Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2019)

Выпуск 18

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2019)

Issue 18

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 29 мая — 1 июня 2019 г.). Вып. 18 (25), 2019.

Сборник включает 61 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2019», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

18-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 25-й международной конференции «Диалог». На основании мнений нашего рецензентского корпуса для публикации в ежегоднике редколлегией был отобран 61 доклад из ста работ, которые были приняты к представлению на конференции в 2019 году.

Работы в сборнике отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, перевод, поиск, саммаризация, генерация, анализ тональности и т. д.)
- Глубокое обучение в NLP (методики применения, содержательная интерпретация)
- Компьютерный анализ Social Media
- Корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов)
- Лингвистический анализ текста (морфология, синтаксис, семантика)
- Лингвистические онтологии и автоматическое извлечение знаний
- Мультимодальная коммуникация (включая лингвистический анализ речи)
- Модели общения и диалоговые агенты
- Компьютерная лексикография

В соответствии с традициями «Диалога», старейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка.

Одно из ключевых событий «Диалога» — подведение итогов технологических соревнований между разработчиками систем лингвистического анализа текстов, Dialogue Evaluation. В этом году состоялись четыре соревнования:

- автоматическая генерация заголовков новостей;
- автоматический анализ малоресурсных языков (для которых очень мало данных для машинного обучения);
- автоматическое разрешение анафоры и определение референциальных цепочек (различных упоминаний одного и того же объекта в тексте),
- автоматическое восстановление слов по контексту (гэппинг-эллипсис).

В сборник включены наиболее оригинальные работы участников этих соревнований.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге, поскольку бумажный вариант пользуется все меньшей популярностью. Сборник, как и в прошлые годы, размещается на сайте конференции и индексируется Scopus.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Филологический факультет МГУ
- Школа прикладной математики и информатики МФТИ

Международный программный комитет

Богуславский Игорь Михайлович	ИППИ РАН, Россия; Мадридский политехнический университет, Испания
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Райгородский Андрей Михайлович	МФТИ, Школа прикладной математики и информатики, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, МФТИ, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	МГУ им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Пиперски Александр Чедович	РГГУ
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	МГУ им. М. В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Родионова Ольга Игоревна, *координатор оргкомитета*	Компания ABBYY
Ульянова Анна Вячеславовна, *секретарь оргкомитета*	РГГУ

Рецензенты

Tania Avgustinova
Vladimir Benko
Anatoly Gersman
Diana Macartney
Preslav Nakov
Piek Vossen
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Артемова (Черняк) Екатерина Леонидовна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Добрынин Владимир Юрьевич
Дроганова Кира Андреевна
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Кутузов Андрей
Левонтина Ирина Борисовна
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Марков Александр Юрьевич
Мисюрев Алексей Владимирович
Недолужко Анна Юрьевна
Новицкий Валерий Игоревич
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Пивоварова Лидия
Пиперски Александр Чедович
Подлесская Вера Исааковна
Смирнов Иван Валентинович
Смуров Иван Михайлович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Сорокин Алексей Андреевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович

Contents*

* The papers are ordered by the surname of the first author in compliance with the English alphabet.

A REUSABLE TAGSET FOR THE MORPHOLOGICALLY RICH LANGUAGE IN CHANGE: A CASE OF MIDDLE RUSSIAN[1]

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

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

The paper discusses the standardization efforts to create a morphological standard for the Middle Russian corpus, which is part of the historical collection of the Russian National Corpus (RNC). To meet the needs of different categories of corpus researchers as well as NLP developers, we consider two styles of the morphological annotation (RNC schema and Universal Dependencies schema). A number of specifications of the feature list proposed to facilitate data reusability, linking and conversion.

Key words: full morphology tagging, pos-tagging, lemmatization, tagset, historical corpora, Russian National Corpus, Universal Dependencies, Old Russian, Middle Russian

МНОГОЦЕЛЕВОЙ МОРФОЛОГИЧЕСКИЙ СТАНДАРТ РАЗМЕТКИ ДЛЯ ЯЗЫКА С МЕНЯЮЩЕЙСЯ ГРАММАТИЧЕСКОЙ СТРУКТУРОЙ: СЛУЧАЙ СТАРОРУССКОГО КОРПУСА

Ляшевская О. Н. (olesar@yandex.ru)

Национальный исследовательский университет «Высшая школа экономики»; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Статья посвящена созданию морфологического стандарта для разметки Старорусского корпуса, который входит в состав исторических корпусов Национального корпуса русского языка (НКРЯ). Для того,

[1] The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

чтобы сделать разметку удобной для лингвистов, работающих с историческими и современными корпусами, а также для разработчиков систем автоматической обработки исторических текстов, мы предусматриваем две параллельные схемы морфологической разметки, в нотации НКРЯ и Универсальных зависимостей (Universal Dependencies). Предлагается ряд спецификаций тагсета для облегчения совмещения разметок разных корпусов, связывания и конвертирования данных.

Ключевые слова: лексико-грамматическая разметка, частеречная разметка, лемматизация, тагсет, исторические корпуса, Национальный корпус русского языка, древнерусский язык, старорусская письменность

1. Introduction

Middle Russian Corpus (MidRus) is part of the Russian National Corpus (http://ruscorpora.ru) included in the collection of historical corpora [Sichinava 2014]. The MidRus contains over 4,700 texts of different genres written mostly between 1,300 and 1,700 (over 7 million words). Up to now, only a simple search for word forms and their parts has been available in the corpus interface. The paper represents the first attempt to develop the full morphology annotation standard for the MidRus.

Tagging the parts of speech, inflectional grammatical categories, and lemmas in historical corpora is a challenging task, since from one period to another, the grammatical structure changes: some grammatical forms drop out of use whereas new categories and grammatical patterns appear, the structure of the intra- and interparadigmatic homonymy varies. Furthermore, grammar and lexicon varies across schools and manuscripts, the texts often have noticeable dialect and stylistic features as well as varying and unstable spelling. While developing the full morphology annotation of the MidRus, we take into account the academic interests of the different categories of users including:

- researchers in the Middle Russian period of the language;
- researchers of the older periods of Russian accustomed to the annotation schemas of the Old Russian RNC corpus (OldRus) and the Old Novgorodian/East Slavic birchbark letters RNC corpus (OldNovg);
- researchers of the modern language who are interested in the micro-diachrony studies and are used to the tagset of the RNC Main corpus (ModernRus);
- NLP researchers who would be likely to use the Middle Russian data in their computational experiments, including comparative ones based on various paleoslavic data collections.

What makes things more challenging is that the annotation standards for the corpora of the earlier period and the modern period of Russian are well established but differ with regard to the lists of tags, the boundaries of lexical classes to which they apply, attested combinations of tags representing particular grammatical forms, and lemmatization rules. Therefore, we need to adopt existing schemas while evaluating contradicting data and clarifying the boundaries of the phenomena.

The last, but not the least issue that deserves attention is data reusability and customization. In recent years, new cross-language standards have gained popularity in NLP as they allow one to accumulate data of different origin and reuse and deploy the language technologies developed in the community.

To meet these new trends, the morphological annotation standard of the MidRus adopts two tagsets in parallel:

- RNC-MidRus: RNC Middle Russian tagset close to those of the Main RNC corpus, Old Russian, and Old Novgorod corpora;
- UD-MidRus: Universal Dependencies (UD) tagset close to those of the UD-Church Slavic and UD-Russian data collections.

As for the tagset customization, we distinguish among the core annotation schema (RNC and UD), an extended schema (RNC-ext and UD-ext), and a simplified schema encompassing only a selection of tags shared by the UD-MidRus and other UD corpora (UD-s).

The paper is structured as follows. **Section 2** outlines the state of the art in the field of historical Russian corpora and available NLP technologies. **Section 3** focuses on the part-of-speech tagging, **Section 4** covers the core grammatical tags, and **Section 5** is devoted to the analytical forms. The optional tags, extended and simplified annotation schemas are discussed in **Sections 6**, **7**, and **8**, respectively. Unless otherwise stated, the paper will refer to the core annotation schema, and the UD tags will be explicitly marked UD, if needed.

2. Historical Russian corpora and tagging methods

In this section, we overview the known historical corpora for Russian and methods for their tagging. Apart from the MidRus, there are three diachronic corpora in the RNC: OldRus, OldNovg, and Church Slavic (ChurchSlav) corpus [Moldovan 2015]. The Old Russian corpus [Mishina, Pichkhadze 2015] is provided with manual lemmatization and morphological annotation. The tool Morphy [Arkhangelsky et al. 2014] suggests annotations known from the texts which were tagged before. The original (Russian) tags are then translated into the (latin) tags used by the RNC search engine. The tagsets of the OldNovg [Sichinava 2018] and ChurchSlav [Dobrushina et al. 2015] are similar to those used in OldRus but differ in details. The annotation of the OldNovg is done semi-manually whereas the ChurchSlav is tagged automatically. An additional annotation of ambiguous word boundaries, fragmented tokens and comments on possible interpretations is available in OldNovg and, to a lesser extent, in OldRus. Moreover, the analyses in the OldNovg are most theoretically motivated, since they are based on the foudational work by [Zaliznyak 2004].

The annotation of the Northern Russian hagiographic corpus SCAT [Alexeeva, Azarova 2013] is done manually and follows an in-house extension of the TEI schema [Alexeev 2011]. The annotation features labeling the declension types.

The web page of the Regensburg Russian Diachronic Corpus mentions a “best bet” method based on the output of three taggers: Regensburg Old Church Slavonic tagger, Regensburg Old Russian guesser, and the modern Russian model of TreeTagger.

[Meyer 2011] adds that the main source is the annotation projection from modern translations.

The corpus Manuscript [Baranov et al. 2007] is partially tagged using a sophisticated rule-based pipeline which is powered by the Old Russian grammatical dictionary, modern grammatical dictionary, and a dictionary of pseudo-units. The tool carries out lemmatization and provides normalized orthographic representations.

The TOROT treebank [Eckhoff, Berdičevskis 2015] is an Old Russian add-on to the PROIEL Old Church Slavonic (OCS) treebank, which uses the same annotation environment and tagset. The texts are tagged manually, lemmas and annotations being provided with the aid of statistical preprocessing [Berdičevskis et al. 2016]. Currently, the data are released offline in MULTEXT-East XML format, and the PROIEL OSC data are also converted into the UD-CONLL format (the Old Russian TOROT data are planned to be released in UD in 2019).

To sum up, the morphological tagsets for many corpora described above are hardly available (see also detailed reviews in [Mitrenina 2014], [Eckhoff forthc.]). The most popular tagset is RNC (which exists in a few slightly different versions); MULTEXT-East and UD schemas are most accessible for NLP purposes due to the open license of the TOROT data.

Among the tagging methods, labeling by precedents, dictionary- and rule-based systems, and the projection of the modern Russian annotations are widely used. However, remarkably, other methods pave the way for the statistical learning: [Berdičevskis et al. 2016] compares the output of the HMM-based probabilistic tagger TnT and a hybrid system that makes use of the grammatical dictionary. [Scherrer et al. 2018] run computational experiments using conditional random fields method (CRF, tagger MarMoT) and deep neural network learning (char-embedding BLSTM). It is worth noting that since the amount of machine readable data is very modest and the historical data do not have a homogeneous structure with respect to their tagsets, this could potentially foster the interest of NLP developers to the material. Thus, the harmonization of data annotation is obviously crucial for improving the quality of tagging.

3. Parts of speech

The lists of part-of-speech (pos) tags and core grammatical features is available at: https://github.com/olesar/UD_MidRussian/blob/master/MidRussianUD.md. The document also reports the mapping between the RNC and UD tags. To evaluate the mismatches in the corpus annotation practice, we compared all attested combinations of pos-tags and features as well as their association with lemmas (lexical coverage) in OldRus, OldNovg, TOROT, UD-Church Slavic, and ModernRus.

In general, the RNC pos-list can be mapped to the UD UPOS list almost straightforwardly. The pos-tags for adjectives (A), ordinal numerals (ANUM), and the most part of predicative words (PRAEDIC, see below) are mapped to ADJ in UD; the pos-tags for adverbs and parenthetic words (ADV, ADVPRO, PARENTH) are mapped to ADV in UD. The noun tags (S in RNC) are mapped to NOUN (common nouns) and PROPN (proper nouns), and the verb tags (V in RNC) are splitted between VERB and AUX (auxiliaries)

in UD. The RNC tag CONJ is splitted between CCONJ (coordinate conjunction) and SCONJ (subordinate conjunction). The non-words (NONLEX) are splitted into X (foreign words, unknown words) and SYM (symbols). Besides that, the punctuation marks are explicitly tagged PUNCT in UD.

In the remainder of the section, we consider the mismatches in the annotation schemas with respect to the lexical coverage of pos categories in RNC and UD.

3.1. Pronominal words

И, е, я are tagged `spro` (UD: `pron`), the same way as in OldRus. Similarly, *иже, еже, яже* are tagged SPRO in RNC and PRON in UD. (In OldRus, they are tagged eigher APRO or SPRO, but we follow the principle to label a lemma uniformly as much as possible).

The relative pronouns *который*, *кыиждо*, *кыиже* are tagged APRO in RNC and PRON in UD. The reason is that they have the morphological properties of an adjective and the syntactic properties of a noun (nominal head), and this solution has already been implemented in the modern Russian UD [Droganova et al. 2018].

The possessives *его*, *ея*, *ихъ*, etc. are tagged as the Genitive forms of *онъ*, *оно*, *она*, *онъ*, *они*: SPRO, `gen` (UD: PRON, `Case=Gen`). (In OldRus, they are tagged as the Genitive forms of *и*; in ModernRus, they are tagged as indeclinable adjectival pronominals *его*, *ее*, *их*).

The list of APRO (UD: DET) includes:

- interrogative, relative, negative adjectival pronouns, quantifiers: *каковый*, *никакий*, *вьсь*;
- deictic (demonstrative) words: *сей*, *овъ*, *таковый*, etc.;
- possessive adjectival pronouns: *мой*, *свой,* etc.

The numeral *одинъ* is tagged ANUM in RNC and NUM in UD. In tagging it ANUM, we follow the practice of ModernRus (*один* has an adjective-like paradigm and is used as an attribute: for example, in the Nominative, it does not govern the Genitive case of the noun phrase compared to other numerals, see [Zaliznyak 2003]). However, in the UD treebanks the pos-tag NUM is applied consistently to the lexical equivalents of *один*. In OldRus, *одинъ* is labeled NUM as well.

3.2. Predicative words

Since there is no general mapping for the RNC PRAEDIC class to UPOS tags in UD, we use the conventions similar to those of the modern Russian UD standard:

- *-о*, *-е/-ь* forms (cf. (*ночью*) *тяпло*, *пригоже*, *явно*) that have corresponding adjectival forms are tagged as the short neutral forms of adjectives (UD: ADJ, `Gender=Neut, Number=Sing, Variant=Short`);
- the modal words—*можно*, *льзх*, *надобно*, *уне*—and the negative existentials *нхтъ*, *нх* are tagged VERB in UD;
- nouns such as *пора* used predicatively (cf. *пора идти*) are tagged as S in RNC and NOUN in UD;
- interjections, onomatopoeic words used predicatively are tagged as INTJ.

3.3. Auxiliaries

AUX in UD is used to tag:

- the auxiliary use of *быти*, *имѣти*, *хотѣти* in the analytical verb forms; this also includes the conditional markers *бы*, *бъ*—originally, the forms of *быти*, too, which got grammaticalized as indeclinable particles by the end of the Middle Russian period;
- the copula use of *быти* in nominal clauses;
- the reflexive markers (clitics) *си*, *ся*.

Only the existential and locative uses of the verb *быти* are tagged VERB in UD.

In the RNC schema, *бы* and *бъ* are subject of a double tagging strategy: they are labeled as verbs (lemma *быти*) and particles.

3.4. Named entities

The patronymics, last (family) names, nicknames and family nicknames and the like are tagged s (UD: PROPN): *Васильевичю*, *Колюбакинымъ*. This also applies to naming formulae with non-agreeing and agreeing possessive forms such as *Ивану Ильину сыну Челищева*, *Семену Васильеву сыну Власьеву*. The only exception are forms with full adjectival endings such as *Борисовую* in *княгиню Борисовую* and *Ондрѣевскую* in *Ефросинию*, *княж Ондрѣевскую жену Ивановича* which are considered adjectives (cf. the same practice in OldRus: *бабы* (*своеи*) *Романовои*). Note that in TOROT, the patronymics are sometimes considered adjectives.

4. Core grammatical tags

This section highlignts only key grammatical categories that distinguish the annotation schema of MidRus from those of OldRus or ModernRus.

4.1. Animacy

Animacy (`anim`; UD: `Animacy=Anim`) is tagged in the Accusative construction in which the form of Accusative is equal to the Genitive form, cf. *брата нашег*[*о*] *молодшег*[*о*]. In OldRus, such forms aree tagged `accgen`. The opposite case, when the Accusative case form is equal to the Nominative form, is not marked in MidRus.

4.2. L-form (indeclinable perfective participle)

L-participles (cf. *взялъ*) are tagged `perf` (UD: `VerbForm=PartRes, Tense=Past`), to distinguish them from other participles (cf. *взявъ*: `past partcp`; UD: `VerbForm=Part, Tense=Past`). The tense tag in UD will allow one to map the MidRus l-forms to the ModernRus past forms. L-forms are used both on their own and within the analytical forms, see below.

4.3. Gerundive (indeclinable adverbial participle)

Following [Zaliznyak 2004], forms such as *уповая*, *слышев* are considered indeclinable gerundives: `ger` (UD: `VerbForm=Conv`).

5. Analytical forms

The analytical forms are annotated as two (or more) tokens cross-linked at the morphological (in OldRus) and syntactic (in UD) level. All tokens are tagged `analyt` (UD: `Analyt=Yes`) and the grammatical features of the analytical form as a whole are labeled on the content word, cf. the annotation of the clause (1) *а будет не дошла* 'And if it won't reach (you)' in RNC (Fig. 1) and UD (Fig. 2).

(1)
```
а       <ana lex="а" gr="CONJ"></ana>
будет   <ana lex="быти" gr="V,3p,act,analyt,fut,indic,intran,sg" gr_ext="IN:FUT2+3312"></ana>
не      <ana lex="не" gr="PART"></ana>
дошла   <ana lex="доити" gr="V,act,analyt,f,fut2,intran,perf,pf,sg" gr_ext="IN:FUT2+3310"></ana>
```

Figure 1: A sample annotation in RNC-MidRus

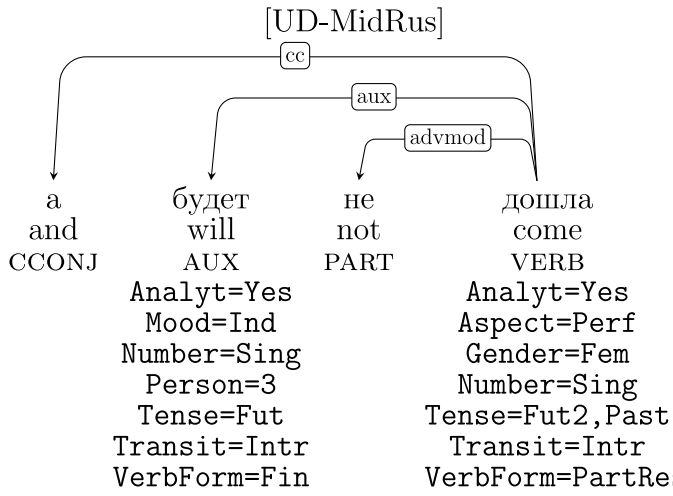

Figure 2: A sample annotation in UD-MidRus

In example (1), number, person, and future tense are labeled on the auxiliary *будет*: `sg`, `3p`, `fut` (UD: `Number=Sing`, `Person=3`, `Tense=Fut`), and gender, number, l-form are labeled on the content verb *дошла*: `f`, `sg`, `perf` (UD: `Gender=Fem`, `Number=Sing`, `Tense=Past`, `VerbForm=PartRes`): these are the intrinsic grammatical values of the tokens. The content word is also labeled by the tense of the whole analytical form `fut2` (UD: `Tense=Fut2`). Furthermore, *будет* is tagged AUX (part of speech) and `aux` (dependency relation) in UD.

The list of analytical forms includes:

- analytical future (new form, attested starting from the 1600s): infinitive + the future form of *быти* (*буду*, *будешь*), cf. *буду просить*: in new future forms, the content verb is tagged `fut` (UD: `Tense=Fut`);
- future 1: infinitive + the auxiliary nonpast forms of *хотѣти* and *имѣти*, cf. имет *обидѣти*: in: in the future 1 forms, the content verb is tagged `fut1` (UD: `Tense=Fut1`);
- future 2: l-form + the future form of *быти* (*буду*, *будешь*), cf. *боудеш*[ь] *послал*, *боудоу задѣла*: in the future 2 forms, the content verb is tagged `fut2` (UD: `Tense=Fut2`). Note that in OldRus, the analytical forms with почати, *начати*, *учати*, *стати*, *яти* are also labeled as the future 1 or future 2 forms, but we do not consider them as such in MidRus;
- analytical perfect: l-form and the 1st and 2nd person auxiliary in the present tense (*есмь*, *еси*, etc.), cf. *взял еси*;
- pluperfect (plusquamperfect): l-form + the perfect form of *быти*, cf. *дал еси был*, the content verb is tagged `pqperf` (UD: `Tense=Pqp`);
- subjunctive (conditional): l-form + *бы*, *бъ*, other aorist forms of *быти* (or conjunctions that incorporate *-бы*: *чтоб*(*ы*), *абы*, etc.), cf. *я бы сталъ*, *чтоб онъ пожаловалъ*: in conditional forms, the content verb is tagged `cond` (UD: `Mood=Cnd`);
- subjunctive (conditional) 2: l-form + *бы еси*, *бы есте* (2nd person forms of *быти*), cf. *держали бы есте* (*веру христианскую*): in conditional 2, the content verb *бы* is tagged `cond2` (UD: `Mood=Cnd2`).

The optative construction (*да* + non-past), the periphrastic comparative constructions of adjectives and adverbs are not considered analytical forms, nevertheless, they can be labeled with specific optional tags.

6. Optional tags

6.1. Features not available in automatic annotation

The following categories used of OldRus and OldNovg can be identified only in a particular context, often with the assistance of encyclopedic knowledge. In MidRus, they are used optionally in manual annotation:

- `as _ S` (UD: `AdjType=Subst`)—substantivized use;
- `as _ persn` (UD: `AdjType=Persn`, `NounType=Persn`)—used as a personal name. In particular, old nicknames such as *Мономахъ* are not counted as the last names and tagged `as _ persn`;
- `as _ topn` (UD: `NounType=Topon`)—used as a toponym;
- `as _ ethn` (UD: `NounType=Ethn`)—used as an ethnonym;
- `as _ ADV` (UD: `NounType=Adv`, `AdjType=Adv`)—used as an adverb, cf. (*придоша Ветроу*) *вечеръ*, (*по*) *готовоу*;
- `as _ PART` (UD: `VerbType=Part`)—used as a particle, cf. *хотя*;

- `as _ PARENTH` (UD: `AdvType=Parenth`; pos-tag PARENTH in RNC-ext, see below)—parenthetical use;
- `as _ PRAEDIC` (UD: `AdjType=Praedic`; pos-tag PRAEDIC in RNC-ext)—predicative use;
- `as _ deb` (UD: `VerbType=Debit`)—used as a debitive, cf. *да не погубиши мьзды своеа*.

The following tags are used optionally and only in the RNC-style annotation:

- `husbn`—distinguishes the name given by husband's name from patronymics, cf. OldNovg (*оу*) *тоудоровъи*;
- `in _ persn`—used within a personal name, cf. *анастасу корсунянину*;
- `in _ ethn`—used within an ethnonym, cf. *Черніи Клобуци*;
- `in _ topn`—used within a geographic name, cf. (в) *Константинъ градъ*;
- `in _ ADV`—used in an adverbial phrase, cf. *тако же*;
- `in _ NUM`—used within a complex numeral, cf. *двѣма на десяте*;
- `in _ CONJ`—used within a multitoken conjunction, cf. *егда како*;
- `in _ PR`—used within a multitoken preposition, cf. *в мѣсто*.

In UD, there are ways to encode most of such cases with the dependency relation tags (e. g. `flat:name` and `fixed`).

6.2. Spelling and non-standard variants

The feature `abbr` (UD: `Abbr=Yes`) is used to tag abbreviated words including those marked by titlo.

- The feature `ciph` is used in RNC schema to label cardinal and ordinal numerals expressed by (Euro-Arabic) digits and Cyrillic letters. In UD-MidRus, the corresponding tag `NumForm=Digit` is used to label cardinal and ordinal numerals expressed by (Euro-Arabic) digits (*за 5 верстъ, 5-ти дней, лета 7030-го июля в 9 день*);
- `NumForm=Cyril`—used to tag numerals expressed by Cyrillic letters (*КЕ ал, по Д чысло*);
- `NumForm=Word`—used to tag numerals expressed by words (*одинъ, первый, лѣта семь тысячь девятаго*).
- The feature `distort` (UD: `Typo=Yes`) is used to label distorted words and words guessed by the editors of the historical manuscrupts. Specific cases include (RNC-style only):
- `damaged`—guessed words (if the text segment is damaged);
- `crossed _ out`—crossed out, cf. OldRus: (*и ко полотьску*)
- `redundant`—redundant word (*не не сподобилъ же еси*). Note that in UD, the feature `Echo` can be used to label various kinds of repetitions.

The feature `anom` (not tagged in UD) is used to tag grammatically anomalous forms. However, what is considered 'grammatically anomalous' in the historical data is controversial and theory-specific. Therefore, this tag should be used with caution.

Finally, `oov` (cf. `bastard` in ModernRus, not tagged in UD) is a specific kind of tags which is used to label words not seen in the training data or the grammatical dictionary of the tagger.

7. Extended annotation schema

We introduce the notion of cross-features (or x-features) that can be added into the schema to make the annotations in different corpora comparable. For example, in micro-diachronic studies, the data of the modern language are compared against the historical data. Even if a certain grammatical category is under development and it is not evident if it is present or absent in the data, x-features allow one to look for the potentially interesting patterns. In the current Middle Russian standard RNC-ext, the x-features include:

- `anim$` and `inan$` (UD: `Animacy[lex]=Anim, Animacy[lex]=Inan`): classifying features that correspond to anim and inan in the ModernRus annotation. This category is not to be mixed with `anim` (UD: `Animacy=Anim`) that is applicable only to the Accusative constructions (see above). There are cases in which the lexically animated nouns (`anim$`) are not tagged as `anim`;
- the transitivity tags `tran` and `intr` (UD: Transit=Tran, `Transit=Intr`). The transitivity is tagged often inconsistently in modern corpora, and the situation is even worse in historical corpora. However, this is an interesting category under development that allows a user to study various morphosyntactic phenomena.

Another example is the use of cross-features to make the data conversion between different formats more straightforward. So, in the intermediate schema UD-ext, an extended list of parts of speech is used which includes ANUM, PRAEDIC, PARENTH. Further, a number of cross-features under the category NounType are introduced in UD-ext to reflect RNC tags such as `persn, patrn, famn, zoon, ethn, topon` (e.g. NounType=Ethn).

8. Simplified annotation schema

An alternative option to make data compatible is reducing the lists of tags. This is particularly useful in NLP evaluation tasks since the dominance of features carefully designed for human research but rarely attested in corpora can cause the drop in tagging performance. In order to make the tagsets of historical corpora available in UD (UD Church Slavic (UD-PROIEL), UD-TOROT and UD-MidRus) compatible, the following features can be excluded from annotation:

- `Aspect` (verb aspect)
- `Reflex` (reflexivity labeled on verbs and pronound)
- `Animacy` (Acc=Gen)
- `PronType` (pronominal type)
- `Variant` (long/short forms)
- `Strength` (a rough equivalen for Variant in UD-PROIEL/TOROT)

Except for Variant/Strength and Animacy, these features are lexical (classifying) and do not add to the identification of which paradigm cell the form fills. Obviously, extended and optional features are out of the simplified list as well.

In addition, the tense forms of aorist (`Tense=Aor`) and imperfect (`Tense=Imp`) should be relabeled as `Tense=Past` according to the universal UD guidelines (and thus mirroring the annotation in UD-PROIEL/TOROT).

9. Conclusion

We have presented the annotation standard for the Middle Russian corpus, detailing guidelines to the tagging of part-of-speech and morphological features in RNC and UD schemas and introducing a mapping between the RNC and UD tags. We distinguish between core, extended and simplified tagsets and show that different categories of users can benefit from them.

The annotation schemas were evaluated and corrected while doing the manual annotation of the MidRus gold standard [Lyashevskaya 2018], on the one hand, and carrying out computational experiments in automatic tagging and training data amplification [Scherrer et al. 2019], on the other hand. The test sample was annotated manually in both standards, RNC and UD, in parallel. After data conversion from RNC to UD-s, the inter-annotator agreement was calculated over a total of 400 tokens. The ratio of equivalent annotations was considerably high (95%).

A pilot version of the gold standard MidRus data is released with open license in Universal Dependencies, v2.4.

Acknowledgements

We are grateful to Irina Juryeva, Roman Ilushin, Maria Skachedubova, Elizaveta Bunina, and Dmitri Sitchinava who contributed to the annotation of the Middle Russian gold standard data and revision of the annotation guidelines. We would also like to thank Anna Pichhadze, Alexandr Moldovan, Vladimir Plungian, Roman Krivko, Yves Scherrer, Achim Rabus, Hanne Eckhoff for fruitful discussion and advice.

References

1. *Arkhangelsky T. A., Mishina E. A., Pichkhadze A. A.* (2003), A tool for the electronic grammatical annotation of Old Russian and Church Slavonic texts and its use in web resources [Sistema elektronnoj grammaticheskoj razmetki drevnerusskikh i tserkovnoslavjanskikh tekstov i jejo ispol'zovanie v veb-resursakh], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference, Sofia, Izhevsk, 2014.

2. *Alexeev V. A.* (2011), Expansion and implementation of the format for describing the grammatical and graphic data of the SKAT corpus [Rasshirenie i realizatsija formata opisanija grammaticheskikh i graficheskikh dannykh korpusa SKAT]. Master's thesis, St.-Petersburg, St.-Petersburg state university.
3. *Alekseeva E. L., Azarova I. V.* (2013), Peculiarities of the morpho-syntactic annotation for the Old Russian hagiographic texts [Osobennosti morfo-sintaksicheskoy razmetki drevnerusskikh agiograficheskikh tekstov], Proceedings of the International conference "Corpus linguistics-2013", St.-Petersburg, pp. 157–164.
4. *Baranov V. A., Mironov A. N., Lapin A. N. et al.* (2007), Automatic morphological analyzer of Old Russian language: linguistic and technological solutions [Avtomaticheskij morfologicheskij analizator drevnerusskogo jazyka: lingvistichekie i tekhnologicheskie reshenija] 10th jubilee international conference EVA 2007, Moscow.
5. *Berdičevskis A., Eckhoff H. M., Gavrilova T.* (2016), The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016", Moscow, pp. 99–111.
6. *Dobrushina E. R., Kravetsky A. G., Poljakov A. E.* (2015), A corpus and a frequency grammatical corpus-based dictionary of Church Slavonic in the collection of the Russian National Corpus [Korpus i chastotnyj grammaticheskij korpusnyj slovar' tserkovnoslavjanskogo jazyka v sostave Nacional'nogo korpusa russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
7. *Droganova K., Lyashevskaya O., Zeman D.* (2018), Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks Proceedings of the 17th IntGernational Workshop on Treebanks and Linguistic Theories (TLT 2018), Oslo, pp. 52–65.
8. *Eckhoff H. M.* (forthc.), Historical corpora and the re-evaluation of Slavonic language history.
9. *Eckhoff H. M., Berdičevskis A.* (2015), Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank, Scripta & e-Scripta, Vol. 14–15, pp. 9–25.
10. *Lyashevskaya O.* (2018), A test dataset for the automatic morphological analysis of the Middle Russian texts [Testovaja kollektsija dlja zadach avtomaticheskogo morfologicheskogo analiza tekstov starorusskoj pis'mennosti], The academic heritage of V. A. Bogoroditsky and the modern vector of research of the Kazan linguistic school [Nauchnoje nasledije V. A. Bogoroditskogo i sovremennyj vektor issledovanij Kazanskoj lingvisticheskoj shkoly], Works and materials of int. conf., Kazan: Kazan University, pp. 131–135.
11. *Meyer R.* (2011), New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations, Russian linguistics, Vol. 35 (2), pp. 267–281.
12. *Mishina E. A., Pichkhadze A. A.* (2015), Old Russian subcorpus of the Russian National Corpus [Drevnerusskij podkorpus Nacional'nogo korpusa russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
13. *Mitrenina O.* (2014), The corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language, Scrinium: Journal of Patrology, Critical Hagiography, and Ecclesiastical History, Vol. 10 (1), pp. 455–461.

14. *Moldovan A. M.* (2015), Old Russian manuscripts in the Russian National Corpus [Pamjatniki drevnerusskoj pis'mennosti v Natsional'nom korpuse russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
15. *Nivre J., De Marneffe M. C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R. T., Petrov S., Pyysalo S., Silveira N., Tsarfaty, R.* (2016), Universal Dependencies v1: A Multilingual Treebank Collection, Proceedings of LREC 2016.
16. *Nivre J., Abrams M., Agić Ž. et al.* (2018), Universal Dependencies 2.3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2895.
17. *Polyakov A. E.* (2012), A stemmer for the pre-reform Russian orthography [Lemmatizator dlja doreformennoj russkoj orfografii], Baranov V. A., Varfolomeyev A. G. (eds.), Proceedings of the international conference Information Technologies and Textual Heritage El'Manuscript-12 [Informatsionnye tekhnologii i pis'mennoe nasledie: materialy IV mezhdunarodnoj nauchnoj konferentsii], Petrozavodsk, Izhevsk, pp. 211–215.
18. *Sichinava D. V.* (2014), Historical corpora of the Russian National Corpus as a tool for diachronic grammatical studies [Istoricheskie korpusa Natsional'nogo korpusa russkogo jazyka kak instrument diakhronicheskikh issledovanij grammatiki], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference. Sofia, Izhevsk, 2014.
19. *Scherrer Y., Rabus A.* (2019), Variation in pre-modern Slavic corpus data and accuracy of neural tagging, Proceedings of the conference "Historical Corpora and Variation", Cagliari, 2019.
20. *Sichinava D. V.* (2014), Historical corpora of the Russian National Corpus as a tool for diachronic grammatical studies [Istoricheskie korpusa Natsional'nogo korpusa russkogo jazyka kak instrument diakhronicheskikh issledovanij grammatiki], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference. Sofia, Izhevsk, 2014.
21. *Sichinava D. V.* (2018), The corpus/database of Old East Slavic birchbark letters, El'Manuscript 2018 Book of Abstracts, Vienna, Krems.
22. *Zaliznyak, A. A.* (2003), A Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo jazyka], Moscow.
23. *Zaliznyak, A. A.* (2004), Old Novgorod Dialect, Moscow, Languages of Slavonic Culture.

that it instantiates obligatory adjunct control by the subject. We hypothesize that the rise of the NACC is driven by the analogy with the existing constructions with EXASPERATE-verbs in standard Russian, and we address several other factors that contribute to the development of the new construction.

AUTOMATIC VOCABULARY POSITIONING IN A THESAURUS

Likhonosov A. (andrew.likhonosov@abbyy.com), **Indenbom E.** (Eugene_I@abbyy.com), **Yudina M.** (maria_yu@abbyy.com), ABBYY, Moscow, Russia

Thesauri are one of the most widely used resources in natural language processing. At the same time, many of them are built manually, which takes a lot of time and, due to human errors, can affect their quality and completeness. We propose a procedure for automatic positioning of vocabulary in the ABBYY Compreno thesaurus using large monolingual corpora, a regular bilingual dictionary and a subset of already positioned words.

ANALYSIS OF PROSODIC FEATURES OF THE EMOTIONAL INTONATION USING "INTONTRAINER" SYSTEM (ON THE EXAMPLE OF RUSSIAN PHRASES)

Lobanov B. M. (Lobanov@newman.bas-net.by), **Zhitko V. A.** (zhitko.vladimir@gmail.com), United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The main results of the update of the IntonTrainer system for the purposes of analyzing and studying the prosodic signs of emotional intonation are described. A distinctive functional feature of the updated system is the creation of an expanded set of prosodic signs of emotional intonation. The paper presents preliminary assessments of their effectiveness using the created experimental database of emotional phrases of Russian speech.

A REUSABLE TAGSET FOR THE MORPHOLOGICALLY RICH LANGUAGE IN CHANGE: A CASE OF MIDDLE RUSSIAN

Lyashevskaya O. N. (olesar@yandex.ru), National Research University Higher School of Economics; Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper discusses the standardization efforts to create a morphological standard for the Middle Russian corpus, which is part of the historical collection of the Russian National Corpus (RNC). To meet the needs of different categories of corpus researchers as well as NLP developers, we consider two styles of the morphological annotation (RNC schema and Universal Dependencies schema). A number of specifications of the feature list proposed to facilitate data reusability, linking and conversion.

POSTPOSITIONAL CONSTRUCTIONS IN TATAR: METHODOLOGIES FOR MEASURING INTRALINGUAL VARIATION

Lyutikova E. A., MSU, MPSU, Pushkin State Russian Language Institute,
Gerasimova A. A., MSU, MPSU, Pushkin State Russian Language Institute

The paper addresses the issue of intralingual variation in Tatar postpositional phrases. The nominal in Tatar postpositional phrases demonstrates differential case marking: the choice between genitive and unmarked case form is determined by the morphosyntactic class of the nominal. With postpositions derived from nouns with locative or abstract semantics variation in case assignment is accompanied by presence/absence of the ezafe marker on the postposition. In this paper we use corpus-based and experimental methods to investigate the distribution of grammatical variants and estimate the current status of the variation. We argue that the existing grammatical descriptions do not capture the current state of affairs.

We show that pronouns and nouns do not form a homogeneous class with respect to case marking in the postpositional phrase. The genitive case marking is common for 1st/2nd person personal pronouns and 3rd person singular personal pronoun. All other pronouns and nouns are primarily used in an unmarked form, an observation supported by both corpus and experimental data.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 18 (25). 2019

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06