ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ т. 10, № 4(43), с. 181–199

ББК Ш111:3813
ГРНТИ 16.31.21, 28.23.13
УДК 004.89:81'322.2

Н. А. Власова, И. В. Трофимов, Ю. П. Сердюк,
Е. А. Сулейманова, И. Н. Воздвиженский

PaRuS — синтаксически аннотированный корпус русского языка

Аннотация. В статье представлен новый аннотированный корпус русского языка PaRuS (Parsed Russian Sentences). Корпус имеет объем свыше 2,5 миллиардов токенов и предназначен для решения задач компьютерной лингвистики методами машинного обучения. PaRuS состоит из предложений русского литературного языка. Каждое предложение снабжено лингвистической разметкой: морфологической в формате MULTEXT-East и синтаксической в нотации СинТагРус. В статье рассмотрена методология создания корпуса, описан гибридный лингвистический конвейер PaRuS_pipe, разработанный для порождения разметки. Обсуждаются вопросы качества аннотирования языкового материала в корпусе PaRuS, выполнена оценка морфологического анализатора конвейера PaRuS_pipe по методологии соревнования MorphoRuEval-2017.

Ключевые слова и фразы: компьютерная лингвистика, корпусная лингвистика, русский язык, языковой корпус, разметка, морфология, синтаксис.

Введение

Всё более широкое распространение методов машинного обучения в современной компьютерной лингвистике порождает потребность в больших объемах аннотированного в том или ином аспекте текстового материала. Лингвистика «моделей» в разработках компьютерных систем постепенно уступает место лингвистике размеченных данных [1].

В период доминирования «модельных подходов» аннотированные данные служили главным образом для исследовательских целей. Размечавшиеся вручную небольшие корпуса в полной мере отвечали этой потребности. Сейчас, когда основой компьютерной лингвистики

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00779.

10.25209/2079-3316-2019-10-4-181-199

стали обучаемые алгоритмы, объем и репрезентативность корпуса становятся ключевыми факторами, определяющими качество решения задачи.

Корпуса большого объема создаются автоматически и полуавтоматически. Качество разметки в корпусе зависит от методологии его создания. Вопрос о том, какой должна быть эта методология, — отдельный научный вопрос, не имеющий очевидного ответа. Как подобрать и подготовить текстовый материал, чтобы имеющийся арсенал средств автоматического аннотирования позволил получить приемлемый результат? Как выявить и устранить систематические ошибки аналитических средств? Как обеспечить верификацию, одновременно минимизируя трудоемкость? Сложность такого рода проблем растет вместе с увеличением объема корпуса.

Еще один фактор, во многом определяющий свойства создаваемого корпуса, заключается в выборе — для каждого уровня разметки — теоретической базы, на основе которой будет выполняться аннотирование. Пока этот выбор не очевиден, имеет смысл создавать корпуса на основе разных теоретических принципов.

Таким образом, для современной компьютерной лингвистики, ориентированной на машинное обучение, актуальны:

(*1*) создание корпусов большого объема, различающихся лежащими в основе разметки теоретическими подходами;

(*2*) поиск эффективных методов создания больших корпусов.

В данной работе описан опыт создания большого русскоязычного корпуса `PaRuS` с морфологической и синтаксической разметкой. Корпус состоит из предложений русского литературного языка. Объем — свыше 2,5 миллиардов токенов (более 150 миллионов предложений). Синтаксическая разметка корпуса выполнена в нотации `СинТагРус`. Лежащий в ее основе формализм создан ведущей отечественной школой теоретической и компьютерной лингвистики в рамках проекта `ЭТАП` [2].

Дальнейший материал изложен в следующем порядке. В первой части статьи приводится краткий обзор аннотированных корпусов русского языка, вторая посвящена особенностям корпуса `PaRuS`, в третьей описан метод его создания.

1. Аннотированные корпуса русского языка

Требованию большого объема в настоящий момент удовлетворяют несколько корпусов русского языка: `Araneum Russicum` [3, 4], `RuTenTen` [5], `ГИКРЯ` [6], `Taiga` [7, 8]. В каждом из перечисленных корпусов присутствует морфологическая разметка.

Синтаксическая разметка в корпусах русского языка представлена далеко не так широко, как морфологическая. Между тем, для обучаемых алгоритмов синтаксическая информация может оказаться полезной. Так, например, первый корпус русского языка с синтаксической разметкой СинТагРус [9], входящий в состав Национального корпуса русского языка [10], несмотря на скромный по современным меркам объем, успешно применялся для тренировки алгоритмов машинного обучения. В частности, на его материале обучена русскоязычная модель синтаксического парсера MaltParser [11]; решалась задача выявления границ сложных именных групп [12].

Первым русскоязычным корпусом большого объема с синтаксической разметкой стал корпус Taiga. Корпус аннотирован в формате Universal Dependencies (UD) [15] — международном универсальном формате морфологической и синтаксической разметки, который был разработан для нивелирования различий в метаязыках описания синтаксической структуры, используемых для типологически различных языков.

Ответ на вопрос, могут ли UD-корпуса служить полноценной заменой корпусам, аннотированным в соответствии с «традиционными» национальными схемами, по-видимому, зависит от решаемой задачи и требует соответствующих исследований на параллельных корпусах (как, например, эксперименты с обучением парсеров для шведского языка [16]).

Разметка в корпусе СинТагРус основана на системе синтаксических отношений, которая используется многоцелевым лингвистическим процессором ЭТАП [13, 14]. В нотации СинТагРус задействовано, по разным источникам, от 67 [17] до 78 [18] типов синтаксических отношений, тогда как формат UD использует всего 40 синтаксических тегов [17][1]. В качестве принципиального отличия между двумя подходами к синтаксической разметке мы бы отметили следующее: в UD тип зависимости определяется поверхностными признаками словоформ, тогда как нотация СинТагРус учитывает особенности модели управления подчиняющих слов[2].

[1] О соответствиях между двумя системами синтаксических отношений см. технический отчет K. Droganova, D. Zeman. *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (technical report)*. UFAL MFF UK, Praha, Czechia. 2016. http://ufal.mff.cuni.cz/techrep/tr60.pdf

[2] В частности, в UD не предусмотрена возможность различения актантных и атрибутивных зависимых имени, различаемых в СинТагРус; предложные группы, подчиненные глаголу, в UD оформляются одинаково независимо от их роли — актантной или обстоятельственной.

С учетом всего сказанного, задача создания большого, объемом в несколько миллиардов токенов, корпуса с разметкой в нотации СинТагРус представляется актуальной.

2. Корпус русского языка PaRuS

При создании корпуса PaRuS преследовались две ключевые цели:

(*1*) обеспечение качественной лингвистической разметки (морфология и синтаксис);

(*2*) достижение объемов, отвечающих потребностям современных обучаемых аналитических алгоритмов.

Этим обусловлен ряд принятых при создании корпуса специфических методологических решений, речь о которых пойдет ниже.

Единицей языкового материала корпуса PaRuS является предложение. Предложение — это языковая единица, в рамках которой проявляются в полной мере морфологические и синтаксические свойства словоформ, а также семантика отдельных лексем и словосочетаний. Поэтому предложение можно использовать для получения полноценной информации об этих языковых единицах и соответствующих языковых уровнях.

Такое нестандартное для корпусной лингвистики решение имеет определенные преимущества по сравнению с традиционным подходом, когда единицей корпуса является текст. Во-первых, на стадии подготовки материалов для корпуса нет необходимости решать задачу определения границ текста. Это непросто, когда речь идет о комментариях к новостям, чатах в соцсетях, диалогах, электронной переписке — такие материалы всё чаще включаются в языковые корпуса для обеспечения репрезентативности.

Во-вторых, работа с предложениями позволяет использовать эффективные алгоритмы фильтрации языкового материала, не применимые к текстам. В частности, без последствий для целостности можно удалять из корпуса предложения, автоматическая обработка которых, вероятно, окажется неуспешной (подозрительно длинные токены, предложения на языках, отличных от русского, «странные» последовательности символов и т.п.).

В-третьих, отпадает необходимость рассмотрения вопросов, связанных с авторскими правами.

Недостатками выбора предложения в качестве единицы корпуса можно считать утрату метаданных и информации о структуре текста, потерю целостности дискурса. Однако для задач, которые

предполагается решать с помощью корпуса PaRuS, такие потери не существенны.

Предложения, составляющие корпус русского языка PaRuS, получены из двух типов источников:

(*а*) произведения художественной, научно-популярной, публицистической литературы, доступные онлайн;

(*б*) новостные сообщения с нескольких десятков новостных сайтов.

В PaRuS не включались предложения из блогов, социальных сетей, форумов, комментариев к новостям и т.п. Такие источники сложны для автоматического анализа, так как часто содержат символы и комбинации знаков, не применяющиеся в литературных текстах, предложения имеют специфическую синтаксическую структуру и пр. Кроме того, в таком языковом материале гораздо чаще, чем в редакционном, встречаются опечатки и ошибки, что требует дополнительных усилий по очистке. Таким образом, PaRuS состоит из предложений современного русского *литературного* языка.

Итак, корпус PaRuS — это большой морфологически и синтаксически аннотированный корпус предложений русского литературного языка. Объем корпуса в настоящий момент превышает 2,5 миллиардов токенов. Цель — более 5 миллиардов. PaRuS можно использовать для обучения алгоритмов, оперирующих в пределах предложения, таких как алгоритмы морфологического и синтаксического анализа, некоторые алгоритмы разрешения лексической неоднозначности, алгоритмы построения дистрибутивных моделей, обнаружения устойчивых словосочетаний и т.п. Данные и техническая документация корпуса PaRuS размещены по адресу https://parus-proj.github.io/PaRuS.

3. Метод создания корпуса PaRuS

В методе создания корпуса PaRuS выделяется три крупных группы операций:

(*1*) отбор и подготовка текстов;

(*2*) лингвистическое аннотирование (разметка);

(*3*) дедупликация, фильтрация и перемешивание предложений.

Рассмотрим подробнее каждую из них.

3.1. Отбор и подготовка текстов

Целью данного этапа обработки было получение материалов для корпуса в виде простых текстовых файлов в кодировке utf-8. Напомним,

что корпус составлялся из текстов двух категорий: художественной и нехудожественной литературы, далее *книг* (*а*) и новостных сообщений (*б*). Каждая из категорий потребовала особой технологии селекции и предварительной обработки текстов.

Книги

Источником литературных произведений послужили открытые онлайн-библиотеки. Загрузка документов не представляла сложности, поскольку многие библиотеки публикуют резервные копии в виде архивов. Дальнейшая обработка включала:

- техническую обработку (форматы файлов, кодировки, разбор метаданных);
- жанровую фильтрацию;
- фильтрацию иноязычной литературы.

Жанровая фильтрация выполнялась при наличии указания на жанр в метаданных книги. В частности, исключались поэтические, драматургические произведения, детские сказки, религиозная литература, книги в жанре «фэнтези», юмор. Драматургические произведения, в силу их специфического оформления, исключались с целью снижения доли ошибок в синтаксической разметке корпуса. Детские сказки, религиозная литература и фэнтези описывают вымышленные миры; в таких произведениях встречается множество лексических новообразований, словосочетаний с нехарактерными отношениями между словами (*фиолетовая листва*, *железная бумага* и т.п.), что нежелательно, например, для дистрибутивного моделирования. Юмористические тексты изобилуют игрой слов. Поэтической речи свойственны нестандартный порядок слов, непроективные синтаксические конструкции и другие особенности, с которыми современные автоматические анализаторы справляются плохо.

Для фильтрации иноязычной литературы потребовалось реализовать следующую многоступенчатую процедуру определения языка текста.

1. Для первичной фильтрации использовалась метаинформация о языке текста, если она присутствовала в документе.
2. Затем применялся классификатор на базе n-граммной статистики [**19**]. Реализация — утилита `mguesser`[3]. Утилита позволила отфильтровать значительную долю иноязычной литературы, а также некоторые многоязычные тексты (в частности, документы с большим количеством цитат на иностранных языках).

[3] `https://github.com/yaoweibin/mguesser`

3. Эксперименты показали, что n-граммного метода оказалось недостаточно. Так, `mguesser` иногда принимает тексты на белорусском, украинском, болгарском и других языках с кириллическим алфавитом за русскоязычные. Для решения этой проблемы был реализован специальный эвристический алгоритм, подсчитывающий в тексте количество слов, частотных в перечисленных языках, но отсутствующих в русском. Если число таких слов в тексте оказывалось выше установленного порога, то текст считался не русскоязычным и исключался из корпуса. Например, для определения текстов на украинском языке использовался поиск по таким высокочастотным в данном языке словам, как *ні*, *ці*, *від*, *вже*, *оце*, *щось*, *тоді*, *мені*, *тобі*, *він*, *щоб*, *його*, *якщо*. Для белорусского языка в качестве индикаторных слов были выбраны *ці*, *калі*, *зусім*, *яшчэ*, *вельмі*, *гэта*, *тады*, *надта*, *мяне*, *цябе*. Для казахского языка — *олар*, *быз*, *кайда*, *кашан*, *ондой*, *анау*, *онда*, *санда*, *сыз*, *казр*, *тагх*, *айт*. Аналогичный прием применялся для фильтрации текстов на русском языке в дореформенной орфографии, на древнерусском и церковнославянском.

Согласно грубой оценке, жанровая и языковая фильтрация привела к удалению более половины литературных произведений, загруженных из онлайн-библиотек.

Новостные сообщения

Новостные сайты, как правило, не размещают архивы новостных сообщений в расчете на их массовое скачивание. Поэтому подзадача загрузки новостных сообщений потребовала создания специализированного краулера. Мы воспользовались тем, что многие издания предоставляют доступ к новостной ленте в формате rss-каналов. Загрузка новостей осуществлялась небольшими (суточными) порциями в течение длительного промежутка времени (около 5 лет). Для отдельных сайтов использовались специальные процедуры коррекции URL новостного сообщения, чтобы получать не страницу новости в контексте новостного сайта, а содержащую меньше посторонней информации «версию для печати». Чтобы охватить больший исторический период, для одного информационного агентства был разработан специализированный краулер, позволивший загрузить новости, относящиеся к 1990–2015 годам.

В общей сложности таким мониторингом было охвачено более 100 новостных ресурсов (государственных и региональных, отечественных и переводных зарубежных, общетематических и специализированных).

Результатом загрузки новостей были html-страницы, которые кроме новостного сообщения содержали постороннюю информацию (комментарии, элементы навигации по сайту, рекламу и т.п.). Страницы необходимо было очистить. Извлечение текста новости из html-страницы осуществлялось алгоритмом `jusText`[4] [20]. Этот эвристический алгоритм имеет высокие показатели точности на наборах размеченных данных `CleanEval` и очень хорошие показатели полноты, не требует обучающего множества и прост в настройке. Мы использовали собственную реализацию алгоритма, которая, кроме очистки html-дерева, осуществляла конвертацию текста в кодировку utf-8.

3.2. Лингвистическое аннотирование

Для порождения лингвистической разметки корпуса был создан гибридный конвейер `PaRuS_pipe`[5], основанный на двух известных русскоязычных лингвистических процессорах: конвейере Шарова-Нивре [21] и `UDPipe` [22]. Новый конвейер решает следующие задачи:

- определение границ слов и предложений,
- морфологический анализ,
- синтаксический анализ.

Работая над `PaRuS_pipe`, мы проанализировали сильные и слабые стороны положенных в его основу конвейеров и с учетом этого попытались создать более эффективный аналитический инструментарий. Ведущая роль отводилась конвейеру Шарова-Нивре (далее `Ш-Н`), а `UDPipe` использовался как вспомогательное средство. В частности для определения границ предложений мы использовали `UDPipe`, так как он имеет более совершенный алгоритм распознавания сокращений. Качественное решение этой задачи важно для успеха последующего синтаксического анализа. В морфологическом анализе `UDPipe` лучше распознаёт ряд граммем (например, признак «имя собственное»), а также частично компенсирует неполноту `Ш-Н` в части лемматизации.

Также в `PaRuS_pipe` реализована группа дополнительных модулей, служащих для:

- адаптации исходного текста к формату, с которым базовые конвейеры (`Ш-Н` и `UDPipe`) работают более успешно;
- коррекции в том или ином аспекте результатов базовых конвейеров.

[4] `https://code.google.com/archive/p/justext/`
[5] `https://hub.docker.com/r/parusproj/parus`

Для повышения качества лемматизации был задействован словарный морфологический анализатор AOT [23], а также ряд эвристик для обработки слов, которые пишутся через дефис.

Структурная схема получившегося конвейера изображена на рисунке 1. Более подробное техническое описание каждого из модулей можно найти на сайте корпуса[6].

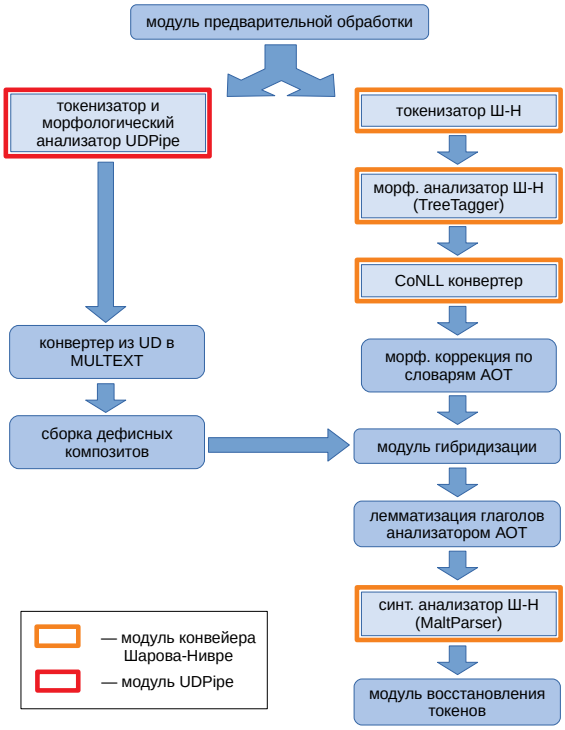

Рисунок 1. Структурная схема конвейера PaRuS_pipe

В нотационном отношении в PaRuS_pipe приняты следующие решения:

– в токенизации: дефисные композиты (*светло-серый*, *слуга-азиат*) представляются одним токеном;

[6] https://parus-proj.github.io/PaRuS/parus_pipe.html

– в морфологическом анализе:
используется нотация MULTEXT-East[7] [24];
– в синтаксическом анализе: строится дерево зависимостей в нотации СинТагРус[8] [14]; количество типов синтаксических отношений — 76.

Качество морфологического анализа в PaRuS_pipe было оценено по методологии MorphoRuEval-2017. Опираясь на свежий обзор по автоматическому морфологическому анализу русского языка [25], приведем оценки точности PaRuS_pipe в сравнении с TreeTagger (из Ш-Н), UDPipe и современным нейросетевым алгоритмом rnnmorph [26] — таблицы 1–2. Колонке 3-in-1 соответствует gold standard соревнований MorphoRuEval-2017 (он состоял из трех частей). Также для оценки использовались два более крупных набора размеченных данных (фрагменты ГИКРЯ и СинТагРус), служивших обучающими множествами в ходе соревнований. Серые пустые ячейки соответствовали бы тестированию на обучающем множестве, поэтому значения в них не приводятся.

Таблица 1. Точность определения граммем по правилам MorphoRuEval-2017

Система	Набор данных		
	3-in-1	ГИКРЯ	СинТагРус
PaRuS_pipe	90,73	90,54	90,40
rnnmorph	**96,28**		93,17
TreeTagger	89,88	89,69	88,97
UDPipe	89,01	88,90	

Таблица 2. Точность восстановления нормальной формы (независимо от успешности определения граммем)

Система	Набор данных		
	3-in-1	ГИКРЯ	СинТагРус
PaRuS_pipe	**97,89**	97,69	97,10
rnnmorph	95,28		94,79
TreeTagger	92,56	92,22	90,68
UDPipe	93,36	93,03	

[7] http://corpus.leeds.ac.uk/mocky/msd-ru.html
[8] http://ruscorpora.ru/new/instruction-syntax.html

Из таблиц видно, что в задаче лемматизации новый конвейер превзошел даже самый современный алгоритм, хотя и уступает ему в задаче определения граммем (из числа тех, что оцениваются по методологии MorphoRuEval-2017).

Качество синтаксического анализа мы не оценивали самостоятельно. Согласно оценке [11], проводившейся на корпусе СинТагРус (т.е. с идеальной морфологической разметкой), использованный в PaRuS_pipe синтаксический анализатор MaltParser может достигать показателей:

– 89% — точность установления синтаксического родителя (UAS, unlabeled attachment score),
– 82% — точность установления и родителя, и типа синтаксического отношения (LAS, labeled attachment score).

3.3. Дедупликация, фильтрация и перемешивание предложений

Проблема дублирования текстовой информации особенно остро стояла для новостных сообщений. Это связано с практикой цитирования новостной информации, а также спецификой современного стиля новостного сообщения (возврат к ранее опубликованной информации с целью напомнить читателю контекст события или обозначить связь с подобными событиями). Технически дедупликация выполнялась на уровне предложений. По тексту предложения вычислялась свертка на базе хэш-функции SHA-256. Уникальность свертки контролировалась в масштабе всего корпуса.

Фрагментарная верификация новостной части корпуса показала, что, несмотря на высокую эффективность алгоритма jusText, в очищенные тексты попало значительное количество посторонней информации. Кроме того, полезность некоторых предложений самого новостного сообщения также оказалась под вопросом. Например, краткие подписи к изображениям и фотографиям, имя и фамилия автора публикации, дата публикации и т.п. Для решения проблемы «бесполезных» предложений был создан специальный инструмент, позволявший специалисту формировать поисковые запросы для их обнаружения и выборочно удалять записи. В общей сложности это позволило отфильтровать около 750 тыс. предложений в новостной части корпуса. Также для повышения качества корпуса выполнено

удаление предложений, в которых содержалась низкочастотная лексика. С этой целью был построен словарь лемм, встречающихся в корпусе, и подсчитаны их абсолютные частоты. На основании словаря был создан список низкочастотных слов (с частотой менее 4). В основном он состоял из опечаток и последовательностей символов, не являющихся словами. Все предложения, в которых встречалось хотя бы одно слово из списка низкочастотных, удалялись из корпуса. Это привело к удалению чуть более 2% предложений.

Последним шагом в создании корпуса было перемешивание предложений в случайном порядке. Такой прием позволил избежать необходимости рассмотрения правовых вопросов, связанных с текстами-источниками.

Заключение

В статье представлен новый корпус русского языка `PaRuS`, состоящий из предложений русского литературного языка, снабженных морфологической и синтаксической разметкой. Для создания корпуса был разработан гибридный конвейер `PaRuS_pipe` на основе двух существующих конвейеров — Шарова-Нивре и `UDPipe`. Итоги тестирования конвейера `PaRuS_pipe` показывают, что качество разметки языкового материала корпуса `PaRuS` достаточно высоко. Таким образом, новый корпус может быть успешно использован для задач компьютерной лингвистики.

Список литературы

[1] С. Ю. Толдова, О. Н. Ляшевская. «Современные проблемы и тенденции компьютерной лингвистики (в зеркале 24-ой конференции по компьютерной лингвистике COLING 2012 Мумбаи)», *Вопросы языкознания*, 2014, №1, с. 120–145. ↑181

[2] Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др. *Лингвистическое обеспечение системы ЭТАП-2*, Наука, М., 1989, ISBN 5-02-006572-2, 296 с. ↑182

[3] V. Benko. “Aranea: yet another family of (comparable) web corpora”, *Text, Speech and Dialogue*, 17th International Conference TSD 2014 (Brno, Czech Republic, September 8–12, 2014), Lecture Notes in Computer Science, vol. **8655**, eds. P. Sojka, A. Horák, I. Kopeček, K. Pala, Springer

International Publishing, Switzerland, 2014, ISBN 978-3-319-10815-5, pp. 257–264. ↑182

[4] В. Бенко, В. Захаров. «Сверхбольшие корпусы русского языка: новые возможности и новые проблемы», По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июня 2016 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **15(22)**, Изд-во РГГУ, М., 2016, с. 79–93 (англ. http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf). ↑182

[5] M. Jakubicek, A. Kilgarriff, V. Kovar, P. Rychly, V. Suchomel. "The TenTen corpus family", Int. Conf. on Corpus Linguistics (Lancaster, 2013). ↑182

[6] В. И. Беликов, Н. Ю. Копылов, А. Ч. Пиперски, В. П. Селегей, С. А. Шаров. «Корпус как язык: от масштабируемости к дифференциальной полноте», По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая–2 июня 2013 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **12 (19)**, Изд-во РГГУ, М., 2013, с. 84–95. ↑182

[7] Т. Шаврина, О. Шаповалова. «To the methodology of corpus construction for machine learning: "Taiga"syntax tree corpus and parser», *Труды международной конференции «Корпусная лингвистика-2017»* (Санкт-Петербург, 27–30 июня 2017 г.), Издательство СПбГУ, СПб., 2017, с. 78–84 (англ.). ↑182

[8] T. O. Shavrina. «Дифференциальный подход к построению веб-корпусов», По материалам ежегодной международной конференции «Диалог» (Москва, 30 мая–2 июня 2018 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **17(24)**, Изд-во РГГУ, М., 2018 (англ.). ↑182

[9] Ю. Д. Апресян, И. М. Богуславский, Б. Л. Иомдин, Л. Л. Иомдин, А. В. Санников, В. З. Санников, В. Г. Сизов, Л. Л. Цинман. «Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы», *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*, Индрик, М., 2005, с. 193–214. ↑183

[10] В. А. Плунгян. «Зачем нужен Национальный корпус русского языка? Неформальное введение», *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*, Индрик, М., 2005, с. 6–20. ↑183

[11] J. Nivre, I. M. Boguslavskii, L. L. Iomdin. "Parsing the SynTagRus treebank of Russian", 22nd International Conference on Computational Linguistics, COLING 2008 (18–22 August 2008, Manchester, UK), 2008, pp. 641–648. ↑183,191

[12] M. Kudinov, A. Romanenko, I. Piontkovskaya. «Conditional random field in segmentation and noun phrase inclination on tasks for Russian», По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **13 (20)**, Изд-во РГГУ, М., с. 297–306 (англ.). URL ↑183

[13] П. В. Дяченко, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, О. Ю. Подлесская, В. Г. Сизов, Т. И. Фролова, Л. Л. Цинман. «Современное состояние глубоко аннотированного корпуса русского языка (СинТагРус)», *Национальный корпус русского языка: 10 лет проекту*, Труды Института русского языка им. В. В. Виноградова, т. **6**, М., 2015, с. 272–299. URL ↑183

[14] I. Boguslavsky. "SynTagRus — a deeply annotated corpus of Russian", *Les émotions dans le discours — Emotions in Discourse*, English and French edition, eds. P. Blumenthal, I. Novakova, D. Siepmann, P. Lang, 2014, ISBN 978-3-631-64608-3, pp. 367–380. URL ↑183,190

[15] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. "Universal Dependencies v1: A multilingual treebank collection", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (May 23–28, 2016, Portorož, Slovenia), pp. 1659–1666. URL ↑183

[16] Ф. А. Антомонов. «Универсальные зависимости: сравнение синтаксического анализа для шведского языка», По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июня 2016 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **15(22)**, Изд-во РГГУ, М., 2016 (англ.), 7 с. URL ↑183

[17] O. Lyashevskaya, K. Droganova, D. Zeman, M. Alexeeva, T. Gavrilova, N. Mustafina, E. Shakurova. *Universal dependencies for Russian: a new syntactic dependencies tagset*, Higher School of Economics Research Paper No WP BRP 44/LNG/2016, 2016. doi ↑183

[18] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin, N. Frid. "Dependency treebank for Russian: concept, tools, types of information", 18th International Conference on Computational Linguistics, COLING 2000 (July 31–August 4, 2000, Universität des Saarlandes, Saarbrücken, Germany), 2000, pp. 987–991. URL ↑183

[19] W. B. Cavnar, J. M. Trenkle. "N-gram-based text categorization", 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR-94 (April 11–13, 1994, Las Vegas, Nevada), pp. 161–175. URL ↑186

[20] J. Pomikálek. *Removing boilerplate and duplicate content from Web corpora*, PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech republic, 2011, 108 pp. URL ↑188

[21] S. Sharoff, J. Nivre. «The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge», По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **10(17)**, Изд-во РГГУ, М., 2011, с. 591–604 (англ.). ↑188

[22] M. Straka, J. Straková. *Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe*, CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Vancouver, Canada, August 2017), 2017, 12 pp. ↑188

[23] А. В. Сокирко. «Морфологические модули на сайте www.aot.ru», По материалам ежегодной Международной конференции «Диалог» (2–7 июня 2004 г.), Компьютерная лингвистика и интеллектуальные технологии, Наука, М., 2004, с. 559–564. ↑189

[24] S. Sharoff, M. Kopotev, T. Erjavec, A. Feldman, D. Divjak. "Designing and evaluating Russian tagsets", 6th International Conference on Language Resources and Evaluation, LREC 2008 (Marrakech, May, 2008), pp. 279–285. ↑190

[25] И. В. Трофимов. «Морфологический анализ русского языка: обзор прикладного характера», *Программная инженерия*, **10**:9–10 (2019), с. 391–399. ↑190

[26] Д. Г. Анастасьев, И. О. Гусев, Е. М. Инденбом. «Улучшение морфологического парсера с помощью вспомогательных задач обучения и представлений слов на символьном уровне», По материалам ежегодной международной конференции «Диалог» (Москва, 30 мая–2 июня 2018 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **17(24)**, Изд-во РГГУ, М., 2018, с. 14–27 (англ.). arXiv 1807.00818 ↑190

Поступила в редакцию 19.11.2019
Переработана 20.12.2019
Опубликована 26.12.2019

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Пример ссылки на эту публикацию:

Н. А. Власова, И. В. Трофимов, Ю. П. Сердюк, Е. А. Сулейманова, И. Н. Воздвиженский. «PaRuS — синтаксически аннотированный корпус русского языка». *Программные системы: теория и приложения*, 2019, **10**:4(43), с. 181–199. 10.25209/2079-3316-2019-10-4-181-199
http://psta.psiras.ru/read/psta2019_4_181-199.pdf

Об авторах:

Наталья Александровна Власова

младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.

0000-0002-7843-6870
e-mail: nathalie.vlassova@gmail.com

Игорь Владимирович Трофимов

старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования.

0000-0002-6903-4730
e-mail: itrofimov@gmail.com

Юрий Петрович Сердюк

старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: параллельное программирование, формальные исчисления процессов, системы типов.

0000-0003-2916-2102
e-mail: Yuri@serdyuk.botik.ru

Елена Анатольевна Сулейманова

научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, одна из разработчиков технологии построения систем извлечения информации.

0000-0002-0792-9651
e-mail: yes@helen.botik.ru

Илья Николаевич Воздвиженский

к. т. н., младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, один из разработчиков технологии построения систем извлечения информации.

0000-0001-8959-3460
e-mail: vozdvin@yandex.ru

CSCSTI 16.31.21, 28.23.13
UDC 004.89:81'322.2

Natalia A. Vlasova, Igor V. Trofimov, Yuri P. Serdyuk, Elena A. Suleymanova, Ilia N. Vozdvizhenskiy. ***PaRuS*** *— syntax annotated Russian corpus.*

Abstract. In this article we present a new annotated Russian language corpus named PaRuS (Parsed Russian Sentences). The corpus containing over 2.5 billion tokens is intended for use in computer linguistics tasks involving machine learning methods. PaRuS is a collection of annotated literary Russian sentences. Our linguistic annotation includes morphological features in MULTEXT-East format, and syntactic information in SynTagRus notation. We consider the methodology of corpus creation and describe PaRuS_pipe, a hybrid linguistic pipe developed for sentence annotation. We also discuss the quality of linguistic annotation in PaRuS and provide an assessment of the PaRuS_pipe morphological analyzer, according to the MorphoRuEval-2017 competition methodology.

Key words and phrases: computer linguistics, corpus linguistics, Russian, language corpus, markup, morphology, syntax.

2010 *Mathematics Subject Classification*: 68T50; 91F20

References

[1] S.Yu. Toldova, O. N. Lyashevskaya. "Contemporary issues and trends in computational linguistics", *Voprosy yazykoznaniya*, 2014, no.1, pp. 120–145 (in Russian). ↑181

[2] Yu. D. Apresyan, I. M. Boguslavskiy, L. L. Iomdin i dr. *Linguistic Support of the ETAP-2 System*, Nauka, M., 1989, ISBN 5-02-006572-2, 296 pp. ↑182

[3] V. Benko. "Aranea: yet another family of (comparable) web corpora", *Text, Speech and Dialogue*, 17th International Conference TSD 2014 (Brno, Czech Republic, September 8–12, 2014), Lecture Notes in Computer Science, vol. **8655**, eds. P. Sojka, A. Horák, I. Kopeček, K. Pala, Springer International Publishing, Switzerland, 2014, ISBN 978-3-319-10815-5, pp. 257–264. ↑182

[4] V. Benko, V. Zakharov. "Very large Russian corpora: new opportunities and new challenges", Po materialam yezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 1–4 iyunya 2016 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **15(22)**, Izd-vo RGGU, M., 2016, pp. 79–93 http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf. ↑182

[5] M. Jakubicek, A. Kilgarriff, V. Kovar, P. Rychly, V. Suchomel. "The TenTen corpus family", Int. Conf. on Corpus Linguistics (Lancaster, 2013). ↑182

[6] V. I. Belikov, N. Yu. Kopylov, A. Ch. Piperski, V. P. Selegey, S. A. Sharov. "Corpus as language: from scalability to register variation", Po materialam yezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 29 maya–2 iyunya 2013 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **12 (19)**, Izd-vo RGGU, M., 2013, pp. 84–95 (in Russian). ↑182

10.25209/2079-3316-2019-10-4-181-199

[7] T. Shavrina, O. Shapovalova. "To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser", *Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2017"* (Sankt-Peterburg, 27–30 iyunya 2017 g.), Izdatel'stvo SPbGU, SPb., 2017, pp. 78–84. ↑182

[8] T. O. Shavrina. "Differential approach to web-corpus construction", Po materialam yezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 30 maya–2 iyunya 2018 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **17(24)**, Izd-vo RGGU, M., 2018. ↑182

[9] Yu. D. Apresyan, I. M. Boguslavskiy, B. L. Iomdin, L. L. Iomdin, A. V. Sannikov, V. Z. Sannikov, V. G. Sizov, L. L. Tsinman. "Syntactically and Semantically Annotated Corpus of Russian: State-of-the-Art and Prospects", *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy*, Indrik, M., 2005, pp. 193–214 (in Russian). ↑183

[10] V. A. Plungyan. "What do we need Russian National Corpus for? An informal introduction", *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy*, Indrik, M., 2005, pp. 6–20 (in Russian). ↑183

[11] J. Nivre, I. M. Boguslavskii, L. L. Iomdin. "Parsing the SynTagRus treebank of Russian", 22nd International Conference on Computational Linguistics, COLING 2008 (18–22 August 2008, Manchester, UK), 2008, pp. 641–648. ↑183,191

[12] M. Kudinov, A. Romanenko, I. Piontkovskaya. "Conditional random field in segmentation and noun phrase inclination on tasks for Russian", Po materialam yezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 4–8 iyunya 2014 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **13 (20)**, Izd-vo RGGU, M., pp. 297–306. ↑183

[13] P. V. Dyachenko, L. L. Iomdin, A. V. Lazurskiy, L. G. Mityushin, O. Yu. Podlesskaya, V. G. Sizov, T. I. Frolova, L. L. Tsinman. "A deeply annotated corpus of Russian texts (SynTagRus): contemporary state of affairs", *Natsional'nyy korpus russkogo yazyka: 10 let proyektu*, Trudy Instituta russkogo yazyka im. V. V. Vinogradova, vol. **6**, M., 2015, pp. 272–299 (in Russian). ↑183

[14] I. Boguslavsky. "SynTagRus — a deeply annotated corpus of Russian", *Les émotions dans le discours — Emotions in Discourse*, English and French edition, eds. P. Blumenthal, I. Novakova, D. Siepmann, P. Lang, 2014, ISBN 978-3-631-64608-3, pp. 367–380. ↑183,190

[15] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. "Universal Dependencies v1: A multilingual treebank collection", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (May 23–28, 2016, Portorož, Slovenia), pp. 1659–1666. ↑183

[16] F. A. Antomonov. "Universal dependencies: a parsing comparison for Swedish", Po materialam yezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 1–4 iyunya 2016 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **15(22)**, Izd-vo RGGU, M., 2016, 7 pp. ↑183

[17] O. Lyashevskaya, K. Droganova, D. Zeman, M. Alexeeva, T. Gavrilova, N. Mustafina, E. Shakurova. *Universal dependencies for Russian: a new syntactic dependencies tagset*, Higher School of Economics Research Paper No WP BRP 44/LNG/2016, 2016. ↑183

[18] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin, N. Frid. "Dependency treebank for Russian: concept, tools, types of information", 18th International

Conference on Computational Linguistics, COLING 2000 (July 31–August 4, 2000, Universität des Saarlandes, Saarbrücken, Germany), 2000, pp. 987–991. ↑183

[19] W. B. Cavnar, J. M. Trenkle. "N-gram-based text categorization", 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR-94 (April 11–13, 1994, Las Vegas, Nevada), pp. 161–175. ↑186

[20] J. Pomikálek. *Removing boilerplate and duplicate content from Web corpora*, PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech republic, 2011, 108 pp. ↑188

[21] S. Sharoff, J. Nivre. "The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge", Po materialam yezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 25–29 maya 2011 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **10(17)**, Izd-vo RGGU, M., 2011, pp. 591–604. ↑188

[22] M. Straka, J. Straková. *Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe*, CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Vancouver, Canada, August 2017), 2017, 12 pp. ↑188

[23] A. V. Sokirko. "Morphological Modules on www.aot.ru Website", Po materialam yezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (2–7 iyunya 2004 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, Nauka, M., 2004, pp. 559–564 (in Russian). ↑189

[24] S. Sharoff, M. Kopotev, T. Erjavec, A. Feldman, D. Divjak. "Designing and evaluating Russian tagsets", 6th International Conference on Language Resources and Evaluation, LREC 2008 (Marrakech, May, 2008), pp. 279–285. ↑190

[25] I. V. Trofimov. "Automatic morphological analysis for Russian: application-oriented survey", *Programmnaya inzheneriya*, **10**:9–10 (2019), pp. 391–399 (in Russian). ↑190

[26] D. G. Anastas'yev, I. O. Gusev, Ye. M. Indenbom. "Improving part-of-speech tagging via multi-task learning and character-level word representations", Po materialam yezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 30 maya–2 iyunya 2018 g.), Komp'yuternaya lingvistika i intellektual'nyye tekhnologii, vol. **17(24)**, Izd-vo RGGU, M., 2018, pp. 14–27. arXiv 1807.00818 ↑190

Sample citation of this publication:

Natalia A. Vlasova, Igor V. Trofimov, Yuri P. Serdyuk, Elena A. Suleymanova, Ilia N. Vozdvizhenskiy. "PaRuS — syntax annotated Russian corpus". *Program Systems: Theory and Applications*, 2019, **10**:4(43), pp. 181–199. (*In Russian*).

10.25209/2079-3316-2019-10-4-181-199

http://psta.psiras.ru/read/psta2019_4_181-199.pdf