***С.О. Савчук***

***S.O. Savchuk***

**НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА
В ЗЕРКАЛЕ СТАТИСТИКИ** [[1]](#footnote-1) **THE RUSSIAN NATIONAL CORPUS
through the prism of STATISTICS**

**Аннотация.** Основными критериями оценки текстовых корпусов, принятыми в корпусной лингвистике, являются объем, репрезентативность, сбалансированность. Особенностью НКРЯ, как большого представительного корпуса, является то, что за все время его существования рост объема осуществлялся пропорционально, с соблюдением баланса в функциональном разнообразии текстов. В связи с планами дальнейшего развития корпуса был проведен масштабный анализ состава текстов НКРЯ по основным параметрам, которые традиционно используются для оценки репрезентативности корпусов. Среди них дата создания текстов, сфера функционирования, жанр и тематика текстов. В статье представлены результаты анализа статистической информации об основных метатекстовых параметрах корпуса.

**Ключевые слова.** *Национальный корпус русского языка, состав текстов, репрезентативность, сбалансированность.*

**Abstract.** The main criteria for evaluating text corpora accepted in corpus linguistics are corpus size, representativeness, and balance. A peculiarity of the RNC design as a large representative corpus is that its size increasing was proportional, while maintaining a balance in the functional diversity of texts. A large-scale analysis of text composition of the RNC was carried out using the parameters traditionally applied for evaluating corpora representativeness. Among them are date of text creation, functional sphere, genre, and text topic. The article presents the results of the analysis of statistical information about the main metatextual parameters of the RNC.

**Keywords.** *The Russian National Corpus, corpus composition, representativeness*, *balance.*

**1. Введение**

Национальный корпус русского языка проектировался в начале 2000 годов как корпус современного русского языка по образцу больших национальных корпусов объемом 100 млн словоупотреблений. Это особый тип большого представи­тельного корпуса, отражающий функционирование националь­ного языка в современный период. За 20 лет существования ядро корпуса увеличилось в объеме и обросло системой специальных корпусов, но при этом, согласно статистике запросов, основной корпус письменных текстов остается самым востребованным корпусом в составе НКРЯ. Но за это время в корпусной лингвистике была налажена технология создания интернет-корпусов очень большого объема, от одного до десятков миллиардов, так что на современном этапе НКРЯ сосуществует с другими корпусами русского языка, значительно превосхо­дящими его по объему. Среди них ГИКРЯ [Беликов и др. 2013], семейство корпусов Araneum Russicum [Benko 2014], корпуса, доступные в системе Sketch Engine, например, RuTenTen [Jakubíček et al. 2013], корпуса русских книг в составе Google Books [Захаров и др. 2014]. Первоначальное увлечение гигант­скими объемами веб-корпусов сменилось более взвешен­ной позицией. Специальные исследования, проведенные на материале корпусов разных типов, объективно, на наш взгляд, оценивают преимущества и недостатки каждого типа [Захаров 2015; Пиперски 2020, Davies 2019, Nesset 2019; Sharoff 2020 и др.].

В связи с возникшей возможностью и необходимостью сравнения разных корпусов в зоне внимания опять оказались такие критерии оценки корпусов, как репрезентативность и сбалансированность. Имеются разные подходы к определению репрезентативности. В частности, есть мнение, что под репрезентативностью пони­мается способность корпуса отражать свойства всей популяции текстов, но поскольку мы не знаем этих свойств в полном объеме, ни один корпус не является зеркалом того, с чем мы сталкиваемся в реальной жизни, поэтому репрезентативность – это внутреннее свойство корпуса. Согласно другой позиции, репрезентативность – это способность корпуса представлять максимальное разнообразие текстов на каком-либо языке. Предлагается ориен­тированный на восприятие взгляд на репрезентативность, хотя замечают, это понятие невозможно рассчитать и описать строго математически. В целом в определении репрезентативности исследователи сходятся в том, что это нечто недостижимое, к чему нужно стремиться [Viana et al*.* 2015: 3-4, 66-68, 102, 118, 160]. Сбалансированность – другая сторона репрезентативности. Это свойство корпуса представлять разнообразные тексты в определенных заданных пропорциях помогает, в частности, избежать субъективизма при отборе текстов при составлении корпуса.

НКРЯ предназначен как для исследований современного языка, так и для изучения микродиахронических изменений. Обладает ли он необходимыми свойствами для этого? Согласно М. Дэвису, корпус для изучения языковых изменений должен соответствовать нескольким критериям. Он должен 1) иметь большой объем (не менее 100 млн словоупотреблений), что позволяет отслеживать редкие языковые явления; 2) включать современные тексты (желательно пополняться ежегодно); 3) содержать тексты, относящиеся к разным жанрам, в определенном соотношении; 4) сохранять жанровый баланс от периода к периоду; 5) иметь архитектуру, которая позволяла бы прослеживать частотные характеристики слов в разные периоды и сравнивать их между собой [Davies 2012].

Как представляется, НКРЯ отвечает всем этим требованиям, что будет показано в дальнейшем. Его текущий объем составляет более 320 млн словоупотреблений, то есть почти в 10 раз превышает первоначальный объем при открытии ресурса на сайте ruscorpora.ru. При этом увеличение объема осуществлялось пропорционально, с соблюдением баланса в функциональном разнообразии текстов.

2. Анализ текущего распределения текстов основного корпуса

В связи с проводимой в настоящее время системными и инфраструктурными изменениями в программном обеспечении корпуса и разработкой перспектив его дальнейшего развития было предпринято исследование текстового состава корпусов по основным метатекстовым параметрам, служащим для оценки его репрезентативности. Программное обеспечение для анализа разработано А. Хаджийской. Были получены данные о распределении текстов по дате создания, сферам функциони­рования, жанрам художественной литературы, типам и основной тематике по состоянию на декабрь 2020 г. Ниже перечислены результаты анализа статистической информации об основных метатекстовых параметрах корпуса письменных текстов.

*2.1. Распределение по дате создания*

Поскольку основной корпус используется для диахро­ни­че­ских исследований, он должен давать возможность прослеживать изменения в значениях слов, сочетаемости и пр. на протяжении определенных периодов. Поэтому в идеальном случае каждый период должен быть представлен приблизительно равным количеством словоупотреблений. На уровне подкорпусов, из которых исторически складывался основной корпус, ‑ XIX-го, 1-й пол. XX-го, 2-й пол. XX-го и XXI-го веков – картина близка к идеальной (см. Рис. 1).



*Рис. 1*. Распределение текстов основного корпуса письменных текстов по периодам

Однако если увеличить масштаб и рассмотреть распределение текстов по 10-летиям, то его нельзя назвать сбалансированным (Рис. 2). Подкорпус XVIII-го века пока очень мал и пополняется медленно из-за сложности подготовки текстов. Возможно, его следует рассматривать отдельно, а для сравнительного анализа использовать больший хронологический интервал, например: 1700-1730, 1730-1760, 1760-1780, 1780-1800.

*Рис. 2*. Распределение текстов основного корпуса письменных текстов по дате создания.

Этому есть свое объяснение. Пик на графике, который приходится на начало 2000-х годов, соответствует той текстовой коллекции, с которой начинался НКРЯ как корпус современных текстов. Дальнейшее развитие корпуса – составление корпусов предшествующих периодов, которое еще не завершено. Особого внимания требуют тексты 1800-1830-х годов: по сложности подготовки они мало отличаются от текстов XVIII в. Другая задача, которую предстоит решить, – создание полноценной коллекции текстов 2010-2020 гг., сбалансированной по разным функциональным сферам.

*2.2. Распределение по сферам функционирования*

Центральным элементом организации текстового состава НКРЯ является распределение текстов по функциональным сферам. Это самая общая типологическая характеристика текста, которая определяет отнесение текста к одной из социально значимых областей общественно речевой практики и используется в соответствии с традицией отечественной функциональной стилистики и типологии текста [Савчук 2005: 70-71]. Для описания текстов НКРЯ выделено 9 функциональных сфер: художественная, публицистика (и массовая информация), церковно-богословская, учебно-научная, официально-деловая, производственно-техническая, реклама, бытовая, электронная коммуникация. Отметим, что содержательно это деление соответствует тому, что называется жанровой классификацией в других национальных корпусах (ср. «художественная литература», «журналы», «газеты», «академические тексты», «разговорные тексты» в британском корпусе или «художественная литература», «журналистика», «научные» и «административные тексты» в чешском корпусе). Результаты распределения текстов по речевым сферам в статике представлены в таблице 1.

*Таблица 1*. Распределение текстов по сферам функционирования

|  |  |  |
| --- | --- | --- |
| **Сфера функционирования** | **Кол-во словоупотр** | **Доля в %** |
| Художественная | 134963105 | 42,75% |
| Публицистика | 123213170 | 39,03% |
| Учебно-научная | 34956396 | 11,07% |
| Официально-деловая | 4653591 | 1,47% |
| Производственно-техническая | 1469354 | 0,47% |
| Церковно-богословская | 4282472 | 1,36% |
| Реклама | 713666 | 0,23% |
| Бытовая | 10915179 | 3,46% |
| Электронная коммуникация | 3534997 | 1,12% |
| Всего словоформ | 315691990 | 100.95% |

Для того чтобы можно было сравнивать частотные характеристики единиц в разные периоды, необходимо, чтобы корпус сохранял примерно одинаковое соотношение текстов, относящихся к разным функциональным сферам в разные периоды. В этом случае различия в частотных характеристиках языковых единиц будут иметь языковую природу, а не объясняться несбалансированностью состава корпуса.

Таблица 2 объединяет две характеристики корпуса и представляет распределение текстов по сферам функциониро­вания в динамике. Соотношение текстов разных функциональных сфер показано в разные периоды – в XIX в., в 1‑й пол. XX в., во 2-й пол. XX в., в нач. XXI в**.**

*Таблица 2*. Распределение текстов разных сфер функционирования по периодам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Сферы функционир. | 1800-1899 | 1900-1950 | 1951-1999 | 2000-2020 |
| Художественная | 49,94% | 44,31% | 49,28% | 31,03% |
| Публицистика | 32,96% | 36,48% | 37,13% | 49,41% |
| Учебно-научная | 10,49% | 11,63% | 8,30% | 12,35% |
| Офиц.-деловая | 1,05% | 1,50% | 0,86% | 1,32% |
| Произв.-техн | 0,02% | 0,42% | 0,88% | 0,42% |
| Церк.-богосл | 0,79% | 1,44% | 2,00% | 0,31% |
| Реклама | 0,01% | 0,06% | 0,09% | 0,69% |
| Бытовая | 6,35% | 5,52% | 1,96% | 0,42% |
| Электр. комм. | 0,00% | 0,00% | 0,00% | 4,32% |
| Всего словоформ | 64314169 | 85872427 | 86120684 | 81804968 |

Как видно из таблицы, в целом баланс распределения текстов по сферам функционирования сохраняется в каждом из выделенных периодов, за исключением отдельных отклонений, которые предстоит исправить.

Для более детального анализа получены данные распределения текстов по десятилетиям каждого периода – как для всего корпуса, так и для каждой текстовой коллекции, представляющей разные функциональные сферы. Анализ позволяет найти участки, которые в первую очередь нуждаются как в количественном пополнении, так и в выравнивании соотношения текстов разных функциональных сфер.

*2.3. Соотношение художественных текстов по жанрам
в разные периоды*

Было исследовано более мелкое членение текстов корпуса на жанры и типы. Для художественных текстов важна отнесенность текста к литературному жанру, она во многом определяет тематическое содержание и языковые особенности текста (ср., например, разницу в колорите научно-фантастической повести и документальной прозы).

В разные периоды соотношение по жанрам должно быть ожидаемо разным, поскольку жанры возникают, развиваются и исчезают или трансформируются. Например, детектив как жанр в русской литературе появился в XX в., его предшественники в XIX в. – авантюрный роман (например, «Петербургские трущобы» В. Крестовского) и судебный очерк. Как показал анализ жанрового состава основного корпуса, нежанровая проза (которая в электронных библиотеках маркируется как «современная проза»), в XVIII-XIX вв. преобладает над всей жанровой, вместе взятой (62,6%), и составляет около 50% в XX-XXI вв. Объясняется это тем, что классификация жанров в корпусе отражает жанровый состав современной литературы, а в XIX в. эта система была иной, какие-то жанры только склады­вались в пределах нежанровой прозы. Жанровое соотношение художественных текстов корпуса в разные периоды представлено в таблице 3.

*Таблица 3*. Соотношение художественных текстов по жанрам в разные периоды

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Жанры худож. лит-ры | **1800-1899** | **1800-1899, %** | **1900-1949** | **1900-1949, %** |
| Детектив, боевик | 197022 | 0,61% | 463317 | 1,22% |
| Детская | 402639 | 1,25% | 2133894 | 5,61% |
| Документ. проза | 1438694 | 4,48% | 3719641 | 9,78% |
| Историч. проза | 4645672 | 14,46% | 5967662 | 15,68% |
| Любовная история | 185250 | 0,58% | 229501 | 0,60% |
| **Нежанр. проза** | **20112241** | **62,61%** | **20444581** | **53,73%** |
| Приключения | 1244519 | 3,87% | 2476796 | 6,51% |
| Фантастика | 474126 | 1,48% | 1821586 | 4,79% |
| Юмор и сатира | 2252473 | 7,01% | 924997 | 2,43% |
| Драматургия | 1351398 | 4,21% | 999619 | 2,63% |
| Прочее | 511438 | 1,59% | 38091 | 0,10% |
| Всего | 32121476 | 102,16% | 38050694 | 103,07% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Жанры худож. лит-ры | **1950-1999** | **1950-1999, %** | **2000-2020** | **2000-2020, %** |
| Детектив, боевик | 1816749 | 4,28% | 5338492 | 21,03% |
| Детская | 3018028 | 7,11% | 510539 | 2,01% |
| Документ. проза | 5624811 | 13,25% | 1238415 | 4,88% |
| Историч. проза | 2083389 | 4,91% | 426530 | 1,68% |
| Любовная история | 139553 | 0,33% | 668577 | 2,63% |
| **Нежанр. проза** | **22315795** | **52,59%** | **14304336** | **56,36%** |
| Приключения | 1398094 | 3,29% | 130203 | 0,51% |
| Фантастика | 4184241 | 9,86% | 2200695 | 8,67% |
| Юмор и сатира | 1353040 | 3,19% | 855915 | 3,37% |
| Драматургия | 861158 | 2,03% | 232524 | 0,92% |
| Прочее | 38189 | 0,09% | 304560 | 1,20% |
| Всего | 42436986 | 100,93% | 25380417 | 103,27% |

*2.4. Соотношение нехудожественных текстов по типам*

Для корпуса значимо соотношение нехудоже­ственных текстов по типам текстов. В НКРЯ используется подробная система кодирования текстовых типов, привязанная к различным сферам функционирования. Первоначальный список типов был открытым: ограничения на состав типов не делалось, поскольку производилось сплошное описание текстов, а не отбор текстов по заданному списку типов. Предполагалось, что помимо распро­страненных типов в корпус будут попадать и менее частотные, так что список будет расширяться и со временем стабили­зируется. Так и произошло, и в настоящее время мы имеем некоторый основной набор типов, представленный на панели отбора текстов, описывающий подавляющее большинство текстов. И есть некоторый набор типов, представленный малым количеством и даже единичными текстами.

Проведенный анализ позволил оценить действующий стандарт разметки текстовых типов и определить, какие типы текстов в пределах каждой функциональной сферы плохо представлены (менее 100 текстов). В частности, в сфере публицистики недостаточно представлены такие публицисти­ческие жанры (типы), как *автобиография* (33), *аннотация* (54), *блог* (18), *доклад* (85), *заявление* (55), *календарь* (14), *лекция* (19), *листовка* (35), *манифест* (13), *описание* (18), *памфлет* (7), *послание* (17), *пресс*-*конференция* (2), *путеводитель* (12), *рецепт* (66). В учебно-научной сфере меньше 100 текстов приходится на типы: *доклад* (59), *задача* (30), *инструкция* (44), *конспект* (20), *лекция* (49), *обзор* (35), *отзыв* (5), *отчет* (80), *справочник*(3), *тезисы* (3), *учебник* (20), *учебное* *пособие* (61), *методические* *материалы* (54) и др. Составлен список таких типов текстов для каждой функциональной сферы. Поиск, обработка и включение в состав корпуса текстов указанных типов рассматривается как одна из первоочередных задач пополнения основного корпуса.

*2.5. Соотношение нехудожественных текстов по тематике*

Включение в состав НКРЯ текстов всех сфер функци­они­рования обеспечило тематическое разнообразие текстов, поскольку каждая сфера бытования языка связана с определен­ным набором тематик. Наиболее широк этот набор в сфере публицистики, всеобъемлющ в художественной сфере (настоль­ко, что кодирование по этому параметру становится бессмыслен­ным) и требует мониторинга и коррекции в учебно-научной сфере, чтобы при отборе текстов обеспечить представленность в корпусе всех отраслей науки в равной мере.

Анализ корпуса по данному параметру позволил определить, какие тематики в пределах каждой функциональной сферы плохо представлены (менее 100 текстов). В частности, в официально-деловой сфере подавляющее число текстов относятся к темати­че­ским областям *политика и общественная жизнь* (913), *админи­страция и управление* (516), *армия и вооруженные конфликты* (480), *бизнес, коммерция, экономика, финансы* (199), и напротив, слабо представлены такие тематики, как *сельское хозяйство* (8), *спорт* (4), *строительство и архитектура* (11), *техника* (13), *транспорт* (12) и др. Поиск, обработка и включение в состав корпуса официальных документов, относящихся к тематическим областям, недостаточно представленным в коллекции деловой речи, будет учтено при составлении плана пополнения основного корпуса. Аналогичные списки и планы пополнения составлены для других функциональных сфер.

Остановимся подробнее на соотношении тематических областей, связанных с разными науками. Отдельно оценивалось общее состояние по всему массиву текстов и соотношение в текстах с 2010 года. Наилучшим образом представлены гумани­тарные науки (*история, политология, религиоведение, право, образование, психология, культурология*), а из прикладных и естественных наук – *медицина, военное дело, техника, биология, геология* и *география*. Сравнительно небольшой объем текстов (до 2 млн словоупотреблений) приходится на такие области, как *химия, физика, строительство и архитектура, сельское хозяй­ство, транспорт*. Совсем слабо, до 1 млн словоупотреблений, представлены *астрономия, информатика, энергетика, стати­стика* и *лесное хозяйство*.

Что касается текстов, созданных после 2010 года, то первая группа научных областей, наиболее полно представленных в текстах последнего десятилетия, по составу в основном совпадает с общей по корпусу. Однако количество тематических областей, недостаточно представленных в новейших текстах, несколько больше – к *информатике*, *статистике*, *лесному хозяйству*, присоединились *геология* и *география*, *математика*, *химия*, *сельское хозяйство, искусствоведение*. На основе оценки состояния были установлены целевые показатели для репрезен­тации разных наук в пределах учебно-научного стиля. Ситуация с современными научными текстами будет исправлена в ближай­шее время, с включением в состав корпуса большой коллекции текстов научных журналов и популярных научных сайтов (Arzamas, postnauka.ru, polit.ru, pro-science.ru, indicator.ru и др.).

3. Заключение

Проведенный анализ основного корпуса письменных текстов по основным метатекстовым параметрам показал, что по состоянию на начало 2021 года состав корпуса отличается достаточным разнообразием. Что касается распределения по хронологической оси, то в количественном отношении удовле­творительно представленным можно считать период со 2‑й пол. XIX в. до наших дней: в этом интервале объем подкорпуса за каждое 10‑летие превышает 10 млн словоупотреблений, причем в отдельные 10‑летия многократно. Задача ближайшего будущего состоит в том, чтобы сделать это распределение более равномерным. Особого внимания требуют подкорпус текстов раннего периода – XVIII – 1-й пол. XIX вв., столь важный для диахронических исследований, а также коллекция новейших текстов, отражающих текущий момент. Если вторая задача решается достаточно просто, то первая, при соблюдении высоких стандартов качества текстов, требует значительных усилий и времени.

Что касается объема корпуса, то не следует забывать о растущем газетном корпусе (период 2000-2019 гг., объем, включая региональные газеты, 350 млн словоупотреблений) и корпусе устных текстов (13,3 млн словоупотреблений), которые представляют современный язык. Историческую часть дополняет старорусский корпус (8 млн словоупотреблений) и поэтический корпус, в котором значительную долю составляет поэзия XIX в. (около 5 млн словоупотреблений). Организация кросс-корпусного поиска объединит все эти коллекции в один массив (объемом около 700 млн словоупотреблений), тем самым частично решив проблему увеличения размера корпуса.

Для того чтобы при пополнении корпуса новыми текстами добиться равномерного распределения по основным текстовым параметрам в пределах каждого периода, предстоит сочетать две стратегии. Традиционная стратегия «естественного» прироста, которая при отборе текстов опирается на библиографические критерии (автор, издание, год издания) без учета жанровой или тематической принадлежности, будет дополнена стратегией «искусственной поддержки» редких жанров, при которой тексты отбираются по текстологическим критериям в заданном количестве, что поможет повысить представленность в корпусе отдельных текстовых типов и жанров. Распределение текстов по функциональным сферам будет осуществляться в соответствии с целевыми показателями баланса корпуса, что обеспечит пропорциональное соотношение текстов разных функциональных сфер по периодам.

Подобная программа развития корпуса обеспечит поступа­тельный рост его объема, который в обозримом будущем может достигнуть 1 млрд словоупотреблений. Что касается «больших данных», то они остаются прерогативой интернет-корпусов русского языка, развитие которых идет как по пути наращивания объемов, так и усложнения структурной организации данных.

**Литература**

*Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А.* (2013)Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог-2013». Вып. 12 (19). Т. 1. С. 84–95.

*Захаров В. П*. (2015) Сочетаемость через призму корпусов // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». Вып. 14(21), М.: РГГУ. С. 667–682.

*Захаров В.П., Масевич А.Ц*. (2014) Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. // Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С. Петерб. ун-та, 2014. С. 303–327.

*Пиперски А. Ч.* (2020) Русский язык и корпусное разнообразие // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». Вып. 19 (26). С. 84–95.

*Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horak, I. Kopeček and K. Pala (Eds.): Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno. Czech Republic. September 8–12 2014. Springer International Publishing, pp. 257–264.

*Davies, M.* (2012) Looking at Recent Changes in English with the Corpus of Contemporary American English (COCA). 21st Century Text. (Peer-reviewed, online journal).

*Davies, M.* (2019) Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In From data to evidence in English language research (Digital Linguistics), eds. Suhr, Carla, Terttu Nevalainen and Irma Taavitsainen. Leiden: Brill. 66–87.

*Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V*. (2013), The TenTen Corpus Family. In Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.

*Nesset, T.* (2019),Big data in Russian linguistics? Another look at paucal constructions. Zeitschrift für Slawistik 2019; 64(2): 157–174

*Sharoff S.* (2020) Topography of internet corpora // Компьютерная лингвистика и интеллектуальные технологии. Вып. 19 (26). Доп. том. С. 1134–1137.

*Viana V., Zyngier S., Barnbrook G.* (eds.). (2011), Perspectives on Corpus Linguistics: Connections & Controversies. Philadelphia: John Benjamins.

**References**

*Belikov V. I., Kopylov N. Yu., Piperski A. Ch., Selegej V. P., Sharov S. A.* (2013) Korpus kak yazyk: ot masshtabiruemosti k differencialnoj polnote [Corpus as language: from scalability to register variation] // Кomp'juternaja lingvistika i intellektual'nye tehnologii: Materialy mezhdunarodnoj konferentsii “Dialog–2013” [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference “Dialog–2013”]. Vol. 12(19), pp. 84–95.

*Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horak, I. Kopeček and K. Pala (Eds.): Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno. Czech Republic. September 8–12 2014. Springer International Publishing, pp. 257–264.

*Davies M.* (2012), Looking at Recent Changes in English with the Corpus of Contemporary American English (COCA). 21st Century Text. (Peer-reviewed, online journal).

*Davies M.* (2019), Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design." In From data to evidence in English language research (Digital Linguistics), eds. Suhr, Carla, Terttu Nevalainen and Irma Taavitsainen. Leiden: Brill. Pp. 66-87.

*Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V*. (2013), The TenTen Corpus Family. Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.

*Nesset T.* (2019),Big data in Russian linguistics? Another look at paucal constructions. Zeitschrift für Slawistik; 64(2), pp. 157–174.

*Piperski A. Ch*. (2020), Russkij yazyk i korpusnoe raznoobrazie [Russian language and corpus diversity]. Кomp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii “Dialog” [Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”]. Vol. 19 (26), pp. 84-95.

*Sharoff S.* (2020), Topography of internet corpora. Кomp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii “Dialog” [Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”]. Vol. 19 (26), pp. 1134-1137.

*Viana V., Zyngier S., Barnbrook G.* (eds.). (2011), Perspectives on Corpus Linguistics: Connections & Controversies. Philadelphia: John Benjamins.

*Zakharov V. P*. (2015), Sochetaemost cherez prizmu korpusov [Set Phrases: a View through Corpora]. Кomp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii “Dialog” [Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”]. Vol. 14 (21), pp. 667-682.

*Zakharov V. P*., *Masevich A. C*. (2014), Diahronicheskie issledovaniya na osnove korpusa russkih tekstov Google Books Ngram Viewer [Diachronic investigations on the base of Russian corpus of Google Books Ngram Viewer] Strukturnaya i prikladnaya lingvistika [Structural and applied linguistics]. Iss. 10. St. Petersburg: St. Petersburg University Publishing house, pp. 303-327.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Савчук Светлана Олеговна**

Институт русского языка им. В.В. Виноградова РАН

**Svetlana Savchuk**

V.V. Vinogradov Russian Language Institute, RAS

***E-mail:*** ***savsvetlana@mail.ru***

1. Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2020-793. [↑](#footnote-ref-1)