Вестник Московского университета. Серия 9. Филология. 2024. № 3. С. 94–112
Lomonosov Philology Journal. Series 9. Philology, 2024, no. 3, pp. 94–112

СТАТЬИ

RUCONST: СИНТАКСИЧЕСКИЙ КОРПУС РУССКОГО ЯЗЫКА С РАЗМЕТКОЙ ПО НЕПОСРЕДСТВЕННЫМ СОСТАВЛЯЮЩИМ

П.В. Гращенков

Московский государственный университет имени М.В. Ломоносова, Москва, Россия; pavel.gra@gmail.com

Аннотация: В статье представлены результаты работы по созданию корпуса русского языка с разметкой по синтаксическим составляющим. После обсуждения основных преимуществ подхода к синтаксическому анализу, построенному на грамматике составляющих, представлены примеры некоторых имеющихся на данный момент ресурсов и сформулированы основные ожидания от корпуса. Далее описан процесс создания корпуса, включавший в себя проработку дизайна представления данных, привлечение ансамбля инструментов морфосинтаксической разметки, фильтрацию ошибочных разборов и совпадающих примеров и т. д. В конце статьи изложены основные принципы работы с корпусом, описаны его базовые характеристики и приведены примеры использования.

Ключевые слова: корпуса текстов; русский язык; синтаксис; морфология

Финансирование: Исследование поддержано грантом РНФ 22-18-00037 и проводится в НИВЦ МГУ имени М.В. Ломоносова.

doi: 10.55959/MSU0130-0075-9-2024-47-03-7

Для цитирования: *Гращенков П.В.* RuConst: синтаксический корпус русского языка с разметкой по непосредственным составляющим // Вестн. Моск. ун-та. Серия 9. Филология. 2024. № 3. С. 94–112.

RUCONST: A TREEBANK FOR RUSSIAN

Pavel V. Grashchenkov

Lomonosov Moscow State University, Moscow, Russia; pavel.gra@gmail.com

Abstract: The paper presents the results of the development of a corpus for Russian with the syntactic markup. Having discussed the main advantages of the constituent grammar approach to the syntax, we present examples of resources, that are currently available on the internet, and highlight basic expectations for a constitu-

ency treebank. Then the paper describes the process of developing the treebank, which includes working out the design of data representation, creation of an ensemble of morphosyntactic markup tools, filtering erroneous parses and matching examples, etc. The paper finishes outlining the basic principles of search in the treebank, describing its main characteristics and giving some examples of use.

Keywords: corpora; Russian; syntax; morphology

Funding: The study is supported by a grant from the Russian Science Foundation (22-18-00037) and is carried out at the Research Computing Center, Lomonosov Moscow State University.

For citation: Grashchenkov P. (2024) RuConst: A Treebank for Russian Language. *Lomonosov Philology Journal. Series 9. Philology,* no. 3, pp. 94–112.

1. Преимущества составляющих перед зависимостями[1]

Синтаксический анализ традиционно проводится в рамках одного из двух формализмов: грамматики составляющих (ГС) или грамматики зависимостей (ГЗ), см. [Тестелец 2001: 61–154] и приводимую там литературу. Несмотря на то, что в прикладной и корпусной лингвистике тотальное распространение получил формализм ГЗ, составляющие имеют ряд существенных преимуществ. Назовем некоторые из них.

i) «Голые» зависимости без специального иерархизирования не могут задавать корректный порядок слов, ср., например: **съесть два эти яблока* vs. *съесть эти два яблока*[2]. С точки зрения ГЗ обе структуры должны быть приемлемы, достаточно того, что в обе-

[1] По справедливому замечанию одного из рецензентов, в большинстве случаев в данной работе речь идет о противопоставлении «синтаксическая группа — синтаксическая вершина». Мы, однако, будем говорить о составляющих, т. к. корпус подразумевает поиск и по группам, и по вершинам. Те преимущества, которые есть у корпуса по сравнению, например, с синтаксическим подкорпусом НКРЯ, относятся к возможностям учета информации именно о группах.

[2] Один из рецензентов заметил здесь: «Автор слишком хорошо думает о русской грамматике. Чтобы не препираться о расхождении интуиций о приемлемости, сверимся с НКРЯ, корпус выдает 132 примера с последовательностью *два эти*: после отсева мусора остается 119 чистых примеров вида ***два** эти X-а*. Конечно, это не столько, сколько при базовом порядке *эти два X-а* (D Num NP), но базовый и инвертированный порядки и не должны быть представлены с одинаковой частотностью, а возможность инверсии D Num NP Ю Num D__NP является интересным свойством русской грамматики».

Если прибегнуть к помощи поисковых систем, то «Яндекс» дает 2 результата против 250 на инвертированный vs. прямой порядок, а Google — 0 результатов против 7340. Как представляется, такие цифры можно считать показателем неграмматичности первого и грамматичности второго вариантов. В любом случае, описание неграмматичных порядков слов представляется более предпочтительным в терминах именно ГС, а не ГЗ.

их *эти, два* зависит от *яблока*. Современный формализм ГС обычно подразумевает указание на вершину составляющей, поэтому он легко преобразуем в ГЗ, см., например, [Gelbukh, Torres, Calvo 2005]. Обратное преобразование возможно, только если кроме указания зависимостей все словоформы проиндексированы с точки зрения порядка их следования.

ii) Отношения связывания между антецедентом и анафорическим средством, перемещенными объектами и их следами и т. п. естественно определяются в терминах составляющих, а не связанных в цепочки словоформ.

iii) Селективные ограничения могут быть представлены в терминах составляющих более экономно, чем в терминах зависимостей. Например, переходные глаголы принимают в качестве комплемента именную группу, вершиной которой может быть как существительное, так и местоимение; определением в именной группе может быть как группа прилагательного, так и группа причастия и т. д. Таким образом, в ГС мы оперируем типами составляющих, а в ГЗ — частями речи. Поскольку одной и той же составляющей могут соответствовать разные части речи (именная группа: существительные, местоимения, атрибуты при эллипсисе и субстантивации и т. д.), мы действительно можем говорить о большей экономности.

iv) В ГС есть возможность различать разные типы вложенных структур: комплементы, адъюнкты, спецификаторы. Важность различения адъюнктов и аргументов (последних — и между собой) очевидна синтаксистам, работающим в разных формальных парадигмах. Но именно в терминах ГС такое различение получает некоторое онтологическое обоснование, опирающееся, например, на тайминг деривации (а не только ярлыки для стрелочек).

v) В некоторых случаях зависимые элементы ориентируются на группу, а не на вершину, по сути имеет «место зависимость через две стрелки». В [Тестелец 2001: 149] в качестве примера приводится предложная группа *прямо над окном*. Формирование зависимости только между *прямо* и *над* было бы невозможно, если бы у *над* не было зависимой именной группы. Можно сказать, что *прямо* в данном случае надстраивается над составляющей (PP). Аналогичны примеры модификаторов типа *слишком*: *слишком низкий для меня потолок*, употребление *слишком низкий потолок* возможно, но подразумевает невыраженного экспериенцера (для кого).

vi) Элементы, имеющие сферу действия, располагаются в определенной позиции относительно всей фразы. Например, *только* находится перед всей именной группой (*только старые московские улицы* vs. **московские только старые улицы*), *не* — перед глагольной группой (*не ест картошку*) и т. д. При этом такие частицы имеют сферу действия либо на всех элементах составляющей, к вершине которой они относятся, либо или на любом ее элементе (*Кот не ест картошку, он мясом питается*).

vii) Именно составляющие (а не вершины) как феномен лежат в основе эффектов «тяжести» [Hawkins 1990], ср. английское: **John gave to Mary some money vs.* ^{Ok}*John gave to Mary everything that he has earned in his entire difficult life*. Синтаксическую тяжесть гораздо проще определять как количество входящих в составляющую словоформ, чем количество стрелок, которые могут зависеть от других стрелок, которые в конце концов выходят из вершины (*everything → earned → life → difficult → …*), см. здесь [Dryer 1992].

Есть и другие свойства, говорящие о том, что синтаксические модели на составляющих гораздо лучше описывают языковую реальность, см. [Тестелец 2001: 107–154]. По этой причине синтаксические корпуса, имеющие разметку по составляющим, приобретают особую ценность.

2. Синтаксические корпуса с разметкой на составляющих

В 1961 г. лингвистами Университета Браун в США был создан корпус Брауна, который включал примерно миллион слов американского варианта английского языка из текстов различных жанров. В 1980-х годах этот корпус получил разметку, включающую частеречные и синтаксические теги, см. [Баранов 2001: 119–120]. Один из наиболее известных синтаксических корпусов, в основе которого лежит разметка синтаксическими составляющими, — Penn Treebank. Он был создан в 1990-х гг. Пенсильванским университетом на основе материала американской деловой газеты *The Wall Street Journal* а также нескольких других корпусов, существовавших на тот момент [Там же].

Для русской корпусной лингвистики большим прогрессом явилось создание Национального корпуса русского языка, в который на определенном этапе вошел подкорпус СинТагРус, получивший разметку в виде синтаксических отношений. СинТагРус содержит чуть более полутора миллиона слов в 107 129 предложениях и является единственным значительным по объему синтаксическим кор-

пусом русского языка со снятой омонимией[3]. При создании СинТагРуса использовался понятийный аппарат проекта ЭТАП, и номенклатура используемых синтаксических отношений оказывается насколько подробной, настолько же и порой плохо предсказуемой и трудной для написания запросов. Кроме синтаксических отношений, поиск по СинТагРусу может быть произведен с использованием лексических функций[4]. Другим ограничением СинТагРуса является его онлайн-формат: он удобен для поиска отдельных конструкций, но использовать его как массив для обучения или для многофакторного поиска невозможно в силу того, что он доступен только в интернете.

Если говорить о корпусах, использующих аппарат синтаксических составляющих, такие корпуса были разработаны для целого ряда языков: английского [Marcus, Santorini, Marcinkiewicz 1993; Taylor, Marcus, Santorini 2003], французского [Abeillé, Clément, Toussenel 2003; Wang et al. 2020], немецкого [Brants et al. 2004; Telljohann, Hinrichs, Kubler 2004], итальянского [Montemagni et al. 2000], испанского [Moreno et al. 2000], португальского [Afonso, Haber, Santos 2002], норвежского [Dyvik et al. 2016], болгарского [Simov, Popova, Osenova 2002], чешского [Hajič J. et al. 2001], турецкого [Kara et al. 2020], китайского [Xue et al. 2005], японского [Horn, Alastair, Yoshimoto 2017], тайского [Chay-intr, Sarakit, Theeramunkong 2017] и др. Объем корпусов варьирует от нескольких тысяч до нескольких сотен тысяч предложений. Для русского языка корпуса с разметкой на составляющих до сих пор создано не было.

3. Требования к функционалу корпуса и базовая структура данных

Для русского языка, таким образом, представляется востребованным создание синтаксического корпуса, размеченного составляющими. Информация о зависимостях также должна быть доступна, равно как и разметка частеречными и другими морфологическими тегами.

Итого, требования к проектируемому корпусу на этапе планирования включали:

i) наличие информации о синтаксических составляющих;

ii) наличие информации о синтаксических зависимостях;

[3] 107 129 предложений, 1 529 501 слово, см. https://ruscorpora.ru/new/search-syntax.html.

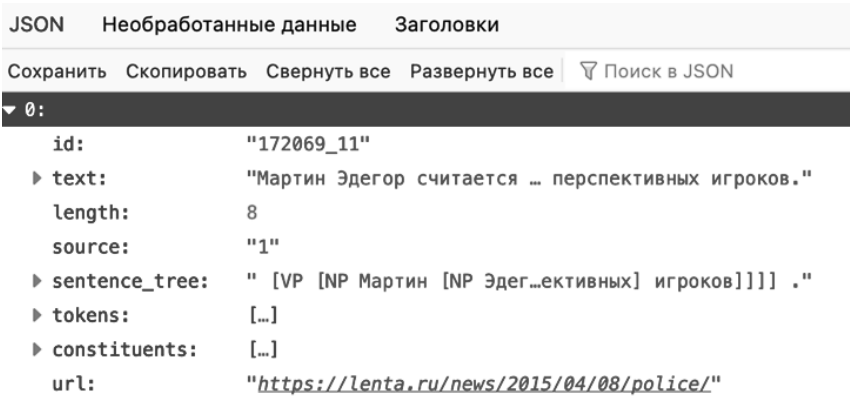
[4] https://ruscorpora.ru/page/instruction-syntax/.

iii) наличие удобной и общепонятной морфосинтаксической нотации;
iv) размер — сотни тысяч предложений (миллионы словоформ);
v) доступность датасета офлайн;
vi) возможность написания сложных поисковых запросов;
vii) возможность выгрузки и анализа данных после отрабатывания запросов.

В качестве морфосинтаксической нотации был выбран ставший общепринятым в последние годы формализм Universal Dependencies (далее — UD) [Nivre 2015][5]. Синтаксические теги в нем имеют достаточно общий и абстрактный вид[6] и соответствуют общелингвистическим представлениям, не требуя глубоко погружения в детали конкретных синтаксических теорий[7].

Для представления информации о предложении было решено использовать достаточно стандартную структуру данных, предполагающих отдельные поля для id, самого текста, информации о длине, источнике (см. ниже) и скобочном представлении разбора составляющей. Кроме этих полей, есть также два больших поля со списками словоформ (tokens) и составляющих (constituents), а также ссылка на пример в интернете:

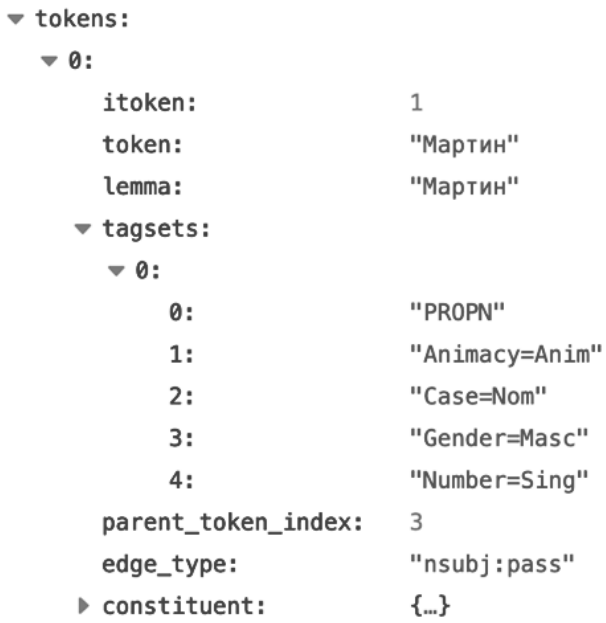
(1) Схема организации данных в корпусе



Ниже приводится пример представления для отдельной словоформы (список в блоке tokens):

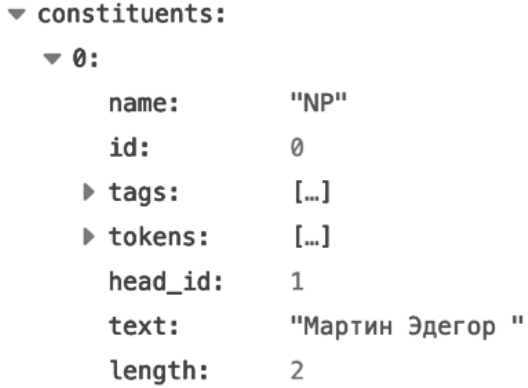
[5] См. также https://universaldependencies.org/.
[6] См. https://universaldependencies.org/u/dep/.
[7] В частности, для поиска по СинТагРусу не всегда прозрачным является то, какое именно из комплетивных, определительных, аппозитивных и др. отношений нужно выбирать, где проходят границы классов разных синтаксических отношений и т. п.

(2)



В следующем примере представлен вход для одной из составляющих (блок constituents):

(3)



Так хранится каждое конкретное предложение в корпусе. Весь корпус после разработки был разбит на 4 части (источник, поле «source»), каждая хранится в отдельном json-файле, где все предложения представлены как список верхнего уровня.

4. Процесс разработки, работа по устранению ошибок

В качестве источника данных был взят архивный массив данных с публикациями портала Lenta.ru. Поскольку этот ресурс хранит публикации прошлых лет, все примеры из корпуса доступны по ссылкам в интернете. Исходный массив был разбит на части, состоящие из 1 млн предложений каждое. Четыре такие части подверглись разбору для добавления в корпус морфологической и синтаксической информации.

Для получения морфосинтаксической разметки были задействованы автоматические системы грамматического анализа. Ошибки парсеров устранялись при помощи так называемой методики «ансамбля» инструментов: текст пропускался через все выбранные парсеры и оставлялись лишь те примеры, анализ которых совпадал.

Были задействованы три наиболее часто используемых нейросетевых инструмента, дающих оптимальное на сегодня качество за приемлемое время работы: Stanza (см. [Qi et al. 2020]), UDPipe (см. [Nguyen et al. 2021]) и Trankit (см. [Straka, Hajic, Strakova 2016]). Поскольку все разрабатываемые к данному моменту нейросетевые решения для русского языка обучаются на корпусе СинТагРус, был риск, что использование подобных инструментов для создания нового датасета может не дать качественного прироста в данных. Существовала возможность, что примеры, которые верно разбираются всеми перечисленными инструментами, будут некоторым образом похожи на те или иные данные СинТагРуса. Напротив, примеры, на которых один или несколько парсеров отрабатывают неудачно, могут являться проблемой из-за недопредставленности конструкций определенного типа в СинТагРусе. Для расширения феноменологической базы создаваемого корпуса было решено использовать также разработанный в рамках одного из коммерческих проектов контекстно-свободный парсер на правилах.

На первом этапе обработки, таким образом, исходные данные проходили через четыре указанных инструмента синтаксического анализа. На следующем этапе на основании построенного четырьмя парсерами дерева зависимостей формировалось скобочное представление в терминах ГС. Далее для устранения ошибок автоматического разбора, потенциально возможного у каждого из инструментов, отбирались лишь те варианты, которые совпадали у трех разных систем. Такие разборы считались валидными, примеры отбирались в датасет. Приведем примеры таких скобочных представлений:

(4)

а) [VP Об [NP этом] сообщает [NP The [NP Hollywood [NP Reporter]]] .]

б) [VP Об [NP этом] сообщает [NP The [NP Hollywood] [NP Reporter]]]

в) [VP Об [NP этом] сообщает [NP The [NP Hollywood] [NP Reporter]]]

Пример а) в данном случае взят из разбора парсера на правилах, б) и в) — из разборов от Stanza / UDPipe соответственно. Как видно, все они совпадают с точностью до знаков препинания[8]. Такие случаи считались тождественными независимо от того, совпадают ли ярлыки синтаксических отношений или нет. Решение игнорировать ярлыки отношений было принято потому, что нотации могут слегка различаться (близкие синтаксические отношения из UD иногда размечаются недостаточно систематично). Кроме того, иногда в разборе какого-либо парсера могла происходить ошибка в определении отношений, чаще всего — путались местами субъект и объект (здесь, например, выделено подлежащее *кубок*: [VP [NP *Первое место* [NP *и кубок* [NP *чемпионов*]]] *завоевал* [NP *МГУ*]]).

Часть 1 (первый миллион предложений) была отобрана при помощи парсера на правилах и систем Stanza и UDPipe, остальные три части — при участии Stanza, UDPipe, Trankit[9].

Следующим этапом была фильтрация схожих предложений. Был выбран порог совпадения в четыре символа, т. е. все предложения, которые различались на четыре символа или менее, фильтровались[10] — оставалось только одно из них, см. следующие примеры:[11]

(5)

==>duplicates found: О причинах смерти ничего не сообщается. О причинах аварии ничего не сообщается.

==>duplicates found: Об этом он заявил в интервью BBC. Об этом он заявил в интервью CNN.

[8] Знаки препинания и разделители игнорировались при сопоставлении.

[9] Парсер на правилах использовался только для отбора первой части, т. к. суммарное время его работы над 1 млн предложений составило около 3 месяцев. Время работы над частями в 1 млн предложений у UDPipe составлялось около 2–3 дней, у Stanza — порядка одной недели, у Trankit — несколько более недели.

[10] Из-за необходимости сопоставлять каждое из сотен тысяч предложений со всеми остальными время фильтрации каждой из четырех частей составляло 1–2 недели.

[11] Появление полных или частичных дубликатов объясняется как перепечатыванием материалов, так и обилием в данных стандартных журналистских клише.

==>duplicates found: Это на 23,3 процента больше, чем в 2015 году. Это на 3,4 процента больше, чем в 2012 году.

==>duplicates found: Врачи оценивают его состояние как стабильное. Врачи оценивают их состояние как стабильное.

==>duplicates found: В результате аварии пострадали 18 человек. В результате аварии пострадали 33 человека.

==>duplicates found: Информации о пострадавших нет. Информации о пострадавших нет.

==>duplicates found: 182 человека пострадали. Три человека пострадали.

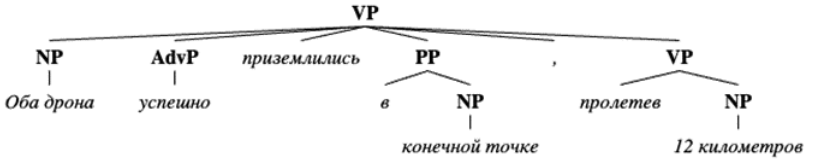
После этого прошедшие фильтрацию примеры добавлялись в части 1–4 корпуса. Часть 1, в создании которой участвовал парсер на правилах, в результате представляет собой наиболее короткие предложения (см. статистику ниже), т. к. контекстно-свободный парсер склонен «набирать» большее количество ошибок с ростом длины предложения.

Скажем несколько слов о принятой структуре составляющих. Составляющей при создании корпуса считалась любая ветвящаяся вершина. Именные вершины и вершины-наречия оборачивались в составляющие даже в отсутствие зависимых.

Будучи построена из ГЗ UD, итоговая структура в результате небинарна, не различает комплементы и адъюнкты, не содержит нулевых элементов и информации о кореферентности и т. д. В действительности почти никакие сколько-нибудь представительные по объему корпуса не содержат указанных сведений. Часть такой информации может быть добавлена в корпус в процессе дополнительной доработки, но это требует очень большой вовлеченности человеческих ресурсов.

В настоящий момент имеющаяся разметка позволяет строить деревья составляющих следующего вида:

(6)



В процессе подготовки корпуса было пройдено еще два важных этапа. Во-первых, из морфологического пакета pymorphy2 [Korobov 2015], являющегося самым популярным для русского языка не нейросетевым парсером, была добавлена в морфологические теги глаголов информация о переходности. Во-вторых, принятая в UD

схема организации зависимости между адлогами и субстантивами (стрелка «case» идет от субстантива к адлогу) была инвертирована и приведена к традиционному виду, при котором вершиной предложной группы в русском языке является предлог[12].

5. Итоговая структура: объем, грамматический состав, оставшиеся неточности

В корпусе на данный момент 628 173 предложений, в которых 6 857 935 слов, средняя длина предложения 10,9 слова. Далее приведены таблицы, отображающие статистику по частям речи, типам составляющих и связям.

(7) Размер частей корпуса в словах и предложениях[13]

Часть корпуса	Предложений	Слов	Ср. длина предложения
1	105 757	738 853	6,99
2	182 670	2 023 648	11,08
3	158 618	1 914 319	12,07
4	181 128	2 181 115	12,04
Всего	628 173	6 857 935	10,9

(8) Части речи в корпусе

POS	Количество	Процент	POS	Количество	Процент
NOUN	2304641	27,94%	CCONJ	147669	1,79%
PUNCT	1342708	16,28%	DET	142571	1,73%
VERB	975200	11,82%	SCONJ	141886	1,72%
ADP	922110	11,18%	PART	111481	1,35%
ADJ	668638	8,11%	PTCPL	44462	0,54%
PROPN	625020	7,58%	AUX	34439	0,42%
PRON	284710	3,45%	X	714	0,01%
NUM	274623	3,33%	SYM	123	0,00%
ADV	226320	2,74%	INTJ	26	0,00%

(9) Типы составляющих в корпусе

Группа	Количество	Процент	Группа	Количество	Процент
NP	3224065	57,65%	NumP	103898	1,86%

[12] Вся работа по созданию корпуса заняла около полугода.

[13] Отметим, что средняя длина предложения в СинTarРусе составляет 14,3 слова, см. https://ruscorpora.ru/new/search-syntax.html.

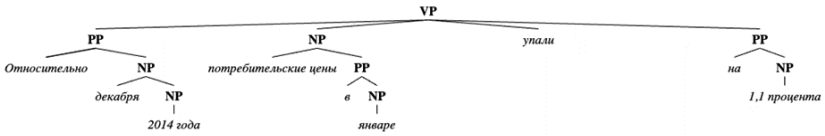
Группа	Количество	Процент	Группа	Количество	Процент
VP	1004649	17,96%	AP	99126	1,77%
PP	920839	16,46%	PartP	7307	0,13%
AdvP	226320	4,05%	CCONJP	4926	0,09%
			SCONJP	1796	0,03%

(10) Синтаксические связи в корпусе

Тип связи	Количество	Процент	Тип связи	Количество	Процент
Punct	1 341 249	16,26%	flat:foreign	73 621	0,89%
Case	935 066	11,34%	flat:name	72 736	0,88%
Nmod	874 606	10,60%	nummod:gov	53 395	0,65%
Obl	789 980	9,58%	flat	50 295	0,61%
Nsubj	717 952	8,71%	iobj	50 076	0,61%
Root	645 244	7,82%	advcl	41 012	0,50%
Amod	616 071	7,47%	cop	34 738	0,42%
Obj	288 669	3,50%	acl:relcl	32 469	0,39%
Advmod	281 353	3,41%	fixed	27 372	0,33%
nummod	223 004	2,70%	nsubj:pass	27 369	0,33%
Conj	204 136	2,48%	acl	18 157	0,22%
Cc	148 397	1,80%	csubj	16 543	0,20%
Mark	139 973	1,70%	compound	5567	0,07%
Det	138 903	1,68%	discourse	1645	0,02%
parataxis	129 295	1,57%	csubj:pass	1643	0,02%
Appos	100 852	1,22%	orphan	1265	0,02%
Xcomp	85 859	1,04%	nummod:entity	69	0,00%
Ccomp	78 744	0,95%	expl	17	0,00%
			aux	2	0,00%

В корпусе еще остаются в некотором количестве ошибки разметки. Чаще всего они связаны со следующими проблемами: а) неверное определение типа связи (субъект вместо прямого объекта); б) неверное определение типа составляющей (NumP вместо NP и т. п.); в) ошибка в уровне прикрепления составляющей. В качестве примера ошибки последнего типа можно привести случай потенциальной омонимии прикрепления адъюнкта к именной либо глагольной группе. В примере ниже вероятнее всего подразумевалась структура [$_{VP}$ [$_{PP}$ *в январе*] *упали*], однако в корпус попал другой потенциально допустимый разбор:

(11)



6. Принципы работы и предполагаемое использование

Корпус снабжен поисковыми средствами и доступен для скачивания в интернете[14]. После скачивания можно создавать поисковые запросы в формате файлов yaml, их функционал подробно описан. Поиск позволяет обращаться к морфологическим и синтаксическим тегам, ярлыкам составляющих, леммам и их спискам, задавать ограничения на порядок слов. Ниже приведен пример правила, задающего поиск субъектных и объектных именных групп, возглавляемых существительными и следующих в порядке «субъект — глагол — прямой объект»:

(12)

```
- SubExample:
    Name: SVO

    Participants:
      - Obligatory: Verb, S, O

    Items:
      - A: S
        Morph: NOUN, Case=Nom
        ConstituentType: NP

      - B: O
        Morph: NOUN, Case=Acc
        ConstituentType: NP

      - C: Verb
        Morph: VERB, VerbForm=Fin

    Links:
      - C, A: nsubj
      - C, B: obj

    Constraints:
      - Order: A, C
      - Order: C, B
```

[14] https://github.com/grapaul/Ru_Const.

Поисковая выдача формируется в виде файлов txt и csv, после чего файл csv может быть открыт и проанализирован в Excel. При задании поиска по составляющим в результат выдачи автоматически включается информация о позиции вершины в предложении и ее длине:

(13) Пример выдачи в формате txt

```
=>Найден пример Clause_WO_Nouns номер 4 типа SVO с текстом
Лидер евроскептиков назвал участников акции подонками.
  =>из предложения #174133_4
   [VP [NP Лидер [NP евроскептиков]] назвал [NP участников [NP акции]] [NP подонками]]
  =>Слоты:
    =>S = Лидер евроскептиков
    =>Verb = назвал
    =>O = участников акции
```

(14) Пример выдачи в формате csv

ex_id	id	example_name	subexamp	text	tree	sent_ler	DO		V		Subj		
1	174115_1	Clause_WO_Noun	SVO	Шамла продолжит работу в центральном офисе Cl	[VP [NP Шамла] продолж	7	работу в цен	3	5	Шамла	1	1	продолжит	2	(
2	174130_11	Clause_WO_Noun	SVO	В начале марта соответствующий законопроект, пр	[VP [PP В [NP начале [NP	11	правительст	12	2	соответствующий з	5	6	одобрило	11	(
3	174131_9	Clause_WO_Pron	SOV_o=N	Я детей не люблю.	[VP [NP Я] [NP детей] не л	4	детей	2	1	Я	1	1	люблю	4	(
4	174133_4	Clause_WO_Noun	SVO	Лидер евроскептиков назвал участников акции под	[VP [NP Лидер [NP евроск	6	участников а	4	2	Лидер евроскептик	1	2	назвал	3	(
5	174135_7	Clause_WO_Noun	SOV	Однако тогда пресс-служба Нацгвардии эту инфор	[VP [AdvP Однако] [AdvP	7	эту информа	6	2	пресс-служба Нацг	3	2	опровергла	7	(
6	174180_6	Clause_WO_Noun	SVO	Прибывшие медики доставили писателя в одну из	[VP [NP Прибывшие меди	9	писателя	4	1	Прибывшие медики	2	2	доставили	3	(
7	174182_10	Clause_WO_Noun	SVO	Затем мужчина вновь направил баллончик на перв	[VP [AdvP Затем] [NP муж	8	баллончик	5	1	мужчина	2	1	направил	4	(
8	174184_9	Clause_WO_Noun	SVO	В середине января парламент Украины утвердил с	[VP [PP В [NP середине [N	9	соответствуе	8	3	парламент Украины	4	2	утвердил	6	(
9	174211_11	Clause_WO_Noun	OSV	Случаи 2010 года диппредставитель назвал единич	[VP [NP Случаи [NP 2010 г	6	Случаи 2010	1	3	диппредставитель	4	1	назвал	5	(
10	174214_9	Clause_WO_Pron	SVO_o=N	Вы определяете ставки.	[VP [NP Вы] определяете	3	ставки	3	1	Вы	1	1	определяете	2	(

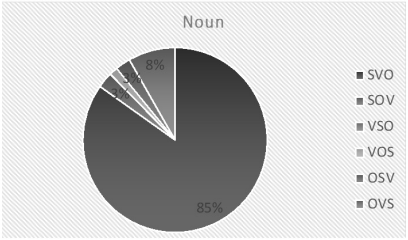
Если после получения поисковой выдачи появилась потребность иметь более полную информацию о каком-либо предложении, оно может быть извлечено из корпуса специальным скриптом.

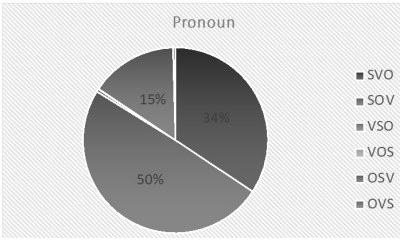
Корпус будет особенно полезен при исследовании порядка слов, изучении проблем, связанных с удобством парсинга (поведение тяжелых составляющих, свойства проекций разных частей речи и т. п.), вариативного согласования, варьирования в падежных стратегиях, при извлечении информации о синтаксическом управлении и синтаксической сочетаемости в целом и т. д.

Приведем результат анализа поисковой выдачи для глаголов с субъектом и прямым объектом. Как можно видеть, стратегии организации таких предложений разительно отличаются в случаях, если вершинами являются существительные или местоимения. При вершинах-существительных 85% примеров приходится на порядок SVO, 8% — на OVS, по 3% — на SOV и OSV. В то же время, если вершинами субъекта и прямого объекта являются местоимения, половина случаев приходится на порядок SOV, что вместе с OSV (15%) дает две трети всех употреблений. Получается, что в 65% случаев, когда аргументами являются местоимения, наблюдается препозиция обоих аргументов — это прекрасно объясняется тем, что местоимения задают известную информацию и, как правило, являются темой.

На «канонический» порядок SVO приходится всего около трети примеров (34%).

(15) Порядок слов в предикациях с аргументами, возглавляемыми только существительными и только местоимениями





7. Заключение

Мы представили синтаксический корпус русского языка, размеченный в терминах грамматики составляющих и грамматики зависимостей. Корпус открыт для скачивания[15] и может быть использован для лингвистических исследований и машинного обучения. Для лингвистов-исследователей корпус может быть полезен при изучении порядка слов, синтаксической и морфологической вариативности, проблем согласования, падежного маркирования и т. д. В дальнейшем планируются работы по устранению оставшихся ошибок в разметке данных, усовершенствование и ускорение поисковых алгоритмов, другие технические работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Баранов А.Н.* Введение в прикладную лингвистику: Учебное пособие. М., 2001. 360 с.
2. *Тестелец Я.Г.* Введение в общий синтаксис. М., 2001. 798 с.
3. *Abeillé A., Clément L., Toussenel F.* Building a Treebank for French // Building and Using Parsed Corpora. Text, Speech and Language Technology, Abeillé, A (ed.). Dodrecht, Boston, London, 2003, pp. 165–187.
4. *Afonso S., Bick E., Haber R., Santos D.* Floresta sintá(c)tica: a treebank for Portuguese // Proceedings of LREC 2002, pp. 1698–1703.
5. *Brants S., Dipper S., Eisenberg P., Hansen S., Konig E., Lezius W., Rohrer C., Smith G., Uszkoreit H.* TIGER: Linguistic interpretation of a German corpus // Research on Language and Computation, Special Issue, 2004, 2 (4), pp. 597–620.
6. *Chay-intr T., Sarakit P., Theeramunkong T.* Iterative Thai Treebank Construction via Interactive Tree Visualization // Proceedings of the 12th International Conference on Knowledge, Information and Creativity Support System (KICSS2017), 2017, pp. 97–101.

[15] https://github.com/grapaul/Ru_Const.

7. *Dryer M.S.* The Greenbergian word order correlations // Language. 1992, 68, pp. 81–138.
8. *Dyvik H., Meurer P., Rosén V., De Smedt K., Haugereid P., Smørdal Losnegaard G., Lyse Gunn I., Thunes M.* NorGramBank: A 'Deep' Treebank for Norwegian // N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.). Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3555–3562, Portorož, Slovenia. ELRA.
9. *Gelbukh A., Torres S., Calvo H.* Transforming a constituency treebank into a dependency treebank // Procesamiento del lenguaje natural. 2005, no. 35, pp. 145–152.
10. *Hajic J., Hladka B., Pajas P.* The Prague dependency treebank: Annotation structure and support // Proceedings of the IRCS Workshop on Linguistic Databases, 2001, pp. 105–114.
11. *Hawkins J.A.* A Parsing Theory of Word Order Universals // Linguistic Inquiry, 1990. Vol. 21, no. 2, pp. 223–261.
12. *Horn S.W., Alastair B., Yoshimoto K.* "Keyaki Treebank segmentation and part-of-speech labelling", 『言語処理学会 第23 回年次大会 発表論文集』, 2017, pp. 414–417.
13. *Kara, N., Marşan B., Özçelik M., Arıcan B.N., Kuzgun A., Cesur N., Aslan D.B., Yıldız, O.T.* Creating a syntactically felicitous constituency treebank for Turkish // 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020b, pp. 1–6.
14. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, 2015, pp. 320–332.
15. *Marcus M.P., Santorini B., Marcinkiewicz M.A.* Building a large annotated corpus of English: The Penn Treebank // Computational Linguistics, 1993, vol. 19, no. 2, pp. 313–330.
16. *Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R.* The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation // Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics, 2000, pp. 18–27.
17. *Moreno A., Grishman R., Lopez S., Sanchez F., Sekine S.* A Treebank of Spanish and its Application to Parsing. In Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000), Athens, Greece, 2000, pp. 107–111.
18. N*guyen Minh Van, Lai Viet, Ben Veyseh Amir Pouran, Nguyen Thien Huu*. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 80–90.
19. *Nivre J.* Towards a Universal Grammar for Natural Language Processing // Computational Linguistics and Intelligent Text Processing, 2015, pp. 3–16.
20. *Qi Peng, Zhang Yuhao, Zhang Yuhui, Bolton Jason, Manning Christopher D.* Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 101–108.
21. *Simov K., Popova G., Osenova P.* HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In: "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", edited by A. Wilson, P. Rayson, T. McEnery; Lincom-Europa, Munich, 2002, pp. 135–142.
22. *Straka M., Hajic J., Strakova J.* UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing //

Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portoroz, Slovenia, pp. 4290–4297.
23. *Taylor A., Marcus M. P., Santorini B.* The Penn Treebank: An Overview // Text Speech and Language Technology, 2003 pp. 5–22.
24. *Telljohann H., Hinrichs E., Kubler S.* The TüBa-D/Z treebank: Annotating German with a context-free backbone // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 2004, pp. 2229–2235.
25. *Wang I., Pelletier A., Antoine J.-Y., Halftermeyer A.* ODIL Syntax, a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees // Proceedings of LREC'2020, Marseille, France, URL: https://hal.science/hal-02523141.
26. *Xue N., Xia F., Chiou F.-D., Palmer M.* The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 2005, 11 (2), pp. 207–238.

REFERENCES

1. Baranov A.N. Vvedenie v prikladnuyu lingvistiku [Introduction to the Applied Linguistics]. Moscow, Editorial URSS Publ., 2001. 360 p.
2. Testelets Ya. G. Vvedenie v obschij sintaksis [Introduction to Syntax]. Moscow, RGGU, 2001. 798 p.
3. Abeillé A., Clément L., Toussenel F. Building a Treebank for French // Building and Using Parsed Corpora. Text, Speech and Language Technology, Abeillé, A (ed.). Dodrecht, Boston, London, 2003, pp. 165–187.
4. Afonso S., Bick E., Haber R., Santos D. Floresta sintá(c)tica: a treebank for Portuguese // Proceedings of LREC 2002, pp. 1698–1703.
5. Brants S., Dipper S., Eisenberg P., Hansen S., Konig E., Lezius W., Rohrer C., Smith G., Uszkoreit H. TIGER: Linguistic interpretation of a German corpus // Research on Language and Computation, Special Issue, 2004, 2 (4), pp. 597–620.
6. Chay-intr T., Sarakit P., Theeramunkong T. Iterative Thai Treebank Construction via Interactive Tree Visualization // Proceedings of the 12th International Conference on Knowledge, Information and Creativity Support System (KICSS2017), 2017, pp. 97–101.
7. Dryer M.S. The Greenbergian word order correlations // Language. 1992, 68, pp. 81–138.
8. Dyvik H., Meurer P., Rosén V., De Smedt K., Haugereid P., Smørdal Losnegaard G., Lyse Gunn I., Thunes M. NorGramBank: A 'Deep' Treebank for Norwegian // N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.). Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3555–3562, Portorož, Slovenia. ELRA.
9. Gelbukh A., Torres S., Calvo H. Transforming a constituency treebank into a dependency treebank // Procesamiento del lenguaje natural. 2005, no. 35, pp. 145–152.
10. Hajic J., Hladka B., Pajas P. The Prague dependency treebank: Annotation structure and support // Proceedings of the IRCS Workshop on Linguistic Databases, 2001, pp. 105–114.
11. Hawkins J.A. A Parsing Theory of Word Order Universals // Linguistic Inquiry, 1990. Vol. 21, no. 2, pp. 223–261.

12. Horn S.W., Alastair B., Yoshimoto K. "Keyaki Treebank segmentation and part-of-speech labelling", 『言語処理学会 第23 回年次大会 発表論文集』, 2017, pp. 414–417.
13. Kara, N., Marşan B., Özçelik M., Arıcan B.N., Kuzgun A., Cesur N., Aslan D.B., Yıldız, O.T. Creating a syntactically felicitous constituency treebank for Turkish // 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020b, pp. 1–6.
14. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, 2015, pp. 320–332.
15. Marcus M.P., Santorini B., Marcinkiewicz M.A. Building a large annotated corpus of English: The Penn Treebank // Computational Linguistics, 1993, vol. 19, no. 2, pp. 313–330.
16. Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation // Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics, 2000, pp. 18–27.
17. Moreno A., Grishman R., Lopez S., Sanchez F., Sekine S. A Treebank of Spanish and its Application to Parsing. In Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000), Athens, Greece, 2000, pp. 107–111.
18. Nguyen Minh Van, Lai Viet, Ben Veyseh Amir Pouran, Nguyen Thien Huu. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 80–90.
19. Nivre J. Towards a Universal Grammar for Natural Language Processing // Computational Linguistics and Intelligent Text Processing, 2015, pp. 3–16.
20. Qi Peng, Zhang Yuhao, Zhang Yuhui, Bolton Jason, Manning Christopher D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 101–108.
21. Simov K., Popova G., Osenova P. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In: "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", edited by A. Wilson, P. Rayson, T. McEnery; Lincom-Europa, Munich, 2002, pp. 135–142.
22. Straka M., Hajic J., Strakova J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portoroz, Slovenia, pp. 4290–4297.
23. Taylor A., Marcus M. P., Santorini B. The Penn Treebank: An Overview // Text Speech and Language Technology, 2003 pp. 5–22.
24. Telljohann H., Hinrichs E., Kubler S. The TüBa-D/Z treebank: Annotating German with a context-free backbone // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 2004, pp. 2229–2235.
25. Wang I., Pelletier A., Antoine J.-Y., Halftermeyer A. ODIL Syntax, a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees // Proceedings of LREC'2020, Marseille, France, URL: https://hal.science/hal-02523141.

26. Xue N., Xia F., Chiou F.-D., Palmer M. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 2005, 11 (2), pp. 207–238.

Поступила в редакцию 04.05.2023
Принята к публикации 16.04.2024
Отредактирована 21.04.2024

Received 04.05.2023
Accepted 16.04.2024
Revised 21.04.2024

ОБ АВТОРЕ

Гращенков Павел Валерьевич — доктор филологических наук, доцент кафедры теоретической и прикладной лингвистики, заведующий лабораторией автоматизированных лексикографических систем НИВЦ МГУ имени М.В. Ломоносова; pavel.gra@gmail.com

ABOUT THE AUTHOR

Pavel V. Grashchenkov — Prof. Dr., Department of Theoretical and Applied Linguistics, Faculty of Philology; Head of the Laboratory for Computational Lexicography, Lomonosov Moscow State University; pavel.gra@gmail.com