence).

Analyzer	Lemma $ER_{100\%}$	POS $ER_{100\%}$
Stanza	36.4%	35.2%
Natasha	46.7%	24.7%
SpaCy	52.8%	28.2%
DeepPavlov	68.5%	39.3%
pymystem3	64.2%	56.1%
UDPipe	21.4%	33.7%
SWC	35.0%	35.4%
PWC	34.8%	33.6%

Table 2: Percentage of (token, POS) with 100% errors (frequency >=4).

DeepPavlov shows most systematic lemmatization errors (68.5%), pymystem3 most systematic POS errors (56.1%). UDPipe has lowest systematic lemma errors. Natasha lowest systematic POS errors. Consensus methods improve systematic errors for many analyzers but do not reach best individual scores.

4.2.3 Disagreement measure

Disagreement values range from 3.38% to 15.75% for lemmatization, and from 2.38% to 10.77% for POS tagging (Figure 1). Tools disagree more on lemmatization than POS tagging. Pairs involving pymystem3 show highest disagreement, confirming its error patterns differ substantially from others. Stanza+UDPipe show lowest disagreement, indicating similar error patterns. POS-stratified evaluation shows ADJ, NOUN, PRON, PROPN, ADV, DET, and NUM are top sources of complementary errors in over 70% of tool pairs.

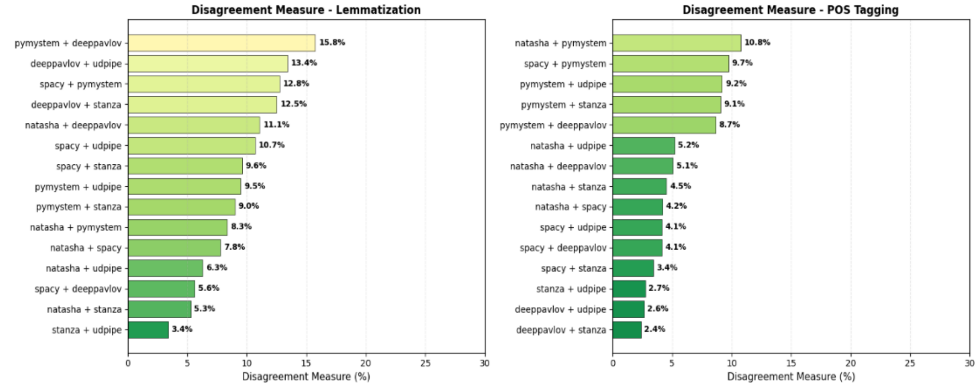

Figure 1. Disagreement Measure for All Tool Pairs (based on 252,702 tokens)

4.3 Qualitative error analysis

To move beyond specific examples, we group errors into recurring linguistic clusters that capture systematic weaknesses shared across analyzers. Appendix B gives high-frequency error clusters for each analyzer.

The most frequent error types include pronoun and determiner ambiguity, lemmatization of comparative and superlative forms, short-form adjectives, inflected proper nouns (especially of foreign origin), and highly ambiguous function words. Many of these errors are concentrated in the “long tail” of the language, including rare lexical items, non-standard forms, and out-of-vocabulary proper names, which remain challenging even for modern neural models.

At the same time, different architectural approaches exhibit distinct failure modes. Neural models tend to be robust for frequent in-context forms but less reliable on rare or irregular items, while dictionary- and rule-based systems struggle with neologisms and unseen wordforms but handle core vocabulary more consistently. This complementarity provides an empirical motivation for ensemble methods.

Natasha's top errors (DET→ADJ, DET→PRON, NOUN→PROPN) are shared with other analyzers. For example, the possessive "его" (his) analyzed as personal pronoun "он" leads to both POS and lemma errors. Natasha tends to nominalize verbs (VERB→NOUN, 5.63%), e.g., "продадите" and "дрожали" misclassified as nouns. PART↔ADV and PART↔CONJ errors show sensitivity to particle ambiguity. Other patterns: short adjectives are classified as verbs ("покоен" → "покоить"), preference for -ый adjective endings ("основное" → "основный"), wrong lemmatization of non-standard nouns ("деньги" → "деньга"), preference for -ие noun endings ("счастьем" → "счастие").

SpaCy shows highest error concentration, with DET confusions accounting for nearly 23% of POS errors. It uniquely shows PRON→SCONJ errors. The model struggles with past tense of "быть" (was), predicting inflected forms instead of canonical "быть". It also shares problems with NOUN (Noun) ↔ PROPN, as well as participles and adverbial participles, for example “соединенных” is lemmatized “соединить” instead of соединенный, “составное” is misclassified as a noun and lemmatized as

“составное” instead of “составной”. -ый ending for adjectives and -ие ending for nouns, as well as patterns for “деньги”, “времен” are also shared with Natasha analyzer.
Stanza exhibits extreme DET confusion (27.37%), with unique PRON→DET errors (3.21%). Words like "этот", "весь", "свой" are frequently tagged as PRON. Particle ambiguity and NOUN↔PROPN problems persist.

UDPipe demonstrates bidirectional DET↔PRON confusion (23.0%) combined with ADJ→NUM errors (8.55%). It uniquely shows bidirectional NOUN/PROPN confusion. X→PROPN errors indicate sophisticated unknown token handling.

DeepPavlov shows DET confusion (26.77%), with DET→NUM (многие, одним), DET→ADJ (других, самые), ADJ→DET (одно). It has the highest ADJ→NUM rate (10.24%). Unique errors include AUX→PART (3.47%) and DET→NUM (2.13%). The particle "бы" and auxiliary "бы" are homonymous; DeepPavlov systematically prefers particle ignoring conditional mood context. The lemmatizer often leaves words in inflected form ("тем" instead of "то", "ему" instead of "он").

Pymystem3 shows entirely different patterns. PROPN→NOUN (13.39%) is most frequent, opposite direction to neural models. Bidirectional NUM↔ADJ confusion (15.70%) involves hyphenated ordinals (1-й, 20-е). PART→CONJ (7.24%) is unique to PyMystem. ADV→CONJ (3.41%) and PRON→CONJ (2.97%) show functional class confusion.

4.4 Statistical significance and consensus performance

Wilcoxon signed-rank tests confirm both SWC and PWC give statistically different evaluations on both datasets ($p < 0.01$ for all comparisons with best individual analyzer). On SynTagRus, bootstrap analysis (1000 resamples) shows PWC mean accuracy 96.91% (95% CI: [96.84%, 96.98%], min = 96.79%, max = 97.02%) for lemmatization and 95.97% (95% CI: [95.89%, 96.05%], min = 95.82%, max = 96.09%) for POS-tagging. PWC outperforms Stanza by 0.46% for lemmas, 0.41% for POS. Error reduction reaches 10.11% for DeepPavlov lemmatization and 4.27% for pymystem3 POS.

On the multi-genre dataset, PWC achieves lemma accuracy 97.77% (95% CI: [97.28%, 98.26%], min = 96.91%, max = 98.56%) for lemmatization and 96.51% (95% CI: [95.87%, 97.13%], min = 95.39%, max = 97.59%) for POS accuracy. Against best individuals (Natasha for lemmas, Stanza for POS), PWC achieves improvements of +1.67 and +0.62 percentage points (absolute differences) respectively. For the RNC dataset, PWC achieves improvement of +2.14 percentage points (absolute differences). These gains are substantial for large-scale corpus annotation.

While the absolute improvement may appear modest, its practical impact becomes evident at scale. For example, in frequency dictionary construction based on a 30-million-token corpus, a +1.67% gain in lemmatization corresponds to approximately 500,000 tokens, while a +0.62% gain in POS tagging corresponds to about 186,000 tokens.

For a frequency dictionary yielding approximately 72,000 frequent lemmas and 384,000 rare lemmas, a +1.67% improvement in lemmatization corresponds to avoiding manual correction for about 1,200 frequent lemmas and 6,400 rare lemmas (over 7,600 lemmas in total). Similarly, a +0.62% improvement in POS tagging reduces manual effort by approximately 2,800 lemmas overall. Assuming a conservative estimate of 10–20 seconds for manual validation of a single lemma (including context inspection and decision-making), the reduction of approximately 10,400 lemmas corresponds to about 40–45 hours of expert work. In practice, high-quality linguistic annotation typically requires validation by at least 2–3 experts to ensure consistency and reliability. That is why the effective effort increases to approximately 80–135 person-hours (2–3 working weeks).

This highlights that even seemingly small improvements in accuracy translate into substantial savings in manual effort and improve the scalability of corpus-based linguistic analysis.

5 Discussion

While the ensemble approach itself is based on standard weighted voting, its effectiveness in our setting stems from the structured and complementary nature of systematic errors in morphological analyzers.

The main contribution of this work lies in identifying and exploiting these patterns, rather than in proposing a fundamentally new aggregation algorithm.

Importantly, the qualitative trends observed are consistent for several test datasets evaluated. The same error types that dominate the large-scale analysis are also reflected in disagreements and corrections observed in the smaller dataset, providing a direct link between error analysis and accuracy improvements.

The results demonstrate that consensus-based methods, particularly PWC, provide robust statistically superior solutions by leveraging complementary strengths while mitigating individual weaknesses.

Numerous experiments we conducted show that by varying the combinations of analyzers within an ensemble, we can reach even higher indexes for specific datasets. Still, the most robust results are obtained both for lemmatization and POS-tagging by the full ensemble including 6 analyzers.

However, consensus cannot eliminate errors shared by all or most tools.

For linguistic applications, specific bias often matters more than overall accuracy. If an analyzer systematically fails on certain parts of speech or produces erroneous lemmatization patterns, this fundamentally distorts frequency distributions, artificially deflating common function words and inflating rare items.

The PWC method has got some limitations:

- Shared errors persist where all tools agree on wrong predictions.
- The method requires running six tools.

Notably, while PWC improves overall accuracy, it does not always reduce the proportion of most systematic errors to the level of the best individual tool for that specific metric (UDPipe for lemmas, Natasha for POS).

Study limitations include:

- Mapping of pymystem3 proprietary tagset to UD is an imperfect approximation.
- Evaluation confined to UPOS and lemmas, excluding full morphological features.
- The fact that several analyzers were partially trained on SynTagRus may lead to optimistic estimates on this dataset. To mitigate this effect, we additionally report evaluation on RNC data.

From a computational perspective, the proposed ensemble requires running all individual analyzers, followed by a lightweight consensus step. Therefore, the total runtime can be approximated as the sum of execution times of the individual tools plus an overhead for aggregation.

While this increases computational cost compared to a single analyzer, it is important to note that many target applications of morphological analysis — such as corpus annotation, frequency dictionary construction, and linguistic resource development — are inherently offline processes. In these settings, annotation is typically performed once, after which the processed data are stored and reused (e.g., in searchable databases or published corpora).

As a result, computational efficiency is less critical than annotation quality and robustness, and the additional cost of the ensemble is justified by the observed improvements in accuracy and reduction of systematic errors.

6 Conclusion

Our study yields several key findings. First, no single universally optimal analyzer emerges, but tools show complementary strengths. While Stanza and Natasha achieve highest aggregate scores, performance varies significantly by genre and part of speech. Aggregate accuracy masks substantial persistent biases, with analyzers failing on every occurrence of certain high-frequency items. These biases can fundamentally distort linguistic tasks like frequency dictionary construction.

Second, disagreement analysis confirms the six tools have complementary error patterns, particularly for lemmatization, providing rationale for ensemble methods. The proposed POS-dependent weighted consensus statistically significantly outperforms all individual analyzers. By leveraging complementary biases, PWC achieves higher accuracy and greater robustness across text genres, providing more reliable foundation for linguistic research.

Third, the detailed error analysis provides a valuable roadmap for future improvements in individual tools, identifying specific architectural weaknesses or fine-tuning existing models.

Acknowledgements

The research was funded by the Russian Science Foundation (project No. 24-18-00570, https://rscf.ru/project/24-18-00570/).

References

[1] Brown T., Mann B., et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS). 10.48550/arXiv.2005.14165. — 2020.
[2] Bender E. M., Gebru T., et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT). — 2021. — P. 610–623.
[3] Kuzmenko E. Morphological Analysis for Russian: Integration and Comparison of Taggers // Communications in Computer and Information Science, Vol. 661. — 2017. — P. 162–171.
[4] Kotelnikov E., Razova E. et al. Close Look at Russian Morphological Parsers: Which One Is the Best? // Proceedings of the International Conference on Analysis of Images, Social Networks and Texts. — Springer, 2018. — P. 131–142.
[5] Trofimov I. Automatic Morphological Analysis for Russian: Application-Oriented Survey // Programmaya Ingeneria, Vol. 10. — 2019. — P. 391–399.
[6] Fenogenova A., Kazorin V. et al. Automatic Morphological Analysis on the Material of Russian Social Media Texts // EPiC Series in Language and Linguistics, Vol. 4. — 2019. — P. 11–17.
[7] Lyashevskaya O., Shavrina T. et al. GRAMEVAL 2020 shared task: Russian Full Morphology and Universal Dependencies Parsing // Computational Linguistics and Intellectual Technologies, Vol. 19. — 2020. — P. 553–569.
[8] Politsyna E., Politsyn S. et al. Analysis of Quality and Extension of Capabilities of Morphological Analysis Tools for Russian Texts // Vestnik VGU. Series: System Analysis and Information Technologies, Vol. 2. — 2023. — P. 171–180.
[9] Afanasev I., Glazkova A., et al. Rubic2: Ensemble Model for Russian Lemmatization // Proceedings of the 10th Workshop on Slavic Natural Language Processing. — ACL, 2025. — P. 157–170.
[10] Perišić A., Vanbelle S., et al. Quantifying Binary Classifier Algorithms Similarity with a Consensus Agreement Approach // Informatica, Vol. 36(3). — 2025. — P. 657–676.
[11] Shamayeva E. D. (2025), Comparison of Neural Network Syntactic Parsers for the Russian Language [Sravnenie neyrosetevykh sintaksicheskikh analizatorov dlya russkogo yazyka] // Computational Linguistics and Computational Ontologies [Komp'yuternaya lingvistika i vychislitel'nye ontologii], Vol. 9, pp. 26–47.
[12] Shavrina T., Belikov V. et al. (2015), A Corpus with Resolved Morphological Ambiguity: Towards a Methodology for Linguistic Research [Korpus s avtomaticheski snyatoy morfologicheskoy neodnoznachnostyu: k metodike lingvisticheskikh issledovaniy] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"].
[13] Touvron H., Lavril T., et al. LLaMA: Open and Efficient Foundation Language Models // 10.48550/arXiv.2302.13971. — 2023.
[14] Devlin J., Chang M.-W., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 10.48550/arXiv.1810.04805. — 2019.
[15] Schick T., Schütze H. Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference // 10.18653/v1/2021.eacl-main.20. — 2021. P. 255–269.
[16] Natasha [Electronic resource] // GitHub. — 2026. — URL: https://github.com/natasha/natasha (accessed: 28.02.2026).
[17] ru_core_news_sm [Electronic resource] // spaCy Models Documentation. — 2026. — URL: https://spacy.io/models/ru (accessed: 28.02.2026).
[18] MyStem [Electronic resource] // Yandex Developer. — 2026. — URL: https://yandex.ru/dev/mystem/ (accessed: 28.02.2026).
[19] DeepPavlov: an open source conversational AI framework [Electronic resource] // DeepPavlov. — 2026. — URL: https://deeppavlov.ai/ (accessed: 28.02.2026).
[20] Qi P., Zhang Y. et al. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. arXiv:2003.07082. — version 2. Access mode: https://arxiv.org/abs/2003.07082
[21] Straka M. UDPipe 2.0. // Proceedings of the CoNLL 2018 Shared Task. — ACL, 2018. — P. 197–207. Access mode: https://dialogue-evaluation.ru/competitions/12.
[22] GRAMEVAL 2020 Shared Task on Morphological Analysis and Lemmatization for Russian [Electronic resource] // Dialogue Evaluation. — 2020. — URL: https://dialogue-evaluation.ru/competitions/12 (accessed: 28.02.2026).

Appendix A. POS accuracy by part of speech for all analyzers (SynTagRus dataset).

POS	deeppavlov	natasha	pymystem	spacy	stanza	udpipe
ADJ	0.837	0.928	0.876	0.895	0.95	0.935
ADP	0.991	0.988	0.996	0.99	0.988	0.99
ADV	0.941	0.928	0.938	0.934	0.953	0.95
AUX	0.667	0.944	0.5	0.789	0.922	0.933
CCONJ	0.977	0.97	-	0.98	0.98	0.983
DET	0.843	0.914	0.9	0.879	0.943	0.907
INTJ	0.583	0.5	0.667	0.333	0.75	0.583
NOUN	0.938	0.971	0.981	0.956	0.96	0.954
NUM	0.901	0.963	0.667	0.895	0.969	0.988
PART	0.986	0.945	0.936	0.95	0.968	0.964
PRON	0.811	0.948	0.923	0.807	0.995	0.98
PROPN	0.85	0.908	0.827	0.889	0.922	0.922
SCONJ	0.962	0.962	-	0.97	0.939	0.962
VERB	0.935	0.941	0.802	0.939	0.954	0.923
X	0.676	0.765	0.971	0.824	0.647	0.647

Appendix B: Most frequent and recurring error clusters for each analyzer.

Analyzer	Dominant error clusters	Additional recurring errors	Notes
Natasha	DET→ADJ, DET→PRON, VERB→NOUN (5.63%)	Proper nouns, short adjectives, function words	Strong performance on frequent in-context forms; reduced robustness for ambiguous categories and long-tail phenomena
SpaCy	Pronoun lemmatization failures (~systematic), DET confusions (~23%)	Comparative forms, foreign words, short adjectives	Performance influenced by limitations of the rule-based lemmatization component, particularly for inflectional variation
Stanza	DET confusion (27.37%)	PRON→DET (3.21%), NOUN↔PROPN	High overall accuracy with remaining challenges in function-word disambiguation
UDPipe	DET↔PRON (23.0%)	ADJ→NUM (8.55%), NOUN↔PROPN	Balanced overall performance with systematic errors in ambiguous and borderline grammatical categories
DeepPavlov	Pronoun lemmatization (near-systematic), DET confusion (26.77%)	ADJ→NUM (10.24%), AUX→PART (3.47%)	Strong contextual modelling with persistent errors in closed-class words and lemmatization of high-frequency forms
pymystem3	PROPN→NOUN (13.39%), NUM↔ADJ (15.70%)	Verb forms, function words	Stable performance for in-lexicon items with reduced coverage of out-of-vocabulary and morphologically complex forms