***Н.Л Аванесян, А.М. Чеповский, Т.Ю. Шерстинова,***

 ***Ф.Н. Соловьев, Д.Ю. Чуйкин***

**Корреляционный анализ частотных словарей лингвистических показателей русской прозы 1900-1930 гг. в динамике[[1]](#footnote-1)**

**Аннотация.** Проведен сравнительный анализ текстовых подкорпусов че­ты­рех периодов русской литературы 1900­-1930 с использованием коэффи­циентов попарной ранговой корреляции частотных словарей лекси­ческих характеристик. Показано, что социальные катаклизмы в мень­шей степени повлияли на употребления глаголов, чем прилагательных и суще­стви­тель­ных. Установлено, что другими показателями, отличающими рассма­три­ваемые времен­ные подкорпуса художественной прозы, являются именные и глагольные группы, а также частотные после­довательности символов.

**Ключевые слова.** Ранговая корреляция, корпусная лингвистика, именные группы (ИГ), глагольные группы (ГГ), литературные тексты.

1. . Введение

Корреляционный анализ традиционно используется в каче­ст­ве меры, позволяющей количественно оценить связь между двумя непрерывными переменными. Предлагаемое исследование напра­вле­но на изучение различных исторических периодов русской прозы 1900-1930 гг., а именно — на сопоставление трех последо­вательных периодов русской литературы. Оно основано на корре­ля­­ционном анализе количественных текстовых показателей, пред­ставленных в виде частотных словарей различных лингвис­тичес­ких данных (в частности, частей речи (ЧР), именных и глагольных групп, псевдооснов слов и т. д.). Целью исследования является оценка эффективности этих количественных методов анализа корпуса художественных текстов.

Материалом послужил корпус русских рассказов первой тре­ти ХХ века, разрабатываемый на филологическом факультете СПбГУ при участии департамента филологии НИУ ВШЭ в Санкт-Петербурге [1-4]. Корпус предназначен для проведения стилемет­рических исследований русской прозы, а также для изучения изменений, произошедших в русском языке в эпоху революций.

Литературные произведения как нельзя лучше отражают со­бы­тия и изменения во всех сферах жизни человека: в культурной, политической, социальной и даже бытовой. В отличие от тради­ционных корпусов, ориентированных на тексты, написанные выдающимися писателями, корпус русских рассказов 1900-1930 предназначен для изучения языка всей лите­ратурной системы рассматриваемой эпохи, поэтому в него вклю­чены рассказы, написанные многими забытыми и второсте­пен­ными писателями. Таким образом, он позволяет в полной мере использовать весь потенциал художественной литературы для исследований в области корпусной лингвистики [3].

Для аннотированной части корпуса было отобрано 300 рос­сийских писателей из общего списка авторов, насчитывающего бо­лее 2800 персоналий. Общий объем аннотированной части состав­ляет 310 рассказов, всего 1 млн словоупотреблений. Корпус разде­лен на 4 временных периода: 1) начало XX века до 1913 г., 2) Пер­вая мировая война до революций (1914-1916), 3) революционные годы (1917-1922) и 4) раннесоветский период (1923-1930). Корпус охватывает рассказы представительного числа авторов для каж­дого периода времени, что позволяет проводить лингвистический и статистический анализ языка и стиля.

В работе [5] для исследования корпуса русских рассказов был использован метод сравнительного анализа корпусов текстов, который позволяет выявить неявные связи между корпусами раз­но­­родных текстов. В основе данной методики лежат методы кор­пусного анализа, реализованные в корпусной платформе TXM: ана­лиз специфичности, позволяющий создать своеобразный «про­филь» подкорпуса на основе определенного свойства (тематика текста, психологическая направленность текста) путем выявления наиболее характерных или нехарактерных для него признаков (начальных форм слова, псевдооснов). В данном исследовании мы применяем метод ранговой корреляции для сравнения частотных словарей различных лексических характеристик подкорпусов.

Представленная методология корпусного анализа может быть использована для сравнения любых текстовых репрезентативных наборов данных.

1. . Метод. Техники анализа текста

Характеристики текстов в данном исследовании определя­лись процедурами автоматизированной обработки текстов на естественных языках, описанными в [6, 7].

Автоматический морфологический анализ словоформ прово­дил­ся на основе компьютерных методов морфологии. Исполь­зуемая для анализа морфологическая модель [7] относит каждое слово к одному из 24 морфологических классов, которые помимо «традиционных» частей речи включают такие категории, как «неизменяемое слово», «аббревиатура», «топоним» и т. д. Каждый из этих морфологических классов характеризуется набором грам­матических характеристик: род, падеж, число, наклонение и т. д. Каждая словоформа содержит свои грамматические характе­ристики и свою каноническую (начальную) форму.

Кроме того, в тексте выделялись именные и глагольные груп­пы слов. Под именной группой (ИГ) мы понимаем группу слов, в которой главным словом является существительное, а другие слова связаны с ним подчиненными синтаксическими связями. При выявлении именных групп решалась задача снятия омони­мической неопределенности, возникающая в результате множе­ственности морфологических анализов отдельных употреблений слов. Методика определения именных групп основана на рас­смотрении всего набора возможных морфологических интер­претаций каждого слова.

Глагольные группы (ГГ) — это словосочетания, основным словом которых является глагол. Связи именных групп с глаго­лами построены на синтаксическом анализе предложения. Гла­голь­ное управление определяется как разновидность синтак­си­ческой подчинительной связи типа управления, в которой главным словом является глагол. Анализируя глагольное управление ос­нов­ного слова (глагола), накладываются ограничения на употре­бление зависимого словосочетания в виде набора вариантов допус­тимых комбинаций грамматических характеристик зависи­мого словосочетания. Анализ глагольного употребления основан на датировке электронного словаря глагольного управления, кото­рый включает первые две тысячи наиболее часто встречающихся глаголов русского языка. В отличие от отдельных слов, выде­ленные именные и глагольные группы несут информацию о конкретных аспектах текстового содержания.

В качестве еще одной лингвистической характеристикой тек­ста рассматривалась псевдооснова, т.е. часть слова, не содержащая ни суффиксов, ни префиксов. Метод автоматического выбора псев­дооснов заключается в сравнении рассматриваемой слово­фор­мы с допустимым в языке набором структур некорневой части слова [7]. Псевдооснова слова выделяется отбрасыванием всех аффиксов, соответствующих определенной структурной схеме, которые описывают максимальное количество комбинаций пре­фик­сов и суффиксов, разрешенное в данном языке. Метод псев­дооснов позволяет анализировать текстовые структуры, а не только точные словоформы.

Сравнительный анализ подкорпусов проводился преимуще­ственно путем попарного сравнения частотных словарей различ­ных лексических характеристик, составленных для исследуемых подкорпусов. Чтобы оценить близость частотных словарей, ранги записей словаря устанавливаются после сортировки характерис­тики, внесенной в словарь, по частоте встречаемости. Сравнение словарей производится путем расчета коэффициента попарной ранговой корреляции для каждой пары словарей разных подкор­пусов. Словарные записи считаются случайными величинами. Связь между наборами таких элементов различных словарей определяется как коэффициент попарной ранговой корреляции значений этих случайных величин.

Коэффициенты попарной ранговой корреляции для любых двух частотных словарей определялись с учетом элементов с одинаковыми частотами. Поскольку размеры словарей могут быть довольно большими, учитывались только первые (в порядке убы­вания частоты) 10 000 записей в каждом из словарей. Если какая-то лексическая характеристика встречается только в одном слова­ре и не встречается в другом, то при вычислении коэф­фициента корреляции мы принимаем ее частоту во втором словаре равной 0.

Коэффициент попарной ранговой корреляции принимает зна­че­ния из интервала [-1; 1]. Значения, близкие к 1, указывают на мо­но­тонную непротиворечивость словарей: если в одном словаре в паре слов одно из них имеет частоту выше, чем другое, то во втором словаре это слово также имеет частоту выше, чем другое. И это правило действует для всех пар слов. Значения, близкие к 1, указывают на обратный эффект: если в одном частотном словаре какое-то слово имеет более высокую частоту, чем какое-то другое слово, то — наоборот — во втором частотном словаре его частота тогда будет ниже; это правило также действует для всех пар слов. Если значение близко к 0, то словари несовместимы: ранговой связи между частотами слов в двух словарях нет.

1. . Сравнительный анализ частотных словарей

Сравнительный анализ хронологических подкорпусов корпу­са рассказов был проведен с использованием частотных словарей лексических характеристик, полученных методами, описанными в предыдущем разделе. Методами автоматического анализа были со­ставлены частотные словари частей речи (существительные, гла­голы, прилагательные) и словосочетаний (именные и гла­гольные группы), размеры которых составляют от 10000 до 60000.

Общая характеристика частотных словарей частей для изу­чаемого временного интервала приведена в статье [8]. Результаты сравнения частотных словарей существительных, прилагательных и глаголов с помощью коэффициента попарной ранговой корре­ляции показывают, что с точки зрения сравнения рангов слов раз­личных частей речи в разных подкорпусах исследуемого корпуса художественных текстов наблюдается слабая положительная ран­говая корреляция для существительных (0,4-0,5) и несколько бо­лее выраженная корреляция между рангами прилагательных (0,5-0,6) и глаголов (0,6-0,7). Таким образом, можно сделать вывод, что социальные катаклизмы в меньшей степени повлияли на частоты употребления глаголов, в то время как существительные показы­вают большие изменения. Это подтверждается выводами, сделан­ными ранее на этом материале [8].

Интересно, что значение коэффициента ранговой корреляции в случае имен­ных и глагольных групп достаточно мало — лежит в интервале [-0.1;0.5]. Для именных групп его значение относи­тельно ста­бильно на уровне слабой положительной корреляции (0,2-0,4), в то время как корреляция между глагольными группами отдельных периодов фактически нулевая (минимум наблюдается для двух групп значений — между советским периодом и началом века, а также между военным и революционным периодом). Интер­пре­тация этих данных требует более детального анализа. Однако мож­но предположить, что это указывает на возможность определения конкретных словосочетаний для каждого подкорпуса и возмож­ность рассмотрения словосочетаний в качестве отличи­тельных признаков при идентификации рассказов для опреде­ленного исторического периода, особенно для глагольных групп.

Сравнение частотных словарей буквосочетаний проводилось для последовательностей из 3-6 букв. Словари буквенных соче­таний длиной 3 практически совпадают по метрике ранговой кор­ре­ляции, что вполне естественно и ожидаемо: трехбуквенные со­че­тания характеризуют язык и все подкорпуса русского литера­турного языка.

Полученные результаты подтверждают утверждение: частота использования буквосочетаний длиной от 1 до 3 определяет язык, на котором написан текст. Следовательно, ранговые порядки этих частотных словари близки, что и показывает коэффициент корреляции.

Сравнение словарей сочетаний букв с длиной 4, 5 и 6 букв по­казало уменьшение совпадения частотных словарей с увели­чением длины сочетаний букв. Возможность разделения подкор­пусов по содержанию и эмоциональной ориентации на основе сравнения словарей сочетаний букв не подтверждается в явном виде этими экспериментами. Но полученные нами результаты указывают на возможность использования длинных буквенных комбинаций для обозначения подкорпусов рассказов.

1. . Заключение

Сравнительный анализ подкорпусов текстов проводился с использованием методики расчета коэффициентов попарной ранговой корреляции частотных словарей, полученных для раз­личных лексических характеристик текста (для отдельных частей речи, именных групп, глагольных групп, n-грамм на уровне букв).

На основе полученных данных было показано, что социаль­ные катаклизмы рассматриваемого исторического периода в мень­шей степени повлияли на частоты употребления глаголов, в то время как прилагательные и особенно существительные показы­вают большие изменения. Далее было установлено, что словосо­четания (особенно глагольные группы) в большей степени, чем отдельные части речи, могут быть использованы для сравни­тельного анализа и идентификации кратких литературных текстов для данных исторических периодов. Важным наблюдением яв­ляется то, что наиболее показательными характеристиками, отли­чаю­щими разные подкорпусы текстов, являются частотные после­довательности символов.

**Литература**

1. Мартыненко Г.Я., Шерстинова Т.Ю., Попова Т.И., Мельник А.Г., Замирайлова Е.В.  О принципах создания корпуса русского рассказа первой трети XX века // Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ʻTEL 2018ʼ. Казань. С. 180–197.

2. Мартыненко Г.Я., Шерстинова Т.Ю., Мельник А.Г., Попова Т.И. Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) // Компьютерная лингвистика и вычислительные онтологии. Выпуск 2 (Труды XXI Межд. объед. конф. "Интернет и совре­менное общество, IMS-2018, Санкт-Петербург, 30 мая - 2 июня 2018 г. Сборник научных статей"). — СПб: Университет ИТМО, 2018. С. 99­-104.

3. Martynenko, G.Y., Sherstinova, T.Y. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL-2019, Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552. pp. 105–120.

4. Sherstinova T. Yu, Ushakova E. O., Melnik A. G. Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy: 2020, pp. 221­–229.

5. Лаврентьев А. М., Рябова Д. М., Тихомирова Е. А., Фокина А. И., Чеповский А. М., Шерстинова Т. Ю. Сравнительный анализ специальных корпусов текстов для задач безопасности // Вопросы кибербезопасности. 2020. № 3(37). С. 58-65 DOI: 10.681/2311-3456-2020-03-58-65.

6. Соловьев Ф.Н. Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, №1. С. 74–82. DOI 10.25205/1818-7900-2020-18-1-74-82.

7. Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015.

8. Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Guka­sian M., Egoshina I., and Turygina M. Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930), In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy: 2020. pp. 366­–373. https://fruct.org/publications/acm27/files/She.pdf

**References**

1. Martynenko, G., Sherstinova, T., Melnik, A., Popova, T., Zamirailova, E. O principakh sozdaniya korpusa russkogo rasskaza pervoj treti XX veka // Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ʻTEL 2018ʼ. Kazan. pp. 180–197. (in Russian).

2. Martynenko, G.Ya., Sherstinova, T.Yu., Melnik, A.G., Popova, T.I. (2018) Metodolo-gicheskie problemy sozdaniya Komp’yuternoj antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoj khudozhestvennoj prozy v ehpokhu revolyucionnykh peremen (pervoj treti XX veka) // Computational linguistics and computational ontologies. Issue 2 (Proc. of the XXI Int. United Conf. The Internet and Modern Society, IMS-2018, St. Petersburg, May 30 - June 2, 2018), ITMO University, St. Petersburg. Pp. 99–104. (in Russian).

3. Martynenko, G.Y., Sherstinova, T.Y. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL-2019, St. Pe­ters­burg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552. pp. 105–120.

4. Sherstinova T. Yu, Ushakova E. O., Melnik A. G. Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy: 2020, pp. 221­–229.

5. Lavrentyev A. M., Raybova D.M., Tikhomirova E.A., Fokina A. I., Chepovskiy A. M., Sherstinova T.Yu. Sravnitelniy analis specialnikh korpu­sov tekstov dlay zadach bezopasnosty // Voprosi kiberbezopasnosti. 2020. № 3(37). С. 58-65. DOI: 10.681/2311-3456-2020-03-58-65. (in Russian).

6. Solovev F. N. Embedding Additional Natural Language Processing Tools into the TXM Platform. Vestnik NSU. Series: Information Technologies, 2020, vol. 18, no. 1, p. 74–82. (in Russian)

7. Chepovskiy A. M. Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. Vtoroye izdaniye, pererabotannoye. M.: Natsional'nyy otkrytyy universitet “INTUIT”, 2015. (in Russian)

8. Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., and Turygina M. Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930), In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy: 2020. pp. 366­–373. https://fruct.org/publications/acm27/files/She.pdf

**Correlation Analysis of Frequency Dictionaries
 of Linguistic Indicators of Russian Prose 1900-1930
in Dynamics**

***N.L. Avanesyan, A.M. Chepovskiy, T.Yu. Sherstinova, F.N.  Soloviev, D.Yu. Chuikin***

**Abstract.** The paper describes a comparative analysis of text subcor­puses referring to four periods of Russian literature 1900-1930, which was carried out using the coefficients of pairwise rank correlation of frequency dictionaries of different lexical characteristics. It is shown that social cataclysms influenced the use of verbs to a lesser extent than that of adjectives and nouns. It has been also established that other indicators that distinguish the considered temporal subcorpuses of fiction are noun and verb groups, as well as frequency sequences of symbols.

**Keywords.** Rank correlation, corpus linguistics, noun phrases (NP), verb phrases (VP), literary texts.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Аванесян Нина Левоновна**

Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва.

**Nina L. Avanesyan**

National Research University Higher School of Economics Moscow, Russia.

***E-mail:*** *nlavanesyan@edu.hse.ru*

**Чеповский Андрей Михайлович**

Российский университет Дружбы Народов (РУДН), Москва.

Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва.

**Andrey M. Chepovskiy**

Peoples Friendship University of Russia (RUDN University), Moscow, Russia.

National Research University Higher School of Economics Moscow, Russia.

***E-mail:*** *chepovskiy-am@rudn.ru*

**Шерстинова Татьяна Юрьевна**

Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Санкт-Петербург.

Санкт-Петербургский государственный университет» (СПбГУ), Санкт-Петербург

**Tatiana Yu. Sherstinova**

National Research University Higher School of Economics, St. Petersburg, Russia.

St. Petersburg State University, St. Petersburg, Russia.

***E-mail:*** *tsherstinova@hse.ru*

**Соловьев Фёдор Николаевич**

Федеральный исследовательский центр «Информатика и менеджмент», Москва.

**Fedor N. Soloviev**

Federal Research Center “Informatics and Management”.

***E-mail:*** *the0@yandex.ru*

**Чуйкин Даниил Юрьевич**

Российский технологический университет (РТУ МИРЭА), Москва.

**Daniil Yu. Chuikin**

Russian Technological University (RTU MIREA). Moscow.

***E-mail:*** *chuykin.d.y@edu.mirea.ru*

1. Создание корпуса русских рассказов поддержано грантом РФФИ № 17-29-09173 офи\_м, описанная статистика получена в рамках гранта РФФИ № 19-07-00806. [↑](#footnote-ref-1)