Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2020»

Москва, 17–20 июня 2020 г.

РУССКИЙ ЯЗЫК И КОРПУСНОЕ РАЗНООБРАЗИЕ[1]

Пиперски А. Ч. (apiperski@gmail.com)
РГГУ / НИУ ВШЭ, Москва, Россия

В статье даётся обзор применений наиболее известных корпусных ресурсов исследования для русского языка. На примере лингвистических публикаций 2019 года демонстрируется, что русистика недостаточно активно использует возможности, которые открываются перед исследователями благодаря наличию широкого разнообразия корпусов. В качестве примеров демонстрируется, какую пользу различные «неклассические» корпуса могут принести в исследованиях, посвящённых анализу явлений на различных уровнях языка: в морфологии и синтаксисе, в словообразовании и лексике, в частности в исследовании субстандартных языковых явлений, а также в сфере конструкций. Обсуждаются достоинства и недостатки отдельных корпусов с точки зрения интерфейса и удобства для использования в различных аспектах.

Ключевые слова: русский язык, корпусная лингвистика, интернет-корпуса, методология корпусных исследований

DOI: 10.28995/2075-7182-2020-19-615-627

RUSSIAN LANGUAGE AND CORPUS DIVERSITY

Piperski A. Ch. (apiperski@gmail.com)
Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-10081). Автор сердечно благодарит анонимных рецензентов за высказанные ими замечания.

This paper discusses the use of most widely-known Russian corpora, namely Russian National Corpus, ruTenTen, General Internet Corpus of Russian, and Araneum Russicum Maximum, for the theoretical study of Russian language. Based on a sample of papers from 2019, I demonstrate that scholars, especially theoretical linguists, tend to ignore the opportunities provided by a wide range of Web corpora, even though these resources are well-known to the NLP community. I present a selection of case studies to show that data from "non-classical" corpora can be used for studying various linguistic phenomena, such as: 1) variation in morphology and syntax; 2) word formation and lexical change; 3) construction grammar. I also claim that the underuse of non-classical corpora is partly due to the fact that they are (perceived as) not quite user-friendly.

Key words: Russian, corpus linguistics, Web as corpus, methodology of corpus-based studies

1. Введение

Современный русский язык обеспечен большим количеством корпусных ресурсов. Разумеется, по их количеству он не может сравниться с мировым лидером — английским языком, но всё же их диапазон чрезвычайно широк. Это приводит к тому, что в разных областях науки используются принципиально различные корпуса, которые порой оказываются незнакомы специалистам из других областей. Так, исследователи, которые занимаются грамматикой русского языка, могут и не знать о существовании корпуса Taiga объёмом 6 млрд словоформ [Shavrina, 2018] или корпуса Omnia Russica объёмом 33 млрд словоформ [Shavrina & Benko, 2019]. В свою очередь, разработчики языковых моделей, которые опираются на большие объёмы текстовых данных, наверняка используют Taiga и Omnia Russica, но едва ли заинтересуются 100-тысячным Хельсинкским аннотированным корпусом (ХАНКО) с его глубокой разметкой [Мустайоки & Копотев, 2003], тем более что этот корпус недоступен для скачивания, что особенно важно для разработки моделей.

Не существует полного перечня русских корпусов, предназначенных для всех разнообразных целей в этом континууме от традиционной русистики до NLP, хотя обзоры корпусных ресурсов делаются достаточно регулярно. Некоторое приближение к такому перечню есть на сайте NLPub[2], однако этот список явно неполон; из первых работ такого рода можно упомянуть [Резникова & Копотев, 2005], а из более новых — [Zakharov, 2013]; [Хохлова, 2016] и список в учебнике [Копотев, 2014].

В статье будут рассмотрены лингвистические корпуса, которые находят применение в современной теоретической русистике; особое внимание уделяется тому, какие плоды может принести более широкое использование некоторых из них. Разумеется, автор отдаёт себе отчёт, что многим читателям этот обзор покажется бесполезным, поскольку они и так знают, как устроены описанные здесь ресурсы, и пользуются ими — однако кажется, что такие читатели всё же не составляют большинства среди русистов.

[2] https://nlpub.ru/Ресурсы#.D0.9A.D0.BE.D1.80.D0.BF.D1.83.D1.81_.D1.82.D0.B5.D0.BA.D1.81.D1.82.D0.BE.D0.B2

2. Использование корпусов в русистике: опыт наукометрического анализа

В англоязычном мире польза корпусов для лингвистов стала очевидной уже в начале 1990-х годов — в частности, благодаря работам [Fillmore, 1992]; [Sinclair, 1991]. В русистике обращение к корпусам стало массовым лишь полтора десятилетия спустя. Так, в лекции В. А. Плунгяна «Почему современная лингвистика должна быть лингвистикой корпусов» [Плунгян, 2009] говорится:

> «Все знают, что есть две вещи, нужные, чтобы овладеть языком, это словарь и грамматика, они нужны и теоретику, и практику. Так вот, если в результате моей лекции вы ничего не поймете и не запомните, это совершенно не страшно. Запомните единственное: теперь для овладения языком человеку нужны не две, а три вещи: словарь, грамматика и корпус текстов данного языка».

В 2009 году эта мысль звучала в популярной лекции, но ещё за пять лет до того она была новой для научного сообщества: [Перцов 2006] сообщает, что В. А. Плунгян говорил это в 2004 году на презентации Национального корпуса русского языка[3] (НКРЯ). Впрочем, довольно скоро лингвистам стало ясно, что теоретику нужен корпус не в единственном, а во множественном числе: НКРЯ со всем разнообразием представленных в нём подкорпусов всё же не может служить единственным авторитетным ресурсом [Беликов и др., 2013]. Во втором десятилетии XX века появилось сразу несколько новых значимых ресурсов для русского языка, созданных в парадигме Web as Corpus [Schäfer & Bildhauer, 2013]: ruTenTen[4], Генеральный интернет-корпус русского языка [ГИКРЯ; Беликов и др., 2013], семейство корпусов Araneum Russicum [Benko, 2014]. Представляется важным оценить, какую роль эти новые, «неклассические» ресурсы играют в современной русистике на фоне единственного широко используемого «классического» корпуса — НКРЯ. Стоит добавить, что ещё одним ресурсом, который обрёл новую жизнь несколько лет назад, стал корпус СинТагРус: он был конвертирован в формат Universal Dependencies [Droganova & Zeman, 2016] и благодаря этому стал широко использоваться в типологических исследованиях.

В качестве инструмента для корпусометрии — или, точнее, для наукометрии, — можно использовать Google Scholar и посмотреть, насколько часто исследователи опираются в своих работах на данные, полученные с помощью тех или иных русских корпусов. Для этого в Google Scholar задавались запросы с названиями и/или характерными фрагментами URL-адресов различных корпусов; поиск вёлся только по публикациям 2019 года. Затем был произведён ручной подсчёт количества найденных публикаций. Полученные результаты представлены в **таблице 1**:

---

[3] http://www.ruscorpora.ru

[4] http://sketchengine.eu

Таблица 1. Число публикаций в Google Scholar за 2019 год с упоминанием различных корпусов русского языка

№	Корпус	Запрос(ы)	Число публикаций
1	НКРЯ	ruscorpora	872
2	ГИКРЯ	ГИКРЯ, GICR corpus	30
3	ruTenTen	ruTenTen, ruTenTen11	25
4	Araneum	“Araneum Russicum”	7

Как можно видеть, доля НКРЯ огромна: именно на этот корпус приходится 872 / 934 ≈ 93% упоминаний. Это и определяет структуру статьи: нисколько не умаляя достоинств НКРЯ, я бы хотел поговорить не о нём, а о других, «неклассических» корпусах — о том, какую пользу они могут принести лингвистам в их исследованиях, и указать на то, почему эти ресурсы всё-таки находят недостаточно широкое применение.

3. Примеры (возможных) исследований на «неклассических» корпусах

3.1. Морфология и синтаксис

«Неклассические» корпуса, такие как ГИКРЯ, Araneum и ruTenTen, предоставляют широкий диапазон возможностей для изучения некодифицированных грамматических явлений, в частности — вариативности. Из недавних исследований именно в таком русле выполнена работа [Nesset, 2019], автор которой строит статистические модели, описывающие устройство русских количественных словосочетаний типа *две серьёзные аварии / две серьёзных аварии*. Применение больших корпусов, отражающих устройство современной живой речи, а в случае ГИКРЯ и позволяющих проследить её социальную вариативность, показывает, что во многих случаях мы имеем дело не просто с вариантами, а с вариантами, частотность которых зависит от взаимодействия множества факторов. Образцом такого исследования может служить доклад [Беликов, 2019], в котором автор демонстрирует, что употребление вокализованных и невокализованных вариантов предлогов, в первую очередь *в(о)*, обусловлено региональными и возрастными характеристиками говорящих: к примеру, в Псковской области доля варианта *во Пскове* составляет 46,8%, а по мере удаления от Пскова она снижается: в Санкт-Петербурге — 32,5%, в Москве — 26,0%, а в Сибири — всего 19,3%; у молодёжи доля невокализованных форм выше, чем у более старшего поколения.

Рассмотрим в качестве примера употребление предлога *с / со* перед словами, начинающимися на *щ*. Категориальный подход к этому явлению позволяет принять два возможных решения: либо назвать один из двух вариантов

правильным в этом контексте, а другой — неправильным, либо признать их равно допустимыми. Так, первый из этих подходов цитируется со ссылкой на «Орфоэпический словарь русского языка» [Аванесов, 1988], а второй принимается самим автором в работе [Иткин, 2007]:

> «Согласно [ОСРЯ: 683], употребление варианта *со* является нормой также перед словами, начинающимися на *щ*. По нашим наблюдениям, это не вполне верно. В спонтанной речи соответствующие конструкции избегаются или вызывают у говорящего неуверенность. Опрос носителей языка выявляет некоторое предпочтение морфа *с* на фоне значительных идиолектных расхождений; нередко информанты признают не вполне удачными обе возможных альтернативы. На наш взгляд, варианты *с* и *со* перед *щ* — *с/со щукой*, *с/со щавелем*, *с/со щек*, *с/со щедрыми дарами* и т. д. — должны трактоваться как равноправные» [Иткин, 2007, с. 81–82]

Однако просмотр частотных списков, извлечённых из большого корпуса — в данном случае использовался Araneum Russicum Maximum, — показывает, что мы имеем дело с весьма нетривиальным распределением. В **таблице 2** приведены 82 словоформы, которые суммарно встречаются с предлогами *с / со* более 100 раз в корпусе; они отсортированы по убыванию доли употреблений с вариантом *с* (буквы *е* и *ё* считаются различными; регистр символов не учитывается).

Таблица 2. Частотность вариантов «*с / со* Х», где Х начинается на букву *щ*, по корпусу Araneum Russicum Maximum

Словоформа	*с / со* Х	Доля «*с* Х» (%)	Словоформа	*с / со* Х	Доля «*с* Х» (%)
щедрым	959	81	*щелочными*	1130	45
Щербаковым	136	76	*щебенкой*	251	41
щебеночным	258	71	*щепками*	100	39
щедростью	499	70	*щупом*	374	39
щедрыми	504	70	*щитками*	199	39
щелевым	265	68	*щёткой*	346	38
щедрой	539	68	*щеткой*	2060	38
щечной	153	67	*щенятами*	170	38
щемящим	200	64	*щелчком*	1168	38
Щелковским	107	64	*щебнем*	1703	38
Щелковского	306	63	*щётками*	171	37
щелкунчиком	115	61	*щитовкой*	292	37
щадящей	264	61	*щетками*	973	37
щадящим	492	60	*щелями*	1196	36
щитовой	110	58	*щёлочью*	150	35
щупальцами	1071	58	*щитка*	203	34
Щербинкой	121	57	*щитовками*	110	34
щитовидкой	2186	57	*щелоком*	110	33
щадящими	232	57	*щуками*	117	32
щемящей	262	56	*щенячьего*	1066	32

Словоформа	*с* / *со* X	Доля «*с* X» (%)		Словоформа	*с* / *со* X	Доля «*с* X» (%)
щучьей	101	55		*щукой*	967	32
щавелевой	156	55		*щелочами*	1022	30
щипцами	422	55		*щелочью*	862	30
щечками	136	54		*щупами*	146	30
щетинками	423	54		*щитком*	585	30
щелевыми	214	54		*щавелем*	1777	29
щеточкой	432	54		*щеками*	317	29
щенячьим	178	54		*щенками*	2100	29
щитовым	109	53		*щёк*	166	28
щелочным	559	53		*щенком*	4148	27
щелевой	433	51		*щепой*	189	26
щеточным	148	50		*щеки*	974	26
щадящих	100	50		*щетки*	175	25
щепоткой	2846	50		*щек*	788	24
щитовидной	4541	50		*щитами*	1753	23
щетиной	1602	48		*щитом*	3626	19
щелчками	174	47		*щекой*	281	19
щелью	592	47		*щенка*	206	17
щелевидными	144	47		*щитов*	269	13
щелочной	919	46		*щита*	541	13
щенячества	104	45		*щами*	450	7

Можно отметить по крайней мере три фактора, которые явно влияют на доли *с* и *со*: частеречная принадлежность, падеж и длина словоформы:

Если построить линейную регрессионную модель, предсказывающую долю «*с X*» в зависимости от свойств словоформы, мы получим следующую зависимость:

$$f(\text{«}с\ \text{X»}) = 13{,}8 + 16{,}7 \cdot Adjective + 10{,}3 \cdot Instrumental + 5{,}3 \cdot Syllables$$

Все эти три переменные: *Adjective* (прилагательное: 1, существительное: 0), *Instrumental* (творительный падеж: 1, родительный падеж: 0) и *Syllables* (количество слогов в словоформе) — вносят значимый вклад в определение доли «*с* X» ($p = 9 \cdot 10^{-6}$, $p = 0{,}0128$ и $p = 0{,}0096$ соответственно). Другие протестированные переменные: место ударения при счёте слогов от предлога и частотность словоформы — значимого влияния не оказывают.

Этот пример демонстрирует, что большой неклассический корпус позволяет практически за любой вариативностью увидеть интересные статистические закономерности. На материале НКРЯ такой анализ сделать было бы

невозможно, поскольку там насчитывается в общей сложности лишь 2030 примеров «*с / со* Х», где Х начинается на букву *щ*.

3.2. Словообразование и лексика

Использование неклассических корпусов даёт широкие возможности для анализа таких явлений в сфере словообразования и лексики, которые не попадают в корпуса типа НКРЯ. Речь идёт в первую очередь про неологизмы, недавние заимствования и жаргонизмы.

В качестве примера работы, которая для исследования таких явлений опирается на различные типы корпусов, можно привести статью [Пиперски, 2019], посвящённую словам типа *толерасты, либерасты*, *флудерасты* и т. п. В ней изложение начинается с примеров из НКРЯ, но там удаётся найти лишь 10 лексем, иллюстрирующих эту словообразовательную модель, поэтому дальнейшее исследование основано на корпусе Araneum Russicum Maximum, где обнаруживается 47 таких единиц, встретившихся хотя бы 5 раз. Собранный материал позволяет показать, что такие образования обычно возникают путём наложения *-раст-* на основы, содержащие *р* после второго гласного (точнее: в конце группы согласных после второго гласного, причём длина этой группы может быть равна 1), однако есть и примеры, где *-раст-* уже не требует такого наложения: ср. *единорасты* ‘члены партии «Единая Россия»’, где *р* в производящей основе присутствует, но не в нужной позиции, и *сталирасты* ‘сторонники Сталина’, где *р* в производящей основе нет вовсе. Из этого делается вывод, что *-раст-* постепенно морфологизуется как продуктивный пейоративный суффикс, а не только как элемент языковой игры.

Однако вывод о такой морфологизации на материале Araneum Russicum Maximum может быть лишь гипотетическим; чтобы его подтвердить или опровергнуть, следует привлечь материал ГИКРЯ, где тексты снабжены хронологической разметкой. Морфологизация суффикса будет подтверждена, если продемонстрировать, что слова с наложением *-раст-* на *р* появляются раньше, чем слова, в которых *-раст-* выделяется без такого наложения. И действительно, поиск **`[word=".*расты"]`** в подкорпусе «Живого журнала» в ГИКРЯ демонстрирует, что в более ранние годы слова на *-раст-* образуются от слов с *р* после второго гласного основы, и лишь несколько лет спустя появляются другие типы. Из частотных слов, которые встретились не менее 10 раз, с 2002 года фиксируются *либерасты* и *гитарасты*, с 2003 года — *байдарасты* и *сладострасты*, с 2004 года — *федерасты* ‘сторонники федерализации’, *флудерасты* и *пиарасты*, с 2005 года — *толерасты*, *модерасты* и *питерасты* ‘жители Санкт-Петербурга’, с 2006 года — *поттерасты* и т. д.; все эти слова соответствуют описанной модели. В то же время только в 2005 году в ГИКРЯ появляется слово *восьмидерасты*, в 2006 году — *едрасты* и в 2007 году — *едирасты* и *единорасты*. Разумеется, оценки веб-корпусов могут и должны проверяться другими источниками: например, первые примеры слова *восьмидерасты* в Google Books встречаются уже в 1993 году, но, тем не менее, грамматикализация *-раст-* в суффикс в целом подтверждается данными ГИКРЯ.

Этот пример показывает, что неклассические корпуса являются ценнейшим источником сведений о лексике, в том числе о неологизмах. Это возможно как благодаря объёму этих корпусов, так и благодаря тому, что они нередко включают субстандартные разновидности языка.

3.3. Грамматика конструкций

Неклассические корпуса дают широкий простор для изучения конструкций, в том числе новых и разговорных, которые подчас остаются вне поля зрения исследователей, опирающихся только на тщательно подготовленные к публикации и вычитанные тексты.

Для примера рассмотрим два недавних доклада на очень близкие между собой темы: [Endresen, 2019] и [Урысон, 2020]. В обоих докладах анализируются конструкции типа с повтором «*X* и *X*». Авторы приходят к схожим выводам о возможных значениях этих конструкций, однако оперируют при этом достаточно небольшим числом примеров, либо порождённых ими самими, либо извлечённых из НКРЯ по запросам типа **`работа и работа`** с конкретными словоформами. В то же время использование неклассических корпусов позволило бы заметно убыстрить извлечение примеров и получить намного более представительную выборку: так, при поиске конструкции «*ну(,) X и X*» в корпусе Araneum Russicum Maximum находится 5045 примеров, которые можно подвергнуть статистическому анализу (например, установить частотность различных типов конструкций), а наиболее интересные из них проанализировать с подробной интерпретацией.

4. Почему неклассические корпуса не становятся классическими?

Примеры, приведённые в предыдущем разделе, демонстрируют множество полезных применений неклассических корпусов. В этой связи возникает резонный вопрос: если неклассические корпуса так хороши, почему же они настолько редко — в 15 раз реже, чем НКРЯ, как показывает поиск в Google Scholar — используются в практике исследователей русского языка?

Ответ на этот вопрос отчасти заключается в силе привычки: из всех перечисленных ресурсов НКРЯ появился в публичном доступе первым, и пользователи привыкли именно к нему. Ещё одна причина, разумеется, кроется в более широком хронологическом охвате текстов, присущем НКРЯ, а также в том, что содержащиеся в нём тексты воспринимаются как более нормативные и «качественные», хотя это суждение едва ли доказуемо. Схожие соображения высказывает [Нессет 2019] в своей работе:

> First, as opposed to the Russian National Corpus, the RuTenTen corpus is not balanced, and it is therefore not clear to what extent it is representative of the Russian language. Second, since the RuTenTen corpus is based on data from the internet, we cannot be sure about the quality of the language in the examples,

> which may for instance involve machine translated text of poor quality and other “noise”. Third, while the Russian National Corpus provides important metadata such as the time the example was created, the genre of the text, the name and gender of the author, etc., no such metadata are found in the RuTenTen corpus.

Первая претензия — к репрезентативности и сбалансированности корпуса ruTenTen — представляется весьма показательной. В статье [Беликов и др., 2013] после обстоятельного обсуждения понятий репрезентативности и сбалансированности авторы приходят к заключению:

> [К]орпус считается представительным и сбалансированным тогда, когда на этот счет имеется негласный договор между его создателями и пользователями.

Хотя со времени публикации этой статьи прошло семь лет, более формального определения репрезентативности и сбалансированности не появилось — но, как видим, создателям веб-корпусов так и не удалось вступить в такой негласный договор с лингвистами. В то же время пользователей не смущает, что, например, 899 из 1808 вхождений слова *удлинённый* в НКРЯ приходится на текст Л. Жильцовой «Веснянки (plecoptera). Группа euhologna tha» (2003).

Сверх этого следует отметить ещё одну важную проблему, которая приводит к недостаточному использованию неклассических корпусов: на многих пользователей они производят пугающее впечатление своей технической сложностью (иногда мнимой). Так, для работы с ruTenTen или Araneum Russicum желательно освоить язык запросов CQL и язык регулярных выражений и понимать, какие возможности они предоставляют. В частности, примеры на *с / со* X, приведённые в **разделе 3.1**, можно найти по запросу `[lc="со?"] [lc="щ.*"]`, а примеры конструкции «*ну(,) X и X*», упомянутые в **разделе 3.3**, по запросу `1:[lc="ну"] 2:[lc=","]? 3:[lc!=","] 4:[lc="и"] 5:[] & 3.lc = 5.lc` (1-й токен — *ну*; 2-й токен — запятая, которая может и отсутствовать; 4-й токен — *и*; 3-й токен совпадает с 5-м, но при этом они не являются запятыми). Такие выражения могут показаться громоздкими и непонятными и отбить желание пользоваться неклассическими корпусами. Здесь некоторую помощь может оказать визуальный интерфейс по образцу НКРЯ, который реализован в ГИКРЯ (но не в ruTenTen и Araneum): он позволяет выбирать необходимые характеристики слов и текстов без составления длинных запросов.

Ещё одна сложность неклассических корпусов для пользователя состоит в том, что большие массивы данных предполагают выгрузку и последующую техническую обработку результатов, что тоже не всегда тривиально и требует умения уверенно конвертировать между собой различные форматы данных: например, выгружать данные в формате *.csv, открывать их в Excel и т. п. Естественно, это отнюдь не непреодолимые преграды, однако исследователям порой даже не приходит в голову, что они могут и должны проводить такого рода обработку материала.

5. Заключение

Рассмотрев, насколько часто используются различные корпусные ресурсы в недавних исследованиях по русистике, мы можем прийти к выводу, что «классический» Национальный корпус русского языка применяется на порядок чаще, чем все прочие «неклассические» корпуса, важнейшими из которых являются Генеральный интернет-корпус русского языка, ruTenTen и Araneum Russicum Maximum. Недостаточно широкое использование неклассических корпусов вызывает некоторое огорчение, поскольку эти корпуса могут пролить свет на явления, которые трудно или невозможно изучать с помощью традиционных источников. Примеры таких явлений и возможные пути их исследования приведены в разделе 3. Однако приходится признать, что у недостаточного внимания к неклассическим корпусам есть свои причины, в частности их неудобство для технически неподготовленного пользователя. Хочется надеяться, что в будущем параллельно будут происходить два процесса: адаптация неклассических корпусов к широким кругам пользователей и адаптация пользователей к тому, что работа с корпусом требует определённого уровня технических навыков. Всё вместе это позволит расширить и углубить наши знания о различных явлениях русского языка.

Литература

1. *Benko, V.* (2014), Yet another family of (comparable) Web corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.), Text, Speech and Dialogue. 17th International Conference, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, Cham, pp. 35–60.
2. *Droganova, K. & Zeman, D.* (2016). Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (TR-2016-60; ÚFAL Technical Report). Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.
3. *Endresen, A.* (2019). Ну работа и работа: Russian constructions with reduplication. Novemberseminaret i russisk, UiT Norges Arktiske Universitet. November 29, 2019.
4. *Fillmore, C. J.* (1992). “Corpus linguistics” or “Computer-aided armchair linguistics.” In J. Svartvik (ed.), Directions in Corpus Linguistics, Mouton de Gruyter, Berlin, pp. 35–60.
5. *Nesset, T.* (2019), Big data in Russian linguistics? Zeitschrift Für Slawistik, vol. 64(2), pp. 157–174. https://doi.org/10.1515/slaw-2019-0012.
6. *Schäfer, R., & Bildhauer, F.* (2013), Web corpus construction. Morgan & Claypool. http://dx.doi.org/10.2200/S00508ED1V01Y201305HLT022.
7. *Shavrina, T.* (2018). Differential approach to Web corpus construction. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. http://www.dialog-21.ru/media/4555/shavrinato.pdf.
8. *Shavrina, T., & Benko, V.* (2019), Omnia Russica: Even larger Russian corpus. In В. П. Захаров (ed.), Труды международной конференции «Корпусная лингвистика—2019», Издательство Санкт-Петербургского университета, Санкт-Петербург, pp. 94–102.

9. *Sinclair, J.* (1991), Corpus, concordance, collocation, Oxford University Press, Oxford.
10. *Zakharov, V.* (2013), Corpora of the Russian language. In I. Habernal & V. Matoušek (Eds.), Text, Speech, and Dialogue. Springer, Berlin; Heidelberg, pp. 1–13. https://doi.org/10.1007/978-3-642-40585-3_1.
11. *Аванесов, Р. И.* (ред.) (1988), Орфоэпический словарь русского языка. Произношение, ударение, грамматические формы, Русский язык, Москва.
12. *Беликов, В. И.* (2019), Статистический взгляд на чередование предлогов. VI Международная конференция «Культура русской речи», Москва.
13. *Беликов, В. И., Копылов, Н. Ю., Пиперски, А. Ч., Селегей, В. П., & Шаров, С. А.* (2013), Корпус как язык: От масштабируемости к дифференциальной полноте. Материалы ежегодной международной конференции «Диалог-2013», с. 84–96.
14. *Иткин, И. Б.* (2007), Русская морфонология, Гнозис, Москва.
15. *Копотев, М. В.* (2014), Введение в корпусную лингвистику, Animedia Company, Prague.
16. *Мустайоки, А., & Копотев, М. В.* (2003), Принципы создания Хельсинкского аннотированного корпуса русских текстов ХАНКО в сети Интернет. Научно-техническая информация, 6, с. 33–37.
17. *Перцов, Н. В.* (2006), К суждениям о фактах русского языка в свете корпусных данных, Русский язык в научном освещении, 1(11), с. 227–245.
18. *Пиперски, А. Ч.* (2019), Экспрессивные неологизмы на *-раст* в русском языке // В. Н. Степанов (ред.), Русская грамматика: Активные процессы в языке и речи. Сборник научных трудов Международного научного симпозиума, ЯГПУ им. К. Д. Ушинского, Ярославль, с. 190–195.
19. *Плунгян, В. А.* (2009), Почему современная лингвистика должна быть лингвистикой корпусов. https://polit.ru/article/2009/10/23/corpus/.
20. *Резникова, Т. И., & Копотев, М. В.* (2005), Лингвистически аннотированные корпуса русского языка (Обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005, Индрик, Москва, с. 31–61.
21. *Урысон, Е. В.* (2020), Об одном типе русских предложений тождества (*Платье и платье, ничего особенного*). Конференция в честь 90-летия Ю. Д. Апресяна, Институт русского языка им. В. В. Виноградова РАН, Москва. 04.02.2020.
22. *Хохлова, М. В.* (2016), Обзор больших русскоязычных корпусов текстов. In Компьютерная лингвистика и вычислительные онтологии: Сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург, 22–24 июня 2016 г., Университет ИТМО, Санкт-Петербург, с. 74–77.

References

1. *Avanesov, R. I.* (ed.) (1988), A pronouncing dictionary of Russian. Pronunciation, stress, grammatical forms. [Orfoèpičeskij slovar' russkogo jazyka. Proiznošenie, udarenie, grammatičeskie formy], Russkij jazyk, Moscow.
2. *Belikov, V. I.* (2019), A statistical study of alternations in prepositions [Statističeskij vzgljad na čeredovanie predlogov], 6th International Conference "Culture of Russian Speech", Moscow. February 23, 2019.

3. *Belikov, V. I., Kopylov, N. Yu., Piperski, A. Ch., Selegey, V. P., & Sharoff, S. A.* (2013), Corpus as language: From scalability to differential completeness [Korpus kak jazyk: Ot masštabiruemosti k differencial'noj polnote]. In Computational Linguistics and Intelectual Technologies: Proceedings of the International Conference "Dialogue 2013", 1, pp. 84–96.
4. *Benko, V.* (2014), Yet another family of (comparable) Web corpora. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.), Text, Speech and Dialogue. 17th International Conference, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, Cham, pp. 35–60.
5. *Droganova, K., & Zeman, D.* (2016). Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (TR-2016-60; ÚFAL Technical Report). Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.
6. *Endresen, A.* (2019). Ну работа и работа: Russian constructions with reduplication. Novemberseminaret i russisk, UiT Norges Arktiske Universitet. November 29, 2019.
7. *Fillmore, C. J.* (1992). "Corpus linguistics" or "Computer-aided armchair linguistics." In J. Svartvik (ed.), Directions in Corpus Linguistics, Mouton de Gruyter, Berlin, pp. 35–60.
8. *Itkin, I. B.* (2007), Russian morphophonology [Russkaja morfonologija], Gnozis, Moscow.
9. *Khokhlova, M. V.* (2016), A survey of big corpora of Russian [Obzor bol'šix russkojazyčnyx korpusov tekstov], In Computational Linguistics and Ontologies [Kompjuternaja lingvistika I vyčislitel'nye ontologii], Proceedings of the 19th International Conference "Internet and Modern Society" (IMS-2016), Saint Petersburg, June 22–24, 2016, ITMO University, Saint Petersburg, pp. 74–77.
10. *Kopotev, M. V.* (2014), Introduction to corpus linguistics [Vvedenie v korpusnuju lingvistiku], Animedia Company, Prague.
11. *Mustajoki, A., & Kopotev M. V.* (2003), Principles of construction of the HANCO Corpus on the Web [Principy sozdanija Xelsinkskogo annotirovannogo korpusa russix tekstov HANCO v seti Internet], Scientific and Technical Information [Naučno-texničeskaja informacija], 6, pp. 33–37.
12. *Nesset, T.* (2019), Big data in Russian linguistics? Zeitschrift Für Slawistik, vol. 64(2), pp. 157–174. https://doi.org/10.1515/slaw-2019–0012.
13. *Pertsov, N. V.* (2006), Judgments on the facts of Russian language in the light of corpus data [K suždenijam o faktax russkogo jazyka v svete korpusnyx dannyx], Russian Language and Linguistic Theory [Russkij jazyk v naučnom osveščenii], issue 1(11), c. 227–245.
14. *Piperski, A. Ch.* (2019), Expressive neologisms in -rast in Russian [Èkspressivnye neologizmy na *-rast* v russkom jazyke], In V. N. Stepanov (ed.), Russian grammar: Active processes in language and speech [Russkaja grammatika: Aktivnye process v jazyke i reči], Yaroslavl State Pedagogical University named after K. D. Ushinsky, Yaroslavl, pp. 190–195.
15. *Plungian, V. A.* (2009), Why modern linguistics must be corpus-based [Počemu sovremennaja lingvistika dolžna byt' lingvistikoj korpusov], https://polit.ru/article/2009/10/23/corpus/.

16. *Reznikova, T. I., & Kopotev, M. V.* (2005), Linguistically annotated corpora of Russian: A survey of publicly available resources [Lingvističeski annotirovannye korpusa russkogo jazyka (Obzor obščedostupnyx resursov)], In Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, pp. 31–61.
17. *Schäfer, R., & Bildhauer, F.* (2013), Web corpus construction. Morgan & Claypool. http://dx.doi.org/10.2200/S00508ED1V01Y201305HLT022.
18. *Shavrina, T.* (2018). Differential approach to Web corpus construction. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". http://www.dialog-21.ru/media/4555/shavrinato.pdf.
19. *Shavrina, T., & Benko, V.* (2019), Omnia Russica: Even larger Russian corpus. In В. П. Захаров (ed.), In Proceedings of the International Conference "Corpus Linguistics'2019", Saint Petersburg State University, Saint Petersburg, pp. 94–102.
20. *Sinclair, J.* (1991), Corpus, concordance, collocation, Oxford University Press, Oxford.
21. *Uryson, E. V.* (2020), On a certain type of Russian equative constructions (*Plat'je i plat'je, ničego osobennogo* 'Just a dress, nothing more') [Ob odnom tipe russkix predloženij toždestva (*Plat'je i plat'je, ničego osobennogo*)]. Conference on the occasion of the 90th birthday of Yu. D. Apresyan. V. V. Vinogradov Institute of Russian Language, Moscow, February 04, 2020.
22. *Zakharov, V.* (2013), Corpora of the Russian language. In I. Habernal & V. Matoušek (Eds.), Text, Speech, and Dialogue. Springer, Berlin; Heidelberg, pp. 1–13. https://doi.org/10.1007/978-3-642-40585-3_1.