The Dynamics of Vocabulary in Russian Prose (Based on Frequency Dictionaries of the Corpus of Russian Short Stories 1900-1930)

Tatiana G. Skrebtsova
Saint Petersburg State University
Saint Petersburg, Russia
t.skrebtsova@spbu.ru

Alexander O. Grebennikov
Saint Petersburg State University
Saint Petersburg, Russia
a.grebennikov@spbu.ru

Tatiana Yu. Sherstinova
National Research University Higher School of Economics
Saint Petersburg, Russia
tsherstinova@hse.ru

Abstract

The paper presents the results of a study that is part of a large-scale project aimed at studying the changes that took place in the Russian language during the first three decades of the 20th century. In the history of Russia, this period was marked by stormy events that led to a radical change in the state system and the formation of a new society. To quantify the scale of changes that occurred in the language in the result of these dramatic events, it is necessary to analyze the representative volume of linguistic data and to compare different chronological periods in dynamics using quantitative methods. The research was carried out on the data of an annotated sample from the Corpus of the Russian Short Stories of 1900-1930, which contains texts by 300 Russian writers. All the texts in the Corpus are divided into three time frames: 1) the pre-war period (1900-1913), 2) the war and revolutionary years (1914-1922) and 3) the early Soviet period (1923-1930). Frequency distribution of significant vocabulary in dynamics was analyzed, which made it possible to identify the main tendencies in the change of individual words and lexical groups frequencies from one historical period to another and to correlate them with the previously identified dynamics of literary themes. The technique used allows to trace the influence of large-scale political changes on the vocabulary of literary language, to note the peculiarities and tendencies of the writers' worldview in a certain historical period, and also makes it possible to significantly supplement the analysis of the dynamics of literary themes in fiction.

Keywords: lexical studies; lexical changes in diachrony; language and style of literary texts; Russian short story; frequency dictionary; corpus linguistics; computational linguistics

DOI: 10.28995/2075-7182-2021-20-646-659

Динамика лексического состава русской художественной прозы (на материале частотных словарей корпуса русских рассказов 1900-1930)

Т. Г. Скребцова
Санкт-Петербургский государственный университет;
Санкт-Петербург, Россия
t.skrebtsova@spbu.ru

А. О. Гребенников
Санкт-Петербургский государственный университет;
Санкт-Петербург, Россия
a.grebennikov@spbu.ru

Т. Ю. Шерстинова
Научно-исследовательский университет «Высшая школа экономики»
Санкт-Петербург, Россия
tsherstinova@hse.ru

Аннотация

В работе представлены результаты масштабного проекта, направленного на изучение изменений, произошедших в русском языке в течение первых трех десятилетий XX века. В истории нашей страны этот период был отмечен бурными событиями, которые привели к радикальному изменению государственного строя и построению нового общества. Для количественной оценки масштаба изменений, которые произошли в языке в результате драматических событий рассматриваемого периода, необходим анализ представительного объема языкового материала и сравнение разных хронологических срезов в динамическом аспекте с применением количественных методов. Исследование проведено на материале аннотированной выборки из Корпуса русского рассказа 1900-1930 гг., в которой представлены тексты 300 русских писателей. Материалы Корпуса делятся на три временных среза: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). Проанализировано частотное распределение знаменательной лексики в динамике, что позволило выявить яркие тенденции в изменении частотности отдельных слов и лексических групп от одного исторического отрезка к другому и соотнести их с ранее выявленной динамикой тем. Используемая методика позволяет проследить влияние крупномасштабных политических изменений на словарный состав языка художественной литературы, отметить особенности и тенденции мировосприятия авторов в определённый исторический период, а также дает возможность существенно дополнить анализ динамики тем художественных произведений.

Ключевые слова: лексика; изменение лексического состава в диахронии; язык и стиль художественных текстов; русский рассказ; частотный словарь; корпусная лингвистика; компьютерная лингвистика

1 О Корпусе русских рассказов (1900-1930) и его периодизации

Настоящее исследование является частью масштабного проекта, направленного на изучение изменений, которые происходили в русском языке в первую треть XX века, – возможно, самый драматический период его развития. Проект включает создание электронного текстового корпуса, содержащего тысячи русских рассказов, написанных в первые три десятилетия прошлого века, и их дальнейший комплексный филологический анализ [9; 10; 12].

В истории нашей страны этот период был отмечен бурными событиями, приведшими к радикальному изменению государственного строя и построению нового общества. Цепь исторических событий, охватывающих Первую мировую войну, Февральскую и Октябрьскую революцию и Гражданскую войну, обусловила масштабные языковые и стилистические сдвиги. Огромный пласт «отжившей» лексики сменился новыми словами, отражающими новые понятия и идеи, многие слова «из прошлой эпохи» приобрели новые значения или коннотации, произошла трансформация общепринятых речевых структур (в частности, поменялись функциональные частоты многих лексических единиц, сменился набор привычных коллокаций, появились новые модели сочетаемости, фразеологические обороты и т. д.). Помимо «естественного» процесса резких языковых изменений, неизбежно сопровождающих любой переломный период, следует отметить и сознательные действия новой власти, направленные на изменение языковых норм, с тем чтобы еще более размежеваться с уходящей эпохой и подчинить языковую политику государства решению новых актуальных задач.

Для количественной оценки масштаба изменений, которые произошли в языке в результате драматических событий первой трети XX века, необходим анализ представительного объема языкового материала и сравнение разных хронологических срезов в динамическом аспекте с применением количественных методов. Для этой цели создается Корпус русских рассказов первых трех десятилетий XX века, насчитывающий несколько тысяч единиц. Выбор жанра рассказа для изучения языковых и стилистических изменений обусловлен тем, что он принадлежит к числу наиболее распространенных жанров художественной литературы. Это позволяет охватить тексты максимального числа авторов, писавших в исследуемую эпоху, – не только ведущих, но и множества второстепенных, – что способствует репрезентативности коллекции и достоверности выводов. Другая причина выбора именно этого литературного жанра связана со способностью рассказов, в силу своего небольшого объема и предназначенности (как правило) для пуб-

ликации в периодических изданиях, чутко реагировать на текущие события и улавливать изменения в общественном сознании.

Историческим центром рассматриваемой эпохи, ее переломом, является Октябрьская революция. Все остальные события и процессы рассматриваются или как преддверие центрального события, или как его последствия. Материалы Корпуса делятся на три временных среза: 1) довоенный период: начало XX века до Первой мировой войны (1900–1913), 2) военно-революционные годы: Первая мировая война, Февральская и Октябрьская революция и Гражданская война (1914–1922) и 3) советский период (1923–1930).

Писатель может быть представлен одним рассказом в каждый временной отрезок, причем в Корпус не включаются рассказы, написанные в эмиграции. Так, покинувший Россию в 1920 г. И. А. Бунин представлен одним рассказом за довоенный период и еще одним – за военно-революционный период.

Аннотированная выборка из Корпуса, в которой представлены тексты 300 русских писателей (приблизительно по 100 за каждый период[1], более 1 млн. словоупотреблений) служит своеобразным полигоном для разностороннего изучения материала. Наряду с творчеством признанных мастеров пера – Чеховым, Буниным, Горьким, Куприным, Вересаевым, Булгаковым, Шмелевым, Грином, Тэффи, Замятиным, Зощенко, Катаевым, Пильняком, Кавериным, Гайдаром, Олешей, Бабелем, Платоновым, Пришвиным и др. – она включает произведения авторов, известных лишь узким специалистам. На основе этой выборки проводятся исследования, затрагивающие не только языковые изменения, но и динамику тем, а также композиционную специфику русских рассказов [7; 8; 12–17].

Настоящее исследование также базируется на данном подкорпусе. Оно посвящено анализу наиболее частотных знаменательных лексем, собранных как отдельно по каждому из трех периодов, так и совокупно по всему историческому отрезку длиной в три десятилетия. Следует подчеркнуть, что выделенные периоды принципиально отличны друг от друга: можно сказать, что второй (военный) противопоставлен первому и третьему (мирным), однако первый и третий кардинально различаются между собой по общественно-политическому устройству. Поэтому их попарное сравнение представляет несомненный интерес, в особенности на фоне выявленной ранее динамики тем [14; 15; 17].

В предыдущих исследованиях, выполненных на материале Корпуса русского рассказа, уже было отмечено, что масштабное сопоставление данных, позволяющее выявить лексическое своеобразие эпохи, обусловленное общественно-политической атмосферой и отражающее тенденции в языковом употреблении, возможно только путем сравнения верхних рангов знаменательной лексики с соответствующими рангами писательских словарей и словаря языка в целом, во-первых, и, что гораздо более желательно, с аналогичным электронным корпусом отечественных литературных произведений, относящихся к какому-либо другому историческому периоду, во-вторых.

В частности, в [11] были наглядно представлены результаты такого сопоставления для первого (довоенного) периода рассказов из нашей выборки. Одновременно установлено, что сравнение со словарями отдельных авторов обнаруживает значительную индивидуально-авторскую вариативность выделенных лексем, затрудняющую выделение общих тенденций [1–5]. Сравнение же со словарем языка в целом неизбежно сталкивается с проблемой представленности в нем множества жанров, и, хотя результаты зачастую интересны и показательны, они искажены значительной долей статистического «шума» [6; 7]. Поэтому в данном исследовании было принято решение от него отказаться.

Напротив, было установлено, что сопоставление рассказов 1900–1913 гг. с русскими рассказами начала XXI века, полученными на основе Национального корпуса русского языка (НКРЯ), обладает значительным стилеразличительным потенциалом [6]. Оно дает возможность подтвердить или опровергнуть справедливость сделанных предположений о том, что именно принадлежность к разным историческим периодам прежде всего обусловливает наблюдаемые различия в частотном распределении лексем.

[1] Было выбрано 300 авторов, но из-за того, что некоторые из них представлены более, чем в одном периоде, выборка составляет не 300, а 310 рассказов.

2 Методология исследования

На материале аннотированной выборки из Корпуса русских рассказов были построены расположенные в порядке убывания частот частотные словари как для выборки в целом, так и для каждого из исторических периодов объемом 24 316 лексем, 376 513 словоформ для первого периода; 24 617 лексем, 303 588 словоформ для второго периода; 30 560 лексем; 383 430 словоформ для третьего периода и 124 081 лексема, 1 077 970 словоформ для выборки в целом.

В качестве объекта исследования были выбраны знаменательные слова, расположенные в верхних зонах частотного распределения (с частотой выше 100). Для каждого периода их количество превысило 200, и составило около 800 для выборки в целом.

При сравнении частотного распределения лексики в разные периоды (см. раздел 3) используется такой параметр, как ранг лексемы (иначе говоря, ее статистический вес). Вследствие некоторой разницы в объемах полученных словарей учет именно рангов, а не абсолютных частот является корректным решением.

В тексте статьи для описания динамики частотности отдельных слов используется следующая нотация. В скобках при слове указывается его текущий ранг; если в предшествующий период оно также входило в рассматриваемую верхнюю зону рангового распределения, указывается и его прежний ранг, и два числа соединяются стрелкой. Например, при сравнении второго периода с первым запись *солдат* (116→62) означает, что данное слово имело 116-й ранг в первый период и переместилось на 62-ю позицию во второй период. Если слово в предшествующий период находилось вне верхней зоны, а в рассматриваемый период в нее вошло, первое число отсутствует, ср. *офицер* (→169). Если же, напротив, в более ранний период слово присутствовало в верхней зоне, а в рассматриваемый период ее покинуло, отсутствует второе число, ср. *красивый* (143→). При сопоставлении данных за три периода возможна ситуация, когда некоторое слово присутствует в верхней зоне в первый и третий периоды, но отсутствует во второй – в таком случае запись выглядит следующим образом: *смех* (257→→234).

Несколько слов следует сказать о специфике работы автоматической программы лемматизации[2]. В частности, парные глаголы совершенного и несовершенного вида рассматриваются в качестве отдельных лексем (что представляется методологически правильным в силу частых расхождений в наборе их значений). То же относится к супплетивным формам (так, ниже отдельно фигурируют формы *ребенок* и *дети*). Имена собственные по понятным причинам были исключены из частотных списков. Некоторые из них из-за графического совпадения с именами нарицательными могут влиять на статистику последних, ср. *Вера – вера*. В подобных случаях соответствующие общие имена также не принимались во внимание.

Заметим, что при автоматической обработке неизбежны некоторые погрешности, связанные с неправильным определением леммы (*большой* и *больший*, *стоять* и *стоить*, *пол* и *пола*, *лес* и *леса* и пр.), омонимией (*мир*, *язык*) и грамматической неоднозначностью (*стать*). Во избежание ошибочных заключений подобные случаи также исключены из рассмотрения.

При интерпретации полученных данных полисемичные слова рассматриваются в совокупности всех своих лексико-семантических вариантов, что, разумеется, ведет к некоторым погрешностям, но является неизбежным следствием применения автоматических методов. Представляется, что эта погрешность невелика при рассмотрении небольших и компактно расположенных исторических отрезков.

3 Динамика частотных лексем по периодам

Примечательно, что шесть самых верхних рангов во всех периодах занимают одни и те же слова (различаясь лишь позициями): *говорить*, *сказать*, *один*, *глаз*, *рука*, *мочь*. Остальной материал демонстрирует как сходства, так и существенные различия. Посредством сравнения более позднего периода с более ранним(и) мы стремимся выделить: 1) новые слова, вошедшие в верхнюю зону частотного распределения; 2) слова, которые ушли из нее и 3) слова, демонстрирую-

[2] Словари строились при помощи программы "UNILEX" (разработка Институт русского языка им. Виноградова). См. Аношкина Ж.Г. Тексто-ориентированная компонента АЛС УНИЛЕКС (УНИЛЕКС-Т) // Альманах «Говор», № 5, 1995, стр. 7-29.

щие выраженную динамику частоты. Мы анализируем эти явления, пытаясь связать их с общественно-политической обстановкой соответствующего времени и выявленной ранее динамикой тем [7–9].

Сравнение второго (военно-революционного) периода с первым (довоенным)

На фоне довоенных рассказов в произведениях второго периода закономерно увеличивается доля военной тематики [17, с. 50] и становится актуальной национальная самоидентификация, что соответственно проявляется во вхождении в верхнюю зону частотного распределения слов *офицер* (→169) и *русский* (→182). Военным временем, по-видимому, объясняется и вхождение слов *дьякон* (→83) и *писать* (→181) – в связи с возросшей ролью церкви, вынужденным разделением семей, тревогой за близких и потребностью в переписке. Указанные корреляции подкрепляются динамикой роста у слов *бог* (89→65), *солдат* (116→62) и *письмо* (190→117), которые были в верхней зоне и в первый период, но переместились на более высокие позиции рангового распределения.

Еще одним индикатором смены эпохи можно считать уход слова *можно* (80→), с одной стороны, и появление в верхней зоне слов *должный* (→76) и *нельзя* (→147) – с другой, что сигнализирует о наступлении более сурового и жесткого времени, связанного с ограничениями, запретами и принуждениями (глагол *мочь*, впрочем, сохраняет свой шестой ранг). В целом, можно сказать, что модальность долженствования вытесняет модальность возможности. В такой период важной оказывается концентрация на текущем моменте – отсюда вхождение слов *сегодня* (→172), *сейчас* (→71), *теперь* (→20), *наконец* (→186).

Вполне закономерным выглядит уход из верхней зоны подавляющего большинства слов, обозначающих положительные эмоции, а именно: *праздник* (268→), *добрый* (265→), *светлый* (236→), *красивый* (143→), *веселый* (152→), *весело* (182→), *смех* (257→), *счастье* (228→), *чувство* (126→), *улыбка* (219→), *улыбаться* (173→), *тихий* (128→), *тишина* (249→). (Некоторые из них, а именно *веселый*, *смех*, *тихий* и *тишина*, возвращаются в третий период.) Этот факт коррелирует с уменьшением значимости широкого круга тем, связанных с любовью, семьей, помощью ближнему, благотворительностью [7, с. 54, 56]. В связи с этим обращает на себя внимание уход слов *ребенок* (213→), *играть* (227→) и снижение частоты слова *дети* (69→107), которое все-таки удерживается в верхней зоне.

Из прочих слов, покинувших верхнюю зону, отметим *пить* (267→) и *пьяный* (241→) – этот факт можно объяснить введением сухого закона и соответственным снижением темы пьянства, которая, очевидно, повышает частоту данных слов [7, с. 53]. Уход слов *барин* (258→), *хозяин* (226→), *работать* (218→), *рабочий* (202→) связан, по-видимому, с актуализацией темы войны, вытесняющей мирный труд. По аналогичной причине из верхних рангов пропало слово *студент* (217→).

Заметим, что большинство «пропавших» слов так и не вернулось в верхние ранги частоты в третий, советский, период.

Справедливости ради следует упомянуть и то, что можно назвать контрпримерами, а именно слова, сохранение или исчезновение которых не получается объяснить выявленной ранее динамикой тем. К примеру, в верхней зоне частотного распределения осталось слово *смеяться*, хотя ранг его и понизился (131→183). Напротив, ушли слова *страшный* (135→), *страх* (253→), *ужас* (166→), *дрожать* (185→), *умереть* (191→), *тоска* (207→), *больной* (210→), которые, казалось бы, гораздо более востребованы в эпоху войн и революций, чем в мирное время. Несмотря на заметный и вполне объяснимый рост «мистических» тем, связанных с видениями, предчувствиями, снами, мечтами [7, с. 56], верхнюю зону покинули слова *показаться* (194→), *похожий* (174→), *странный* (155→).

Сравнение третьего (советского) периода с предшествующими

Характерной особенностью третьего периода является вхождение в верхнюю зону частотного распределения большого числа конкретных существительных, связанных с сельской жизнью: *дед* (→114), *старуха* (→166), *ребята* (→183), *хлеб* (→152), *поле* (→238), *куст* (→255), *трава*

(→245), *собака* (→165), *конь* (→204), *птица* (→230) – и техническим прогрессом: *машина* (→250), *поезд* (→256), *вагон* (→173), *ход* (→177). Это коррелирует с выраженным ростом соответствующих тем [17, с. 51-52]. Число абстрактных существительных, напротив, сокращается.

Список частей тела человека, и так широко представленных в верхней зоне частотного распределения, в советский период увеличивается чуть ли не вдвое. Так, к уже имеющимся единицам *рука*, *глаз*, *голова*, *лицо*, *губа*, *зуб*, *нога*, *тело*, *плечо*, *палец*, *волос* добавились слова *нос* (→121), *ухо* (→184), *лоб* (→216), *шея* (→275), *щека* (→264), *борода* (→268), *бок* (→252), *колено* (→205).

Расширился также набор числительных – к *один*, *два*, *три*, *первый* добавились лексемы *четыре* (→262), *пять* (→218), *второй* (→258). На фоне общего уменьшения числа прилагательных возвращается слово *синий* и появляются такие цветообозначения, как *желтый* (→221) и *зеленый* (→222). (*Белый* и *черный* стабильно занимают высокие позиции, а по поводу *красный* см. ниже).

Появилось слово *вперед* (→170), что, по-видимому, объясняется актуализированными темами технического прогресса и светлого будущего [17, с. 51]. Политика ликвидации безграмотности и культпросвета обусловила вхождение в верхнюю зону слова *книга* (→208).

Вновь появилось пропавшее во второй период слово *можно* (80→→95). *Нельзя* (→147→) ушло, но *должный* (→76→67) сохранилось. Слово *хотеться* (101→150→227) демонстрирует последовательное снижение ранга, при том что *хотеть* (19→23→23) на протяжении всех трех периодов занимает примерно одинаковое высокое положение в частотном распределении.

Показателем наступившего мирного времени можно считать возвращение в верхнюю зону частотного распределения слов *работать* (218→→109), *рабочий* (202→→112), *веселый* (152→→232), *смех* (257→→234), *тихий* (128→→196), *тишина* (249→→231), *играть* (227→→207), *разговор* (186→→209).

Обращает на себя внимание вхождение в верхнюю зону слов *ружье* (→278), *рота* (→218) и *кровь* (→105). На первый взгляд, более естественным казалось бы их появление в предыдущий, военно-революционный, период. Возможное объяснение заключается в том, что в нашей выборке за третий период рассказов про Гражданскую войну оказалось примерно в два раза больше, чем за второй, т. е. тогда, когда эта война шла. Это «отставание» литературы от жизни обусловлено рядом факторов. Из наиболее очевидных упомянем тот банальный факт, что на написание рассказа требуется время, затем проходит еще какое-то время до его публикации, которая сама по себе затруднена в условиях затянувшейся войны, политических волнений и экономической разрухи. К тому же, для осознания столь масштабных событий, приведших к радикальному изменению общественной жизни, необходима дистанция («большое видится на расстоянии»). Отсюда своеобразный отсроченный эффект: чем крупнее историческое событие, тем дольше оно сохраняет свою значимость, в том числе в литературе и искусстве.

Вполне закономерно верхнюю зону частотного распределения покинули слова *офицер* (→169→) (в Красной Армии воинские звания были упразднены), *солдат* (116→62→) и *дьякон* (→83→). Слово *бог* (89→65→211) осталось, но его ранг заметно снизился, что обусловлено антирелигиозной политикой советской власти.

Примечателен уход из верхней зоны слов *гость* (165→) и *знакомый* (193→), присутствовавших там и в первый, и во второй периоды. Социальные отношения теперь сводятся в основном к семейным и трудовым, причем семейные связи угасают: у слов *муж* (146→166→236), *жена* (52→84→139), *дети* (69→107→198) наблюдается снижение ранга, а слово *ребенок* так и не вернулось в верхнюю зону.

В числе прочих слов, присутствовавших в верхней зоне в оба предшествующих периода, но ушедших в советское время, отметим *господин* (145→146→) (дореволюционная тематика практически сошла на нет), *толпа* (79→103→), *милый* (122→176→), *любовь* (134→112→), *письмо* (190→117→).

Общая динамика частот на протяжении трех периодов

В этом разделе мы сосредоточимся на словах, присутствующих в верхней зоне частотного распределения на протяжении всех трех периодов. Прежде всего нас интересуют случаи ярко выраженного последовательного изменения ранга. Они разделяются на 1) случаи поступательного роста ранга и 2) случаи поступательного снижения ранга. Эти примеры, как мы предполагаем, обусловлены внешним контекстом – коренным переворотом политического строя, сломом прежних социальных отношений и построением нового общества.

Возможно, наиболее ярким примером первого типа может служить заметная активизация слова *товарищ* (198→105→38) как индикатора новой советской власти. Здесь же уместно упомянуть схожую динамику слова *красный* (163→110→54). Помимо этого, наблюдается существенное повышение ранга у слов, которые можно прямо или опосредованно связать с жизнью на селе, ср. *деревня* (250→192→150), *народ* (224→198→164), *мужик* (252→144→61), *баба* (→143→81), *сын* (183→161→126), *брат* (255→136→120), *утро* (100→81→55), *лошадь* (176→139→93), *бежать* (169→134→85), *дорога* (209→168→134), *ветер* (216→149→113).

Примеры второго типа включают слова, обозначающие внутреннюю жизнь человека, ср. *любить* (39→43→69), *чувствовать* (54→152→197), *смеяться* (131→183→189), *душа* (42→35→151), *мысль* (62→102→130), а также его связи с близкими, ср. *жена* (52→84→139), *муж* (146→166→236), *дети* (69→107→198). Эта же динамика характерна для слов *молодой* (58→59→147), *казаться* (14→51→58) и *смерть* (106→163→223).

Последовательное уменьшение ранга зафиксировано у слова *деньги* (133→178→249), что обусловлено снижением покупательной способности денег в военно-революционный период, эмиссией разнообразных бумажных знаков, имевших сомнительную ценность, а затем денежными реформами советской власти (деноминациями). Заметим, что соответственно снижается частота темы, связанной с оппозицией богатства и бедности: после Гражданской войны в Советской России просто не осталось богатых людей [17, с. 53].

Кроме случаев поступательного изменения ранга (будь то рост или падение) имеются слова, у которых динамика употребления может быть представлена в виде ломаной линии. Иными словами, они имеют точку перелома во втором периоде, а показатели первого и третьего периодов достаточно схожи. Однако их перечень, как нам кажется, не дает основания для каких-либо обобщений и корреляций с исторической ситуацией. В связи с этим мы опускаем данные о динамике рангов, ограничиваясь простым перечислением. Так, возрастание ранга во второй период фиксируется у слов *живой*, *читать*, *легкий*, *просить*, *высокий*, *поднять*, *бояться*, *ждать*, *дать*. Напротив, падение ранга во второй период наблюдается у слов *девушка*, *длинный*, *тяжелый*, *угол*, *воздух*, *подумать*, *свет*.

Достаточно большое число слов вообще не имеют значительных колебаний ранга, образуя своеобразные «инварианты». Они расположены преимущественно в пределах верхних 80 рангов – ниже разброс, как правило, довольно велик. (Хотя и в этих рамках иногда случаются резкие колебания: достаточно указать на слова *красный* и *товарищ*, см. выше.)

К словам, стабильно характеризующимся высокой частотой, относятся основные глаголы движения (*идти*, *ходить*, *выйти*, *пойти*, *уйти*), позы (*сидеть*, *стоять*, *лежать*), речи (*говорить*, *сказать*, *спросить*, *молчать*), чувственного восприятия (*видеть*, *смотреть*), а также глаголы *жить*, *взять*, *спать*, *хотеть*, *любить*, *казаться*. Из высокочастотных прилагательных упомянем *последний*, *большой*, *маленький*, *старый*, *новый*, *белый*, *черный* и примыкающие к ним местоимения *другой* и *каждый*. Существительные верхних рангов включают такие группы слов, как *город – улица – дом*, *комната – стол*, *человек – люди*, *время – год – час*, названия частей тела (*голова*, *лицо*, *глаз*, *нога*, *рука*), времен суток (*день*, *ночь*, *утро*, *вечер*), а также лексемы *жизнь*, *место*, *сторона*, *земля*, *дело*, *отец*, *сердце*.

4 Заключение

В настоящей статье проанализировано частотное распределение знаменательной лексики на материале выборки из Корпуса русских рассказов (1900-1930). Внимательное рассмотрение верхней зоны частотного распределения позволило выявить яркие тенденции в изменении ча-

стотности отдельных слов и лексических групп от одного исторического отрезка к другому и соотнести их с ранее выявленной динамикой тем.

Среди возможных направлений дальнейшего анализа наиболее интересным и перспективным представляется сопоставление полученных данных с частотностью лексических единиц в русских рассказах начала XXI века. Оно позволяет в более широком ракурсе оценить языковые и стилистические изменения в языке литературных произведений и задуматься об их причинах. Мы рассматриваем это в качестве самостоятельного направления дальнейших исследований.

В целом, анализ наших данных показывает, что частотные распределения, построенные на материале представительных выборок из масштабного корпуса текстов, могут служить хорошим индикатором динамики лексического состава художественной прозы произведений отдельной эпохи. Используемая методика позволяет проследить влияние крупномасштабных политических изменений на словарный став языка художественной литературы, отметить особенности и тенденции мировосприятия авторов в определённый исторический период, а также позволяет существенно дополнить анализ динамики тем произведений.

Acknowledgements

Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

References

[1] Grebennikov A.O. (2019), Measures of lexical similarity of frequency dictionaries [Mery leksicheskogo skhodstva chastotnykh slovarey], Structural and Applied Linguistics [Strukturnaya i Prikladnaya Lingvistika], No. 12, Saint Petersburg, pp. 61–68.

[2] Grebennikov A.O., Martynenko G.Ya. (Ed.) (1999), Frequency Dictionary of the Short Stories by Anton P. Chekhov [Chastotnyy slovar rasskazov A.P. Chekhova], Izdatelstvo Sankt-Peterburgskogo Universiteta, Saint Petersburg.

[3] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2003), Frequency Dictionary of the Short Stories by Leonid N. Andreev [Chastotnyy slovar rasskazov L.N. Andreeva], Izdatelstvo Sankt-Peterburgskogo Universiteta, Saint Petersburg.

[4] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2006), Frequency Dictionary of the Short Stories by Alexander I. Kuprin [Chastotnyy slovar rasskazov A.I. Kuprina], Izdatelstvo Sankt-Peterburgskogo universiteta, Saint Petersburg.

[5] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2011), Frequency Dictionary of the Short Stories by Ivan A. Bunin [Chastotnyy slovar rasskazov A.I. Bunina]. Izdatelstvo Sankt-Peterburgskogo universiteta, Saint Petersburg.

[6] Grebennikov A.O., Marusenko N.M. (2020), Corpus of the Russian story of the early XX century. An example of linguistic statistical analysis [Korpus russkogo rasskaza nachala XX veka. Primer lingvostatisticheskogo analiza], Computational Linguistics and Computational Ontologies: Proceedings of the XXIII Joint Conference "Internet and Modern Society" (IMS–2020) [Komp'yuternaya Lingvistika i Vychislitel'nye Ontologii. Trudy XXIII Mezhdunarodnoj Ob"edinennoj Konferencii «Internet i Sovremennoe Obshchestvo» (IMS–2020)], Saint Petersburg, pp. 21–29.

[7] Grebennikov A.O., Skrebtsova T.G. (2019), Yazykovaya kartina mira v russkom rasskaze nachala XX veka [Linguistic picture of the world in the Russian story of the early XX century], Philosophy and Humanities in the Information Society [Filosofija i gumanitarnyie nauki v informatsionnom obschestve], No. 3, pp. 82–92.

[8] Martynenko G., Sherstinova T. (2018), Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture, Digital Transformation and Global Society. Communications in Computer and Information Science, Vol. 859, Saint Petersburg, pp. 299–309.

[9] Martynenko G.Ya., Sherstinova T.Yu, Popova T.I., Melnik A.G. (2018), Methodological Issues Related with the Compilationof DigitalAnthology of Russian Short Stories (the First Third of the 20th Century) [Metodologicheskie 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: Proceedings of the

XXI Joint Conference "Internet and Modern Society" (IMS–2018) [Komp'yuternaya Lingvistika i Vychislitel'nye Ontologii. Trudy XXI Mezhdunarodnoj Ob"edinennoj Konferencii «Internet i Sovremennoe Obshchestvo» (IMS–2018)], Saint Petersburg, pp. 99–104. Access mode: https://openbooks.itmo.ru/ru/file/8421/8421.pdf

[10] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamirajlova E.V. (2018), On the principles of creation of the Russian short stories corpus of the first third of the 20th century [O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka], Proceedings of the XV International Conference on Computer and Cognitive Linguistics "TEL 2018", Kazan, pp. 180–197.

[11] Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina M. (2020), Frequency word lists and their variability (the case of Russian fiction in 1900-1930), Proceedings of the 27th Conference of FRUCT Association, Helsinki, № 27, pp. 366–373.

[12] Sherstinova T., Martynenko G. (2020) 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. R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia. CEUR Workshop Proceedings, Vol. 2552, pp. 105–120. Access mode: http://ceur-ws.org/Vol-2552/.

[13] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. (2020), Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction, Advances in Computational Intelligence. MICAI 2020. Lecture Notes in Computer Science, vol. 12469, Springer, Cham, pp. 134–151. Access mode: https://doi.org/10.1007/978-3-030-60887-3_13.

[14] Sherstinova T., Skrebtsova T. (2020), Russian literature around the October revolution: A quantitative exploratory study of literary themes and narrative structure in Russian short stories of 1900-1930, Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, pp. 117–128. Access mode: http://ceur-ws.org/Vol-2813/rpaper09.pdf

[15] Skrebtsova T. (2020) Thematic tagging of literary fiction: the case of early 20th century Russian short stories, Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, pp. 265–276. Access mode: http://ceur-ws.org/Vol-2813/rpaper20.pdf

[16] Skrebtsova T.G. (2019), The structure of the narrative in the Russian story of the early XX century [Struktura narrativa v russkom rasskaze nachala XX veka], Proceedings of the International Conference "Corpus linguistics–2019" [Trudy Mezhdunarodnoy Konferentsii "Korpusnaya Lingvistika–2019"], St. Petersburg, pp. 426–431.

[17] Skrebtsova T.G. (2020), Dynamics of themes of Russian stories of the early XX century [Dinamika tem russkikh rasskazov nachala XX veka], Philosophy and Humanities in the Information Society [Filosofija i gumanitarnyie nauki v informatsionnom obschestve], No. 3, pp. 45–60.

Приложение. Ранги знаменательной лексики по частоте употребления в каждом из трех исторических периодов, верхняя зона частот (250 слов).

СЛОВО	1-й ПЕРИОД (1900-1913)	2-й ПЕРИОД (1914-1922)	3-й ПЕРИОД (1923-1930)
	РАНГ	РАНГ	РАНГ
сказать	1	2	3
один	2	3	4
глаз	3	4	2
говорить	4	1	5
рука	5	5	1
мочь	6	6	6
лицо	7	9	12
знать	8	7	8
другой	9	12	13
голова	10	10	7
идти	11	8	9
жизнь	12	19	33
человек	13	14	11
казаться	14	51	58
думать	15	22	20
люди	16	15	29
время	17	37	24
голос	18	26	30
хотеть	19	23	23
видеть	20	30	22
дом	21	32	37
большой	22	16	17
смотреть	23	38	31
два	24	17	14
ночь	25	21	26
раз	26	34	28
дело	27	27	18
сидеть	28	29	21
слово	29	28	40
пойти	30	31	19
нога	31	25	15
земля	32	40	27
друг	33	42	88
комната	34	56	56
окно	35	45	42
дверь	36	33	25
есть	37	36	46
белый	38	39	32
любить	39	43	69
стоять	40	50	44
черный	41	46	36
душа	42	35	151
первый	43	60	43
темный	44	118	101
спать	45	58	53
спросить	46	49	57
сердце	47	54	77
час	48	70	71

место	49	44	34
молчать	50	85	59
маленький	51	48	65
жена	52	84	139
новый	53	61	66
чувствовать	54	152	197
жить	55	53	64
минута	56	101	146
сторона	57	78	35
молодой	58	59	147
старый	59	77	60
уйти	60	94	62
выйти	61	57	45
мысль	62	102	130
стена	63	126	80
сделать	64	108	117
стол	65	64	51
взять	66	52	39
делать	67	100	87
отец	68	55	75
дети	69	107	198
леса	70	47	73
каждый	71	90	79
город	72	79	47
ходить	73	73	76
улица	74	74	70
свет	75	148	115
лежать	76	72	72
нужный	77	68	49
вечер	78	69	48
толпа	79	103	
можно	80		95
последний	81	86	83
тяжелый	82	157	143
вода	83	123	52
небо	84	92	116
слышать	85	116	129
стоить	86	104	92
женщина	87	67	91
понять	88	125	100
бог	89	65	211
понимать	90	135	127
почтить	91	142	156
сила	92	128	90
лета	93	106	145
ответить	94	122	111
самый	95	120	86
три	96	98	68
мать	97	113	102
год	98	80	74
воздух	99	164	137
утро	100	81	55
хотеться	101	150	227
глядеть	102	82	123
губа	103	115	82

тело	104	133	84
больший	105	99	63
смерть	106	163	223
конец	107	96	89
кричать	108	129	94
начало	109	145	144
ряд	110	119	106
слушать	111	93	118
длинный	112	188	158
угол	113	167	110
вера	114		
грудь	115	124	103
солдат	116	62	
ждать	117	87	142
подумать	118	170	122
далекий	119	121	125
остаться	120	131	108
взгляд	121	153	213
милый	122	176	
увидеть	123	91	107
солнце	124	97	78
остановиться	125	155	133
чувство	126		
волос	127	173	155
тихий	128		196
дать	129	63	98
плечо	130	109	96
смеяться	131	183	189
шаг	132	158	179
деньга	133	178	249
любовь	134	112	
страшный	135		241
прийти	136	89	157
дорогой	137	132	97
подойти	138	130	128
холодный	139		246
слеза	140	162	182
бояться	141	95	149
девушка	142	191	171
красивый	143		
целый	144	154	200
господин	145	146	
муж	146	166	236
отвечать	147		
хороший	148	127	172
работа	149	114	50
высокий	150	138	186
пройти	151	199	176
веселый	152		232
становиться	153		
посмотреть	154	177	160
странный	155		
войти	156		185
встать	157	187	178
давать	158	174	124

звук	159		
сильный	160		
сон	161	184	175
иногда	162		
красный	163	110	54
огромный	164		201
село	165	171	104
ужас	166		
крикнуть	167		180
широкий	168		168
бежать	169	134	85
вид	170		254
продолжать	171		
плакать	172	156	237
улыбаться	173		
похожий	174		
уходить	175	151	154
лошадь	176	139	93
огонь	177		135
заметить	178		190
знакомый	179	193	
иметь	180		167
полный	181		169
весело	182		
сын	183	161	126
выходить	184	185	243
дрожать	185		276
разговор	186		209
найти	187	137	132
спрашивать	188	189	265
тонкий	189		229
письмо	190	117	
умереть	191		253
палец	192	159	99
серый	193	194	119
показаться	194		
снег	195		140
бледный	196		
поднять	197		181
товарищ	198	105	38
черта	199		192
забыть	200		
мир	201		174
рабочий	202		112
синий	203		136
стараться	204		
двор	205	180	153
начинать	206		273
тоска	207		
помнить	208		
дорога	209	168	134
больной	210		
сталь	211		161
живой	212	175	199
ребенок	213		

сень	214	140	
бросить	215		141
ветер	216	149	113
студент	217		
работать	218		109
улыбка	219		
батюшка	220		
близкий	221		
радость	222	196	191
движение	223		
народ	224	198	164
дерево	225		225
хозяин	226		274
играть	227		207
счастье	228		
ехать	229		159
почувствовать	230		
чужой	231		272
оставаться	232		
оставить	233		
поднять	234	141	
придти	235		
светлый	236		
гореть	237		
десять	238		
доктор	239		233
река	240		224
пьяный	241		
яркий	242		
вернуться	243		187
рубль	244		
крик	245		
легкий	246	195	226
сестра	247		
тень	248		
тишина	249		231
деревня	250	192	150