Санкт-Петербургский государственный университет

АССЕЛЬ Анастасия Никитична

Выпускная квалификационная работа

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

Уровень образования: бакалавриат

Направление 45.03.02 «Лингвистика»

Основная образовательная программа СВ.5106. «Прикладная, компьютерная и математическая лингвистика (английский язык)»
Профиль «Прикладная, компьютерная и математическая лингвистика (английский язык)»

Научный руководитель:
доцент, Кафедра математической лингвистики,
Гребенников Александр Олегович
Рецензент:
кандидат филологических наук, доцент,
Национальный исследовательский университет "Высшая школа экономики",
Шерстинова Татьяна Юрьевна

Санкт-Петербург
2022

Аннотация

В данной работе описывается изучение изменений, произошедших в русском языке в течение первых трех десятилетий XX века и их сравнение с языком первых десятилетий XXI века. За тридцать лет для России произошло много значительных событий, таких как Первая мировая война, распад Российской империи, Октябрьская революции и т.п. Исследование проведено на материале частотных словарей из Корпуса русского рассказа 1900-1930 гг., в котором представлены тексты 300 русских писателей, а также на материале рассказов Национального корпуса русского языка в период с 2000 по 2022 год. Материалы Корпуса русского рассказа делятся на три временных периода: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). В данной работе было проанализировано частотное распределение знаменательной лексики, а также проведено сравнение полученных результатов на разных временных срезах. Таким образом, были обнаружены тенденции в изменении частотности некоторых слов и семантических групп слов. Сравнение лексики первых тридцати лет XX века и лексики XXI века было проведено благодаря параметру относительной встречаемости – ipm. С помощью данного исследования можно проследить влияние важных исторических вех на лексический состав языка художественной литературы, оценить тенденции, которым подвергся язык, а также выявить ведущие темы начала XX века и их отличия от тем XXI века.

Ключевые слова: лексико-статистический анализ; семантика; художественные тексты; частотный словарь; корпусная лингвистика; компьютерная лингвистика

Abstract

This paper describes a study of the changes that occurred in the Russian language during the first three decades of the 20th century and their comparison with the language of the first decades of the 21st century. Over thirty years many significant events occurred for Russia, such as the First World War, the collapse of the Russian Empire, the October Revolution, etc. The research is conducted on the material of frequency dictionaries from the Russian short stories corpus 1900-1930 which are divided into three different time periods and present the texts of 300 Russian writers, as well as on the material of stories from the National Russian Corpus in the period from 2000 to 2022. In this paper the frequency distribution of the significant vocabulary was analyzed, and comparisons of the obtained results at different time slices were made. Thus, the tendencies in the change of frequency of some words and semantic groups of words were found. The comparison of the vocabulary of the first thirty years of the XX century and the vocabulary of the XXI century was carried out due to the parameter of relative frequency - *ipm*. With the help of this study it is possible to trace the influence of important historical events on the lexical composition of the language of fiction, to evaluate the trends to which the language has been subjected, and to identify the leading themes of the early 20th century and their differences from the themes of the 21st century.

Keywords: lexical studies; semantics; Russian short story; frequency dictionary; corpus linguistics; computational linguistics; statistical analysis

Введение

Исследование динамики лексического состава русской художественной прозы на материале частотных словарей корпуса русских рассказов 1900-1930 годов является важным вкладом в анализ художественной литературы данного периода. Благодаря ему представляется возможным сравнить динамику лексического состава языка и оценить тенденции, которые происходили в языке под влиянием исторических событий вышеуказанного периода. В данном исследовании применялись частотный и семантический подходы, благодаря которым можно проследить изменения не только отдельных слов, но и целых семантических групп – тем, которые имели важное значение для людей того времени, влияли на их речь, употребление терминов и определенных слов, а также на литературу и другие аспекты их жизни.

В данной выпускной квалификационной работе описывается изучение изменений, произошедших в русском языке в течение первых трех десятилетий XX века. За эти тридцать лет для России произошло много значительных событий, таких как Первая мировая война, Октябрьская революция, возникновение СССР вместо Российской Империи и т.п. Исследование проведено на материале аннотированной выборки из Корпуса русского рассказа 1900-1930 гг., где представлено более 300 текстов русских писателей, как популярных, так и малоизвестных. Материалы Корпуса делятся на три временных периода: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). Часть данного исследования была построена на произведении частотного анализа лексем, выделенных в верхние зоны каждого из периодов и выделение определенных тенденций и тем, которые можно было отметить благодаря подобному анализу.

Другая часть представляет собой сравнение лексики начала XX века с лексикой XXI века. Лексика двадцать первого века была взята из рассказов Национального корпуса русского языка (2000-2022 гг.). Так как объемы обоих корпусов оказались различны, для сравнения использовалась относительная

частота встречаемости – *ipm*. Таким образом, были выделены тенденции определенных выделенных раннее семантических групп и даны объяснения подобным закономерностям.

Целью данного исследования является выделение тенденций, происходящих во течение первых трех десятилетий XX века, а также сравнение частотных характеристик лексических единиц русского рассказа начала XX и XXI века.

Для достижения поставленной цели необходимо было решить следующие теоретические и практические **задачи**:

1. Построить на базе выборки из «Корпуса русского рассказа 1900-1930 гг.» частотные словари его подкорпусов.
2. Статистически выделить и семантически проанализировать верхние зоны частотного распределения лексем для данных подкорпусов.
3. На основании семантического анализа верхних зон частотных распределений лексем выделить основные темы, которые главенствовали в обществе в данные периоды времени и нашли свое отражение в анализируемых текстах.
4. Сравнить частотности входящих в выделенные тематические группы лексем с частотностями аналогичных лексем в современных русских рассказах.
5. Выявить тенденции в изменении частотности отдельных слов и лексических групп от одного исторического отрезка к другому и постараться их проинтерпретировать.

Объектом исследования является наиболее частотная лексика состава русской художественной прозы, выделенная в верхнюю зону частотности и распределенная по семантическим группам. В данной работе используются такие **методы**, как частотный и семантический анализ, а также диахронический подход к применению данных методов. **Материалом** работы являются

частотные словари Корпуса русских рассказов 1900-1930 годов и русские рассказы Национального корпуса русского языка (2000-2022 гг.).

Теоретическая значимость работы определяется разработкой методики выделения семантических групп и выявление зависимостей между частотностью употребления слов и историческими, социальными и иными явлениями, оказывающими сильное влияние на общество, общественное мнение, а также на темы, обсуждаемые большинством. **Практическая значимость** данной работы заключается в том, что полученные результаты и методики указывают на возможность применения такого метода не только для конкретного выделенного временного промежутка, но и любых других исторических срезов, которые исследователь может найти достаточно интересными для проведения такого рода исследования.

Содержание

Корпус русских рассказов 1900-1930 гг.

Предметом данной работы является корпус русских рассказов 1900-1930 гг., который разрабатывается в Санкт-Петербургском государственном университете. Корпус служит репрезентативным собранием текстов литераторов того времени, и на его основе представляется возможным выявить те или иные тенденции, наблюдаемые на протяжении трех десятилетий начала двадцатого века.

Корпус является объектом изучения такого научного направления в лингвистике, как корпусная лингвистика. «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий. Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» (В.П. Захаров, 2005). В данном случае лингвистической задачей является наблюдение за происходящими в различные временные периоды изменениями, такими как изменения частотности определенных слов, изменение частотности семантических групп, исчезновение слов из верхней частотной зоны и появление новых слов, как наиболее частотных.

Цель создания именно корпуса текстов, нежели некоторого собрания, заключается в определенных предпосылках:

1) большой, или репрезентативный, объем корпуса дает большую вероятность типичности данных, а также обеспечивает максимально полное представление всего спектра языковых явлений;

2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;

3) любой созданный массив данных может использоваться множество раз, не только одним, но многими исследователями, причем его можно использовать в совершенно различных целях.

Одно из самых главных свойств корпуса, которое уже было не раз упомянуто выше, является его репрезентативность. Так как задачей тех, кто создает корпус, является обладание наибольшим количеством текстов, которые могут относиться и относятся к подмножеству языка, для чьего изучения корпус и создавался, можно сказать, что корпус является в некотором роде уменьшенной моделью языка. При создании корпуса необходимо понимать, что его используют не только из-за обычно довольно большого количества языкового материала, но и из-за его пропорциональности. Таким образом, можно дать определение репрезентативности, как необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. (В.П. Захаров, 2005).

Необходимо понимать, что непосредственное успешное собрание текстов в корпус – лишь первый шаг к тому, чтобы с помощью его решать лингвистические задачи. Помимо этого, существует разметка. Корпус с разметкой называется размеченным корпусом. Разметка (или tagging) дает текстам дополнительную лингвистическую и экстралингвистическую информацию. Чаще всего она заключается в приписывании меток (tags) компонентам текстов или же текстам в общем – это могут быть как сведения о самом тексте (жанр, год создания, тематика), так и сведения об авторе – имя, пол и другие подробности, которые могут представлять для исследователя интерес, - в данном случае это называется метаразметкой. Так же метки могут быть структурные, то есть, как следует из названия, они указывают на структуру текстов – абзацы, предложения и т.д. И последние – это лингвистические, которые чаще всего описывают лексические, грамматические и другие характеристики текста, а точнее его элементов. В последней упомянутой лингвистической разметке чаще всего выделяют морфологическую, синтаксическую и семантическую разметку.

Синтаксическая разметка подразумевает описание различных связей между единицами текста, от лексических до синтаксических (таких, как придаточное предложение или глагольное словосочетание).

Семантическая разметка наиболее сложно реализуема в плане автоматизации. Она обозначает семантические категории слов или словосочетаний, а также возможных подкатегорий, которые могут относится к некоторому более узкому значению.

Порядок создания корпуса может разниться, но в общем виде можно выделить восемь главных шагов. Самым первым является создание списка источников. В случае приведенного в данной работе корпуса среди множества текстов с 1900 по 1930 годы было необходимо выделить лишь те, которые являлись рассказами. Всего было отобрано 310 рассказов, представляющих литературные произведения 300 русских писателей, среди которых всемирно известные писатели, такие как Антон Чехов, Лев Толстой, Максим Горький, большая группа относительно известных писателей - Андрей Белый, Владимир Короленко, а также менее известные или почти забытые авторы, например, Владимир Ленский, Евгений Опочинин. При этом в корпус не были включены рассказы писателей, которые эмигрировали; каждый писатель представлен одним рассказом за каждый временной отрезок.

Затем необходимо удостовериться, что все тексты оцифрованы, то есть необходимо тексты, существующие только на бумажных носителях, преобразовать в компьютерную форму. После этого происходит предобработка текстов – их корректировка и выверка. Помимо этого, проводят перекодировку и удаление или изменение различных нетекстовых элементов, таких как графики или рисунки, что не является обязательным шагом, но тем не менее необходимым для некоторого типа текстов. После проводится разметка текстов: она может быть организована вручную, а может быть и автоматической. Автоматическая разметка проводится с помощью специальных программных средств, таких как парсеры (parsers) или тэггеры (taggers). Стоит отметить, что хотя ПО (программные обеспечения) разметки постоянно улучшаются и уровень

качества проводимой ими работы несомненно растет, тем не менее по ряду причин они могут совершать ошибки. Из-за этого следующим шагом является непосредственная корректировка результатов автоматической разметки, если таковая требуется. Такая проверка и последующая корректировка может проводиться не только из-за ошибок, но и из-за неоднозначностей (полисемии, омонимии), которые приходится снимать вручную или полуавтоматически.

Заключительным, но еще не последним, этапом является преобразование полученных результатов в упорядоченную структуру специализированной лингвистической информационно-поисковой системы (corpus manager), обеспечивающей быстрый многоаспектный поиск и статистическую обработку (В.П. Захаров, 2005). Последний шаг – создание открытого доступа к корпусу, ведь одна из главных причин, по которой создается тот или иной корпус – возможность пользования им различными исследователями для своих научных работ.

Существуют различные типы корпусов. Например, порой противопоставляют корпуса, которые относятся ко всему языку в целом (например, НКРЯ – Национальный корпус русского языка), и корпуса подъязыка – корпуса, которые относятся или к определенному литературному жанру, или к временному периоду, или автору и т.д. Помимо этого выделяют корпуса, которые разделяются по типу синтаксической разметки. Таким образом можно сказать, что корпус, на котором основа данная исследовательская работа является корпусом подъязыка, основанного на рассказах русских писателей периода 1900-1930 гг.

Частотные словари

Как было упомянуто выше, данное исследование основано на материале частотных словарей Корпуса русских рассказов. Частотные словари (или частотные списки) уже давно являются частью стандартной методологии работы с корпусами. Синклер отмечает, что «любому, кто изучает текст, скорее всего, потребуется знать, как часто в нем встречаются различные формы слов» (Sinclair, 2004). Триббл и Джонс описали методику использования текстов в языковом классе, предложив, что наиболее эффективной отправной точкой для понимания текста является список слов с частотной сортировкой, где записывается количество раз, которое каждое слово встречается в тексте (Tribble, Jones, 1997). Поэтому частотный словарь может предоставить интересную информацию о словах, которые встречаются (и не встречаются) в тексте. Список слов может быть расположен в порядке первого появления, в алфавитном порядке или в порядке частоты. Порядок первого появления служит быстрым руководством по распределению слов в тексте, алфавитный список строится в основном для целей индексирования, а частотный список выделяет наиболее часто встречающиеся в тексте слова.

Наиболее активно за последнее десятилетие частотные словари стали использоваться для анализа лексического состава языка того или иного автора. Согласно частоте используемых слов можно выделить определенные темы, которые раскрывают лейтмотивы авторских произведений, в то же время выявляя общую картину мира писателя, волнующие его темы и то, какой точки зрения придерживается автор на их счет. Примерами таких работ являются частотный словарь рассказов Л.Н. Андреева (А.О. Гребенников, Г.Я. Мартыненко, 2003), частотный словарь рассказов А.И. Куприна (А.О. Гребенников, Г.Я. Мартыненко, 2006), частотный словарь рассказов А.И. Бунина (А.О. Гребенников, Г.Я. Мартыненко, 2011), а также частотный грамматико-синтаксический словарь языка художественных произведений А.П. Чехова с электронным приложением (О.В. Кукушкина, Е.В. Суровцева, Л.В. Лапонина, 2012). Последний из приведенных словарей является наиболее

репрезентативным для понимания роли частотных словарей в лингвистических (и не только) исследованиях.

Словарь А.П. Чехова состоит из четырех разделов. Первый раздел – непосредственно лексический состав языка писателя, который содержит более одного миллиона словоупотреблений. Они собирались с использованием грамматико-семантического принципа, то есть в корпусе использовалась не только частеречная разметка, но и методы семантического распределения на группы. Начало словаря отведено под имена собственные и их производные, что позволяет увидеть насколько мир А.П. Чехова индивидуализирован. Остальные же слова разделены на четыре группы – нарицательные и местоименные существительные; глаголы и глагольные предикаты; прилагательные, наречия, компаративы, именные предикативы; количественные слова. Второй раздел посвящен семантическим группам, которые могли бы быть интересны не только с лингвистической, но литературоведческой и когнитивной точек зрения, например «Наименования лиц» или «Отображения звучащей речи». Третий раздел содержит результаты количественного анализа лексики А.П. Чехова, то есть дан перечень наиболее регулярных слов, которые были использованы писателем в его произведениях (в основном те, которые употреблялись в ста и более произведениях). Также есть перечень слов с наибольшим числом употреблений, точнее первая сотня. В четвертом лежит описание электронного приложения к словарю, который содержит корпус художественных текстов А.П. Чехова. На его основе и был создан словарь, а также программное средство, с помощью которого можно проводить самостоятельное исследование собранных текстов в корпусе.

Несомненно, данный пример не является единственным способом, как можно представить частотный словарь писателя или автора, но тем не менее он является одним из наиболее подробных.

Одной из главных возможностей частотных словарей является не только выявления картины мира того или иного писателя, но также предоставление некоторой вероятности того, что частоты употребления слов определенным

автором могут проводить связь между изменениями тем писателя и реальными изменениями в обществе. Это продемонстрировали авторы «Статистического словаря языка Достоевского» при разработке теории лексических маркеров (то есть характерных слов текстов автора). Таким образом можно прийти к выводу, что возможности частотных словарей не останавливаются на конкретных авторах, но могут охватывать гораздо большие объемы информации. Одним из подобных примеров является серия частотных словарей рассказов русских писателей кафедры математической лингвистики филологического факультета СПбГУ, который содержит рассказы не одного, а 300 писателей определенного периода начала двадцатого века. Подобный корпус может рассказать не только о влиянии реальных событий на отдельного автора и его видение происходящего, но проследить общие мысли, идеи и видения мира на тот момент, сравнивая не двух и не трех, а множество писателей между собой.

Жанр рассказа

За период трех десятилетий начала двадцатого века произошли события, которые бесспорно оказали большое влияние на жизнь русского народа в целом, в частности на писателей. Такие события, как русско-японская война (1904-1905 гг.), так называемое Кровавое воскресение, ставшее началом первой российской революции (1905-1907 гг.), вступление России в Первую мировую войну в 1914 и участие в ней до 1918 года, Февральская революция (1917 г.), Гражданская война (1917-1922 гг.) и последующее образование СССР (1922 г.) – все это не могло не оставить свой след не только на простой социальной жизни любого из жителей России того времени, но и на литературу. Очевидно, что под влиянием таких исторических событий, будут меняться и история, и язык, появляться новые слова, исчезать некоторые другие.

Первоначальная идея рассмотрения событий и отношения людей к историческим происшествиям подобного глобального масштаба принадлежит русской формальной школе, а именно Ю.Н. Тынянову. Именно он выдвинул концепции синхронических и диахронических литературно-художественных систем. Синхронической системой являлась совокупность произведений определенного периода, эпохи, это могли быть как десятилетия, так и столетия. Под диахронической же системой подразумевалась определенная последовательность синхронических систем, следующих друг за другом. Так, например, в данной работе можно наглядно увидеть применение данной концепции – первый период с 1900 по 1913 год сменяется следующим - 1914-1922 гг. и завершается рассказами 1923-1930 гг.

Благодаря активному развитию исследований, основанный на так называемой big data, то есть исследований, которые работают с большими массивами данных, теперь представляется возможным увидеть изменения не только на произведениях одного писателя, но многих писателей целых эпох, а также возможность рассматривать не только авторов, но и целые литературные жанры того времени. В данной работе акцент сделан на жанре рассказа.

Рассказ – художественное повествовательное произведение небольшого размера, обычно в прозе (Толковый словарь Ушакова, 1940). Выбор именно этого жанра обусловлен тем, что за счет необязательности быть громоздким произведением, рассказ наиболее быстро и полно реагирует на происходящие события, отвечая на требования эпохи, покрывая большую часть ее разнообразных нюансов и аспектов, а порой даже и предвидя их. Так же благодаря малому объему рассказ позволяет привлечь к исследованию гораздо большее число авторов и их произведений, нежели иные прозаические жанры. К тому же именно рассказ наиболее остро реагирует на современность, ее действительность и язык, отражая все стороны реальной жизни, ее богатство и разнообразие индивидуальных стилевых систем (Г.Я. Мартыненко, Т.Ю. Шерстинова, 2018). Собрав различные рассказы определенного периода от различных авторов можно получить ту самую целостность картины мира, которая преобладала в той или иной исторической эпохе.

Составление списка авторов определенного временного периода базируется на основе существующих библиографий, энциклопедий, словарей писателей (в том числе и частотных), принимаются во внимание уже существующие корпуса и интернет-коллекции. После формирования такого списка каталогизируются их произведения.

НКРЯ

Одним из самых выдающихся корпусов является Национальный корпус русского языка (НКРЯ) – это собрание независимых корпусов, каждый из которых предназначен для решения определенных лингвистических задач [НКРЯ]. Это один из самых больших корпусов, каждая из коллекций текстов НКРЯ является репрезентативной и объемной. НКРЯ состоит из нескольких корпусов – основной, газетный, синтактический, параллельный, обучающий, диалектный, поэтический, устный, акцентологический, мультимедийный и исторический. НКРЯ содержит не только русский язык, но также древнерусский, церковнославянский и старорусские языки, как предки русского, белорусского и украинского языков – тексты на перечисленных древних языках можно найти в историческом корпусе. Хотя в большинстве корпусов тексты входят лишь на одном языке, есть исключения в виде параллельных корпусов, которые содержат не только русскоязычные тексты, но также их переводы на другие языки или же иноязычные тексты, переведенные на русский. В НКРЯ входит несколько десятков русско-иноязычных языковых пар и многоязычный корпус, в котором один и тот же текст может быть переведен на несколько языков.

Основной корпус состоит из русских письменных прозаических текстов, написанных после 1700 года. Отсчет времени в корпусе начинается с петровской эпохи, поэтому можно сказать, что он является примером отражения русского языка Нового времени. Основной корпус является представительным (репрезентативным) с точки зрения письменного языка каждой эпохи и включает в себя в определенных пропорциях различные жанры (художественные, научные тексты, публицистику, религиозные тексты, технические тексты, частную переписку).

Все тексты, которые входят в основной корпус, проходят процедуру метаразметки и морфологической разметки.

Морфологическая разметка для русского языка осуществляется с помощью специальных программ автоматического морфологического анализа.

Для многих текстов это адаптированная для корпуса система MyStem. Ряд частотных словоформ (в том числе архаичных, просторечных и т. п.), что могут встречаться в текстах, но анализ которых данной системой по тем или иным причинам невозможен, получает индивидуальный разбор, заданный списком. Для ряда устаревших морфологических вариантов анализатор дополнен автоматическими правилами. Тексты в старой орфографии анализируются тоже автоматически (однако леммы даются в новой орфографии).

Для около 6 миллионов словоупотреблений в корпусе было произведено ручное снятие омонимии, а также дополнительная коррекция результатов работы программы автоматического морфологического анализа DiaLing [DiaLing]. Эти словоупотребления образуют корпус со снятой омонимией, который может служить удобной базой для тестирования различных программ поиска, морфологического анализа и автоматической обработки текстов, а также для исследований современной русской морфологии, требующих повышенной точности поиска.

Обращаясь к метаразметке основного корпуса стоит обратить внимание на ее составляющие, а именно о сведениях о названии текста, дате его создания, имени, годе рождения и поле автора (если таковой известен), месте и дате публикации, источнике, по которому дается текст, его сфере функционирования, жанре и типе текста, хронотопе художественных произведений и мемуаров, специфике аудитории (массовость, возраст), орфографии, снятой или неснятой омонимии.

Все словоформы текстов, входящие в основной корпус, также получают автоматическую семантическую разметку, основанную на наборе дискретных семантических характеристик, приписываемых в словаре.

Помимо этого, благодаря метаразметке, в НКРЯ можно задать собственный подкорпус, задав необходимые параметры. Подкорпус уменьшает объем документов, подходящих для задачи исследователя и тем самым в некотором смысле упрощает исследование и делает его результаты более репрезентативными.

В данной работе большую роль играет семантическое свойство корпусов. Можно сказать, что корпус представляет собой пространственную метафору, в целом весьма характерную для языковой репрезентации многих основополагающих концептов сознания (С.Б. Кураш, 2015). Таким образом, корпусная лингвистика может играть большую роль на стыке нескольких наук или же их направлений, в том числе и самостоятельно кооперировать с ними и их методическим и методологическим аппаратом, что может помочь в получении некоего нового знания о тех объектах действительности, которые казались досконально изученными.

Семантика

Одним из этапов работы в данном исследовании является выделение семантических групп слов. Выявление тем в литературных произведениях, относящихся к художественному жанру, а не к академическому или медийному дискурсу, является довольно сложной задачей. Главная проблема заключается в том, что будучи экспрессивными текстами, описывающими не только происходящие события, но, в том числе, и их оценку, эмоциональное состояние героя или героев, отношения между персонажами, литературные тексты полнятся неявными смыслами, метафорами, образами. Из этого следует, что нет таких общих вычислительных или статистических методов, которые могли бы помочь в выделении подобных тем. Поэтому необходимо тщательно анализировать и качественно интерпретировать материалы по крайней мере на начальном этапе. Тем не менее из этого не следует, что автоматическое извлечение тем невозможно – они могут быть рассмотрены и разработаны позже, когда будет наработан определенный объем данных. Однако и тогда не стоит полностью на них полагаться. В данной работе выделение семантических групп было произведено вручную.

Уже в самом определении семантики может возникнуть некоторая неоднозначность. Несомненно, многие лингвисты согласились бы, что предметом семантики является значение, однако при этом встает вопрос, что подразумевается непосредственно под значением. И.М. Кобозева определяет семантику как раздел языкознания, изучающий содержание единиц языка и тех речевых произведений, которые из этих единиц строятся (И.М. Кобозева, 2000). Н.Н. Болдырев в контексте когнитивной семантики, являющейся одним из первых и центральных разделов когнитивной лингвистики, дает такое определение: «…когнитивная семантика – это многоуровневая теория значения, специфика которой заключается в том, что значения языковых единиц в ней анализируются в контексте всех знаний и опыта человек, а не только языковых явлений.» (Н.Н. Болдырев, 2014).

Таким образом, можно сделать вывод, что семантика занимается смыслом и значением единиц языка и произведений в непосредственной связи со знанием и опытом человека. Особенно в таком ключе стоит обратить внимание на когнитивную семантику, так как она уделяет внимание не только значению как таковому, но и тому, как картина мира отдельного индивидуума или даже общества может влиять на восприятие того или иного слова. Именно эта связь играет большую роль в данном исследовании, так как здесь реализуется попытка проследить изменения в том, как люди меняли свое отношение к предметам реального мира (например, что лошадь из категории транспорта перешла в категорию простого животного).

Когнитивная семантика

Термин «когнитивный» произошло от слова когниция (англ. cognition), который довольно затруднительно перевести на русский язык, но Н.Н. Болдырев предлагает обратиться к определению, данному в Краткой философской энциклопедии: когниция – это знание, познание. Это такое понятие, которое принимает во внимание как теоретическое познание, так и обыденное и не всегда осознанное постижение мира человеком, приобретенного через опыт – чувственный, сенсорный или же телесный, при взаимодействии с окружающим миром, восприятии этого мира, наблюдении за ним, попытках категоризации, что включает в себя мышление, речь, воображение и многие другие психические процессы или даже их совокупность. Проблема, которую решает когнитивная лингвистика, заключается в том, как понимание и видение мира человеком отражено в его сознании.

Лексику в общем виде можно представить в виде совокупности частных систем, которые называются семантическими полями или группами, где слова связаны определенными отношениями. Теория семантического поля впервые появилась благодаря немецкому ученому Иосту Триру. Согласно ей на каждое семантическое или «понятийное» поле накладываются слова, которые в свою очередь образуют «словесное» поле. В современном же языкознании семантическое поле определяется как совокупность языковых единиц, объединенных общность содержания и отражающих понятийное, предметное или функциональное сходство обозначаемых явлений (И.М. Кобозева, 2000). Выделяют несколько основных свойств семантического поля:

1. наличие семантических отношений между единицами семантического поля;
2. данные отношения должны иметь системный характер;
3. лексические единицы семантического поля могут взаимно определять друг друга;
4. поле автономно;

5. семантические поля взаимосвязаны в пределах лексической системы.

Существуют различные отношения (корреляции) между словами в семантическом поле:

- Синонимия – автомобиль, машина, авто.
- Гипонимия, то есть родо-видовые отношения между словами. Выделяют гипонимы, гиперонимы и согипонимы, где гипероним – слово, которое выражает общее понятие в корреляции, гипоним – частный случай указанного типа объектов, а когипонимы (или согипонимы) обозначают слова, которые имеют один гипероним. Например, гипероним «дерево» имеет гипонимы «береза», «ель», «пальма», которые являются согипонимами друг для друга.
- Выделяют несовместимость, как одно из семантических отношений, например мать и отец. Несовместимость в данном случае выражается в том, что эти слова не могут в один и тот же момент времени относится к одному и тому же объекту или выражать одно и то же явление.
- Учитывается корреляция «часть-целое» - связь предмета с наименованиями некоторых его составных частей. Так, в примере с деревом такая корреляция будет реализована со словами «ветка», «листья», «корни» и т.д.
- Нельзя не упомянуть про антонимию – тепло-холодно, утро-вечер и т.п., то есть слова, выражающие противоположные понятия.
- Конверсивность – связь между словами, которые обозначают одну ситуацию, но с разных точек зрения участников данной ситуации: выиграть – проиграть.
- Может быть корреляция семантической производности – это связи между словами на том основании, что такие слова оказываются часто формально связанными словообразовательным отношением

производного и производящего слова, как слушать – слушатель или идти - шел.

- Ассоциативные отношения. Наличие таких отношений было выявлено в ходе психолингвистических экспериментов, во время которых испытуемым предлагалось перечислить слова, приходящие им в голову в связи с некоторым словом-стимулом. Например, в русском языке при слове-стимуле «осел» могут прийти в голову такие ассоциации, как «глупость», «упрямство», «лень».

В контексте данной работы отношения в семантических группах будут комбинироваться в зависимости от первых значений использованных в ней слов.

Проблемы, с которыми можно столкнуться при выделении семантических групп, заключаются в таких явлениях как полисемия и омонимия. В данных случаях происходит варьирование означаемого у одного и того же означающего, что связано с некоторыми экстралингвистическими факторами – определенной обстановкой или участниками речевого акта. Ю.Д. Апресян выделяет речевую многозначность – это представленность узуального значения в речевом употреблении двумя и более вариантами, выбор между которыми обусловлен экстралингвистическим контекстом, в частности, знаниями о мире. Существуют такие связи, которые позволяют лингвистам объединить узуальные значения одного означающего, связанные между собой, как представляющие собой виртуальное значение одно и того же, хотя и многозначного слова – такие случаи называют полисемией. Однако если у узуальных значений одного и того же словесного означающего не существует какой-либо общей части, то хоть они по форме и тождественны, но воспринимаются как реализации разных слов. Это явление называется омонимией.

Важно разграничивать полисемию и омонимию. Это разграничение сводится к установлению наличия или отсутствия достаточной степени сходства между значениями (И.М. Кобозева, 2000). Однако даже при такой интерпретации понятие сходства является довольно размытым и неясным. Некоторые факты могут быть интерпретированы разными лингвистами по-разному и как

омонимия, и как полисемия. Именно поэтому необходимо уточнить само понятие сходства значений. Само определение было дано Ю.Д. Апресяном: «Значения a_i и a_j слова А называются сходными, если существуют такие уровни семантического описания, на которых их толкования или коннотации имеют нетривиальную общую часть, и если она выполняет в толкованиях одну и ту же роль относительно других семантических компонентов» (Ю.Д. Апресян, 1974). Тем не менее возможность адекватного разграничения полисемии и омонимии по большей части зависит от того, насколько точны описания лексических значений, которые имеются в нашем распоряжении.

Корпус и лемматизация

Лексика является важнейшим компонентом любого естественного языка. Списки частотных слов - удобное представление функциональной активности слов в языке в целом или в конкретном тексте. Списки или словари частотных слов находятся в центре внимания специалистов по NLP (Natural Language Processing), поскольку они используются в многочисленных исследованиях, связанных с определением авторства, автоматической кластеризацией и классификацией текстов.

Частотные списки, на которых проводилась данная исследовательская работа, были составлены на собранном Корпусе русских рассказов в период с 1900 по 1930 год. Всего было отобрано 310 рассказов, представляющих литературные произведения 300 русских писателей, среди которых всемирно известные писатели, такие как Антон Чехов, Лев Толстой, Максим Горький, большая группа относительно известных писателей - Андрей Белый, Владимир Короленко, а также менее известные или почти забытые авторы, например, Владимир Ленский, Евгений Опочинин. При этом в корпус не были включены рассказы писателей, которые эмигрировали; каждый писатель представлен одним рассказом за каждый временной отрезок.

Корпус разделен на три подкорпуса, которые относятся к основным историческим периодам рассматриваемой эпохи. Так как ожидается, что социальный фон этих исторических периодов сильно отличается, то можно предположить, что частотные списки слов, отражающие язык этих периодов, также будут отличаться. В результате имеется три зоны:

1. Начало века (1900-1913 гг.). Из главных событий в этот период можно выделать русско-японскую войну и первую российскую революцию.
2. Военное время (1914-1922 гг.) – вступление России в Первую мировую войну, помимо этого Февральская и Октябрьская революции и Гражданская война.

3. Советское время (1922-1930 гг.) – как результат Октябрьской революции и Гражданской войны – последующее образование СССР.

На материале исследуемой выборки из Корпуса были построены частотные словари, которые располагались в порядке убывания частот. Данные частотные словари были составлены и для выборки в целом, 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. Для проведения лемматизации и составления частотных словарей была использована система UNILEX-T, предназначенная для обработки текстов (подготовки частотных словарей и автоматических конкордансов, эксплуатации автоматических конкордансов как автоматических словарей) (Ж.Г. Аношкина, 1995).

77 текстов были написаны почти забытыми или "редкими" авторами, поэтому они были специально оцифрованы для включения в корпус. Тексты выделенного подкорпуса написаны в разные годы и имеют разный размер. Отбор текстов происходил случайным образом, поэтому данные факторы не были приняты во внимание.

Корпус был аннотирован на лексическом, морфологическом и, выборочно, на синтаксическом и ритмическом уровнях. Тексты подверглись обработке и стали разделены на речь рассказчика и речь персонажей. Была произведена редактура результатов автоматической обработки текстов - удалена грамматическая омонимия.

Программа использует следующие POS категории: S (существительное), V (глагол), PR (предлог), CONJ (союз), A (прилагательное), ADV (наречие), PART (частица), NUM (числительное), INTJ (междометие), COM (часть составного слова). В данной работе количество POS категорий было сокращено до пяти, а их наименования были заменены на: С (существительное), Г (глагол), П

(прилагательное), Н (наречие) и ЧИСЛ (числительное). Такой первоначальный вид данных можно наблюдать в Таблице 1.

Таблица 1 – Частотный список слов первой зоны

Ранг	***Слово***	***Часть речи***	***Частота***
1	СКАЗАТЬ	(Г)	1222
2	ГЛАЗ	(С)	1147
3	ГОВОРИТЬ	(Г)	1135
4	РУКА	(С)	1077
5	МОЧЬ	(Г)	1010
6	ЛИЦО	(С)	809
7	ЗНАТЬ	(Г)	790
8	ГОЛОВА	(С)	687
9	ТЕПЕРЬ	(Н)	683
10	ЖИЗНЬ	(С)	646
11	ДЕНЬ	(С)	640
12	ЧЕЛОВЕК	(С)	603
13	КАЗАТЬСЯ	(Г)	587
14	ВДРУГ	(Н)	559
15	ДУМАТЬ	(Г)	547
16	ЛЮДИ	(С)	532
17	ВРЕМЯ	(С)	507
18	ГОЛОС	(С)	498
19	ХОТЕТЬ	(Г)	498
20	ВИДЕТЬ	(Г)	496
21	ДОМ	(С)	471
22	БОЛЬШОЙ	(П)	470
23	СМОТРЕТЬ	(Г)	469
24	ДВА	(ЧИСЛ)	461

Таблица 1 содержит следующие данные: ранг леммы (то есть ее статистический вес), саму лемму, ее частеречную принадлежность и относительную частоту.

Как уже упоминалось выше, частотные списки лемм и словоформ были составлены для каждого рассказа, для корпуса в целом и, что наиболее важно, для каждого из трех подкорпусов. Таким образом, просто сравнив первые двадцать пять слов всех трех зон, уже можно заметить, что есть некоторые частотные сдвиги и изменения, которые могут дать более полную и детальную картину мира того времени. Это отмечено в Таблице 2:

Таблица 2 – Первые 25-ть слов частотных списков зон

I зона			**II зона**			**III зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
1	СКАЗАТЬ	1222	1	ГОВОРИТЬ	962	1	РУКА	1156
2	ГЛАЗ	1147	2	СКАЗАТЬ	846	2	ГЛАЗ	1153
3	ГОВОРИТЬ	1135	3	ГЛАЗ	803	3	СКАЗАТЬ	1080
4	РУКА	1077	4	РУКА	786	4	ГОВОРИТЬ	979
5	МОЧЬ	1010	5	МОЧЬ	631	5	МОЧЬ	866
6	ЛИЦО	809	6	ЗНАТЬ	571	6	ГОЛОВА	760
7	ЗНАТЬ	790	7	ИДТИ	569	7	ЗНАТЬ	720
8	ГОЛОВА	687	8	ЛИЦО	562	8	ИДТИ	715
9	ТЕПЕРЬ	683	9	ГОЛОВА	528	9	ДЕНЬ	665
10	ЖИЗНЬ	646	10	ДЕНЬ	506	10	ЧЕЛОВЕК	616
11	ДЕНЬ	640	11	СТАТЬ	472	11	ЛИЦО	607
12	ЧЕЛОВЕК	603	12	ЧЕЛОВЕК	471	12	ДВА	555
13	КАЗАТЬСЯ	587	13	ЛЮДИ	419	13	НОГА	553
14	ВДРУГ	559	14	БОЛЬШОЙ	407	14	БОЛЬШОЙ	503
15	ДУМАТЬ	547	15	ДВА	407	15	ДЕЛО	480
16	ЛЮДИ	532	16	ЖИЗНЬ	393	16	ПОЙТИ	475
17	ВРЕМЯ	507	17	ТЕПЕРЬ	380	17	ДУМАТЬ	459
18	ГОЛОС	498	18	НОЧЬ	368	18	СИДЕТЬ	448
19	ХОТЕТЬ	498	19	ДУМАТЬ	366	19	ВИДЕТЬ	447
20	ВИДЕТЬ	496	20	ХОТЕТЬ	363	20	ХОТЕТЬ	436
21	ДОМ	471	21	НОГА	339	21	ВРЕМЯ	432
22	БОЛЬШОЙ	470	22	ГОЛОС	336	22	ДВЕРЬ	427
23	СМОТРЕТЬ	469	23	ДЕЛО	334	23	ТЕПЕРЬ	426
24	ДВА	461	24	СЛОВО	327	24	НОЧЬ	425
25	НОЧЬ	449	25	СИДЕТЬ	325	25	ЗЕМЛЯ	421

Можно увидеть, что хоть и большая часть слов во всех трех зонах все так же находится среди первых 25-ти наиболее частотных слов, тем не менее заметны изменения в рангах как на самых первых, так и на последних позициях. Так, первое место в первой зоне занимает слово *сказать*, во второй – *говорить*, а в третьей – *рука*. Слово *сказать* теряет свое первенство уже во второй зоне, получая второй ранг, а в третьей вовсе уходит на третью позицию. Некоторые слова, такие как *жизнь* или *идти* находятся среди первых 25-ти слов только в двух зонах из трех, а некоторые либо получают ранг ниже 25 (*казаться, дом, смотреть*), либо наоборот появляются, заняв место среди 25 первых (*слово, пойти, земля*). Таким образом явственно видно, что за определенный

исторический период даже у наиболее частотных слов происходят изменения, которые могут быть объяснены определенной тенденцией или закономерностью.

Выделение семантических групп

Было решено выделить семантические группы отдельно у каждой части речи – имени существительного, глагола, имени прилагательного, наречия и числительного. Таким образом каждый частотный список определенного периода был распределен на пять подгрупп с сохранением их статистических значений, таких как частота и ранг. Так исследование тенденций в период первых трех десятилетий начала двадцатого века было более детализированным, поскольку можно было отметить и выявить такие семантические группы, которые могли бы «потеряться» при анализе всего частотного списка без частеречного разделения. Более того, так можно было заметить общие лингвистические тенденции, характерные скорее в целом для языка, чем для определенной эпохи, а именно – превалирующее количество слов, относящееся к определенной части речи. На Рисунке 1, Рисунке 2 и Рисунке 3 видно, что лидирующую позицию занимают имена существительные, глагол является вторым по количеству употреблений, имена прилагательные же используются почти вполовину меньше, чем глаголы.

Рисунок 1 – Распределение слов по частям речи (I зона)

Рисунок 2 – Распределение слов по частям речи (II зона)

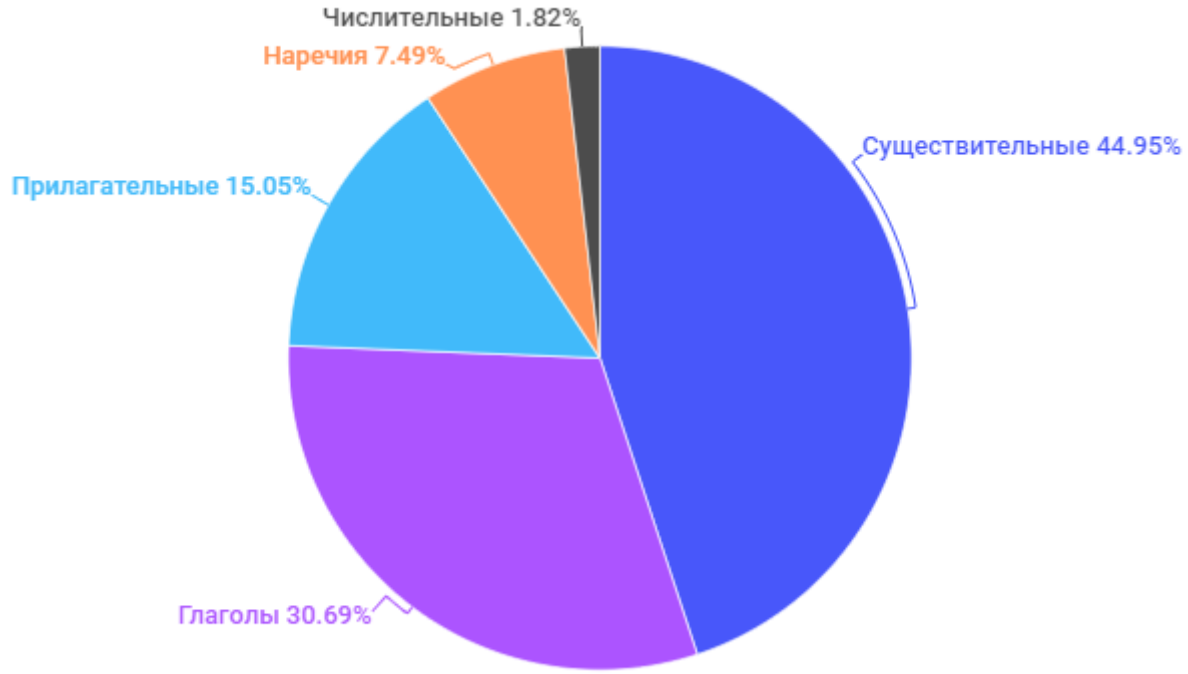

Рисунок 3 – Распределение слов по частям речи (III зона)

Семантические группы были выделены вручную, так как на данный момент не было выведено общих статистических или вычислительных методов, которые можно было бы использовать для автоматизации данного процесса. Другая трудность, связанная с выделением тематического содержания такого художественного текста как рассказ, заключается в том, что он обычно содержит несколько тем, причем они не расположены иерархически, так что нельзя однозначно обозначить одну из них как доминирующую, а другие - как подчиненные.

Так как тематическое маркирование представляет собой выделение максимального количества семантических групп, то каждый рассказ получает некоторый набор тем, хотя стоит отметить, что данный набор не определяет полностью сюжет рассказа. В данной работе на основе рассказов каждого периода были составлены отдельные семантические группы – то есть наборы тем. Некоторые темы могли уходить из одной зоны и появляется в другой, некоторые появлялись только в последнем периоде, количество слов, относящейся к одной группе, могло увеличиться в зависимости от исторических событий, произошедших в определенный временной период, а другое наоборот могло уменьшиться. Всего тематический набор для всего подкорпуса, вне зависимости от зоны, содержит 86 тем. Некоторые темы оказались общими для нескольких частей речи, например, *война, жизнь, свет, эмоции*, другие же оказались конкретными темами именно определенной части речи. Так, у существительных – «Транспорт», «Крестьянский быт», «Абстрактные существительные»; у глаголов – «Глаголы состояния», «Действия над кем-то, чем-то», «Насильственные действия»; у прилагательных – «Цвет», «Классовая характеристика», «Отношение к группе/представителю группы»; у наречий – «Причина», «Оценка действия», «Усиление/уменьшение и бОльший охват». У числительных же не было выделено никаких тем кроме самой темы «Числительные».

Важно отметить, что темы, относящиеся к социально-политической составляющей рассказов, могут и будут возникать в художественной литературе

не сразу в тот же период, в который произошли, а скорее всего позже, спустя время после этих событий. Таким образом, ожидалось, что большая часть лексики, относящейся к революции, будет находиться во второй зоне, хотя первая революция в России началась в выделенном первом временном периоде (1900-1913 гг.).

Слова были распределены по семантическим группам, на основании их первых или же вторых значениях в электронном Толковом словаре С.И. Ожегова и Н.Ю. Шведовой, если первое значение определенно не подходило ни под одну выделенную семантическую группу. Проблема, с которой пришлось столкнуться на данном этапе заключалась в том, что некоторые слова могли относиться сразу к некоторым семантическим группам. Так, например, слово *привезти* во второй зоне находилось в семантической группе «Движение», однако в третьей зоне было перемещено в группу «Производить действие над кем-то/чем-то». Такие изменения в семантических группах являются не ошибкой, но лишь ярким примером проблемы полисемии.

Некоторые семантические группы, выделенные в первой и второй зонах, оказались несостоятельными в третьей. Так, например, в группе «Профессия» из зоны первого периода часть слов ушла в семантическую группу «Статус», а другая часть вышла из частотного списка (см. Таблицу 3 и Таблицу 4).

Таблица 3 – «Профессия» (I зона)

Профессия		
Ранг	***Слово***	***Частота***
259	ДОКТОР	115
473	ИЗВОЗЧИК	69
594	СТОРОЖ	57
610	НАЧАЛЬНИК	55
722	ХУДОЖНИК	49
726	ГОРНИЧНАЯ	48
847	КУПЕЦ	42
884	ЧИНОВНИК	41
942	ПРИКАЗЧИК	39
1095	ЖАНДАРМ	33

Таблица 4 – «Статус» (II зона)

Статус		
Ранг	***Слово***	***Частота***
148	ГОСПОДИН	127
168	ГОСТЬ	115
210	ПОЭТ	99
259	ДАМА	85
313	ГУБЕРНАТОР	75
327	ХОЗЯИН	73
399	ПАН	62
401	ПРЕДСЕДАТЕЛЬ	62
472	**ДОКТОР**	53
488	**КУПЕЦ**	51
510	МАТРОС	49
517	БАРЫШНЯ	48
539	**НАЧАЛЬНИК**	46
559	ДИРЕКТОР	44
639	**ИЗВОЗЧИК**	39
657	**ЧИНОВНИК**	39
681	ВОР	37
703	БАРИН	36
713	ГРАЖДАНИН	36
726	СОСЕД	36
767	КАЗНАЧЕЙ	34
848	ВРАГ	31
963	РАЗБОЙНИК	28
1052	НАЧАЛЬСТВО	26
1056	ОХОТНИК	26
1078	ХОЗЯЙКА	26

Семантическая группа «Нейтральные эмоции» в Таблице 5 оказались как в первой, так и в третьей зоне, но не присутствовала во второй, возможно потому, что эмоциональный фон таких событий как революции и войны не мог быть нейтральным.

Таблица 5 – «Нейтральные эмоции» (I и III зоны)

Нейтральные эмоции					
I зона			**III зона**		
Ранг	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***
976	УДИВЛЯТЬСЯ	38	1049	УДИВИТЬСЯ	34
1035	УДИВИТЬСЯ	36	1050	УДИВЛЯТЬСЯ	34

Что касается новых появляющихся семантических групп, то во втором периоде можно отметить появление таких тем как «Нации и страны», «События» и «Символы», а в третьем – «Смерть», «Этапы работы/деятельности» и группу «Статус» в наречиях. Наименее изменяемыми частями речи оказались наречия и числительные. Единственное изменение, произошедшее на уровне семантических групп с наречиями, было появление группы «Температура», в то время как для числительных наблюдается единая группа, где происходят лишь сдвиги в частотности и, соответственно, рангах, как можно видеть в Таблице 6:

Таблица 6 – Числительные всех трех зон

Числительные								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
24	ДВА	461	15	ДВА	407	12	ДВА	555
42	ПЕРВЫЙ	324	56	ПЕРВЫЙ	213	43	ПЕРВЫЙ	324
50	НЕСКОЛЬКО	295	94	ТРИ	164	69	ТРИ	261
100	ТРИ	208	153	ОБА	123	118	НЕСКОЛЬКО	197
258	ДЕСЯТЬ	115	233	СКОЛЬКО	92	190	ОБА	140
288	МАЛО	105	237	ТРЕТИЙ	91	220	СКОЛЬКО	126
304	СКОЛЬКО	100	272	МАЛО	82	233	ПЯТЬ	121
319	ПЯТЬ	96	280	ПЯТЬ	81	273	ВТОРОЙ	108
367	ТРЕТИЙ	86	339	ВТОРОЙ	70	277	ЧЕТЫРЕ	106
471	ДВАДЦАТЬ	69	353	ДЕСЯТЬ	67	322	МАЛО	94
500	ВТОРОЙ	66	519	ДВОЕ	48	323	ТРЕТИЙ	94
575	ЧЕТЫРЕ	59	527	ДВАДЦАТЬ	47	359	ДЕСЯТЬ	88
791	СТОЛЬКО	46	656	ТРИДЦАТЬ	39	421	ДВАДЦАТЬ	78
816	НЕМНОЖКО	44	702	ЧЕТЫРЕ	37	599	ДВОЕ	55
831	ДВОЕ	43	760	ШЕСТЬ	35	658	ТРИДЦАТЬ	51
970	СЕМЬ	38	810	ТРОЕ	33	743	ТЫСЯЧА	46
974	ТРИДЦАТЬ	38	840	СТО	32	759	ШЕСТЬ	45
975	ТЫСЯЧА	38	863	СТОЛЬКО	31	786	СЕМЬ	44
1060	СТО	35	947	ВОСЕМЬ	28	859	ТРОЕ	41
			966	СЕМЬ	28	890	СТО	40
			973	ТЫСЯЧА	28	891	СТОЛЬКО	40
			1041	ДВЕНАДЦАТЬ	26	925	ЧЕТВЕРТЫЙ	39
						986	СОРОК	37
						1215	ВОСЕМЬ	31
						1216	ПЯТЫЙ	31

Для описания динамики частотности слов в имеющихся частотных списках и в семантических группах, а также в различных зонах, используется

специальная нотация. При слове в скобках указывается его текущий ранг, например, *два* (15). Если слово только появилось в n зоне и его ни разу не было в частотном списке до этого периода, то пишется: *голодный* (→ 819). Если слово изменяет свой ранг, переходя из одной зоны в другую, то это записывается как: *солдат* (123→58), что означает, что в первой зоне у него был ранг 123, а затем переместился на 58-ю позицию во втором периоде. Если слово было в первой зоне, не было во второй и появилось в третьей (то есть, можно сказать, что слово «пропустило» вторую зону), запись выглядит так: *твердый* (580→→796). Наконец, если в более ранний период слово присутствовало в верхней зоне, а затем его покинуло, то выглядит это как: *атаман* (1068→). При указании общих тенденций по всем трем зонам, нотация такова: *радость* (240 → 202 → 204).

Стоит остановиться на отдельных группах, которые явственно выражают закономерности, наблюдаемые в изменении состава лексики литературного языка русских рассказов, основанные на событиях, происходящих в указанные временные периоды.

Война и/или революция

Нельзя подвергать сомнению тот факт, что военные события в любой форме, будто то война или революция, оказывают огромное влияние на жизнь индивида и общества в целом, причем не только на уровне мыслей, слухов и новостей, но и экономическом, государственном и социальном. Тем не менее, как уже было упомянуто ранее, стоит ожидать, что чем значительнее произошедшие события, тем позже, скорее всего, они будут описаны в художественной литературе, ведь речь идет не только об исторических явлениях, но и об эмоциональном фоне, отношении к данным событиям героев и самого автора.

Ниже можно увидеть количество слов трех частей речи (имени существительного, глагола и имени прилагательного), распределенных по трем зонам. Так, имен существительных несравненно больше, чем иных частей речи, причем, как можно увидеть на Рисунке 4, хотя первые военные действия в виде революции (1905-1907 гг.) и начались в первом периоде, более полно данная тема раскрывается даже не во втором, а третьем периоде, где находится наибольшее количество существительных и глаголов и несколько прилагательных, относящихся к теме войны и революции.

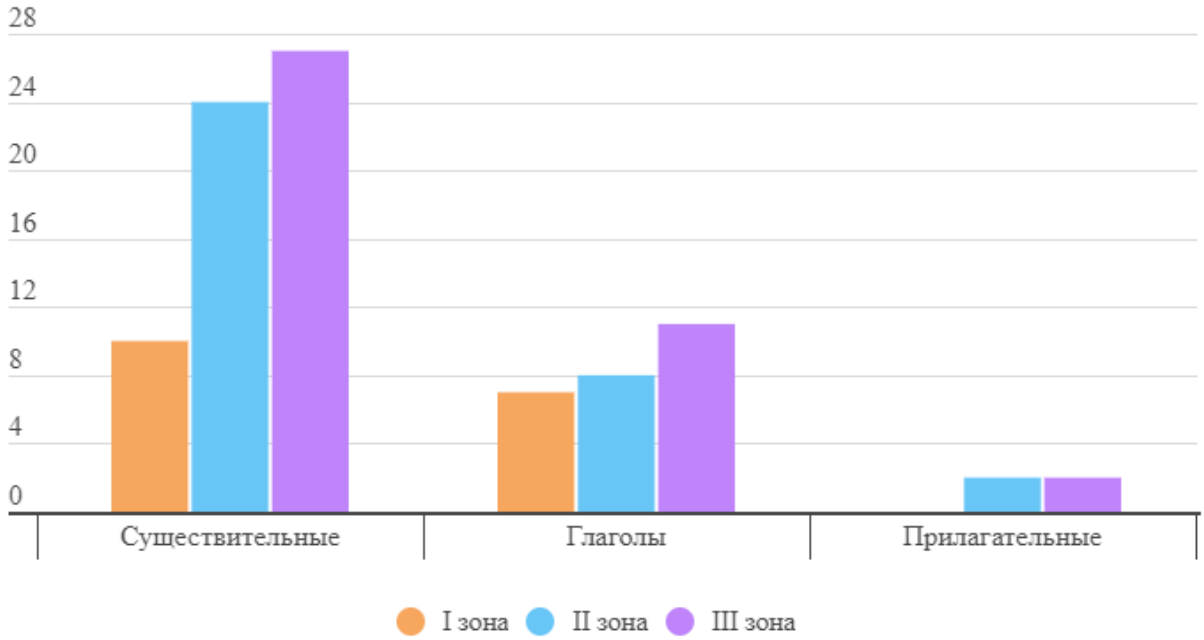

Рисунок 4 – Количественное распределение слов по частям речи (тема «Война и/или революция»)

Однако количество слов в семантической группе еще не дает полного представления о том, насколько важна эта тема в данный временной период. Необходимо посмотреть, как ведут себя ранги в разных зонах, насколько они малы или велики, то есть, насколько частотны или не частотны эти слова (Таблица 7).

Мы видим, что хотя в третьей зоне находится больше слов, относящихся к теме «Войны и/или революция», ранг и частота употребления таких слов как *солдат* (58 → 232) и *офицер* (172 → 411) чуть ли не в три раза чаще используются в рассказах второй зоны, чем третьей. Однако *революция* (889 → 363) очевидно является гораздо более часто употребляемой в третьем периоде, помимо этого появляется слово *красноармеец*, так как от слова *солдат* на определенный период отказались, как от «контрреволюционного», хотя на тот период эти изменения еще не вступили в силу, так как частотность слова *солдат* (300) в третьей зоне все еще несравнимо выше, чем у слова *красноармеец* (851).

Таблица 7 – Война и/или революция (все зоны)

Война и/или революция								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
123	СОЛДАТ	191	58	СОЛДАТ	212	232	РОТА	121
325	КРОВЬ	95	172	ОФИЦЕР	114	296	РУЖЬЕ	100
347	КАПИТАН	89	290	РОТА	80	300	СОЛДАТ	99
535	РОТА	63	322	ВОЙНА	73	363	РЕВОЛЮЦИЯ	87
546	ОФИЦЕР	62	356	ОКОП	67	371	ВОЙНА	85
569	БОРЬБА	59	372	ВИНТОВКА	64	389	КАЗАК	83
669	БАРАК	51	416	ПОРУЧИК	59	410	КОМАНДИР	80
697	ПОЛКОВНИК	50	493	РУЖЬЕ	51	411	ОФИЦЕР	80
1001	БАРРИКАДА	36	540	ПОЛКОВНИК	46	416	ВИНТОВКА	79
1068	АТАМАН	34	612	ПУЛЯ	41	440	ОТРЯД	75
			651	ПРАПОРЩИК	39	444	ВЫСТРЕЛ	74
			762	ВЫСТРЕЛ	34	584	ГЕНЕРАЛ	56
			823	КАЗАРМА	32	593	КАПИТАН	55
			844	ФРОНТ	32	631	АРМИЯ	52
			851	КОМАНДИР	31	677	ФРОНТ	50
			889	РЕВОЛЮЦИЯ	30	721	ПОЛКОВНИК	47
			906	ГЕНЕРАЛ	29	824	БОЙ	42
			912	КАЗАК	29	851	КРАСНОАРМЕЕЦ	41
			952	КАПИТАН	28	888	ШАШКА	40
			959	ОТРЯД	28	922	ПУЛЕМЕТ	39
			1017	ПУЛЕМЕТ	27	942	БОРЬБА	38
			1034	ФЕЛЬДФЕБЕЛЬ	27	1007	ПУЛЯ	36
			1081	БОРЬБА	25	1142	ВОЕНКОМ	32
			1191	БОЙ	23	1188	ЗНАМЯ	31
						1197	ОРУЖИЕ	31
						1365	МИНОНОСЕЦ	28
						1370	ПОЛК	28

Религия

Следующей темой, на которую необходимо обратить особое внимание, является «Религия». Во время войны, когда исход следующего дня неизвестен, люди в большинстве своем пытаются найти утешения у Бога, в церкви. Ожидалось, что в период второй зоны произойдет увеличение частотности слов, связанных с религией и верой, однако в третьей будет наблюдаться спад из-за государственной идеологии атеизма во время советского периода (Рисунок 5).

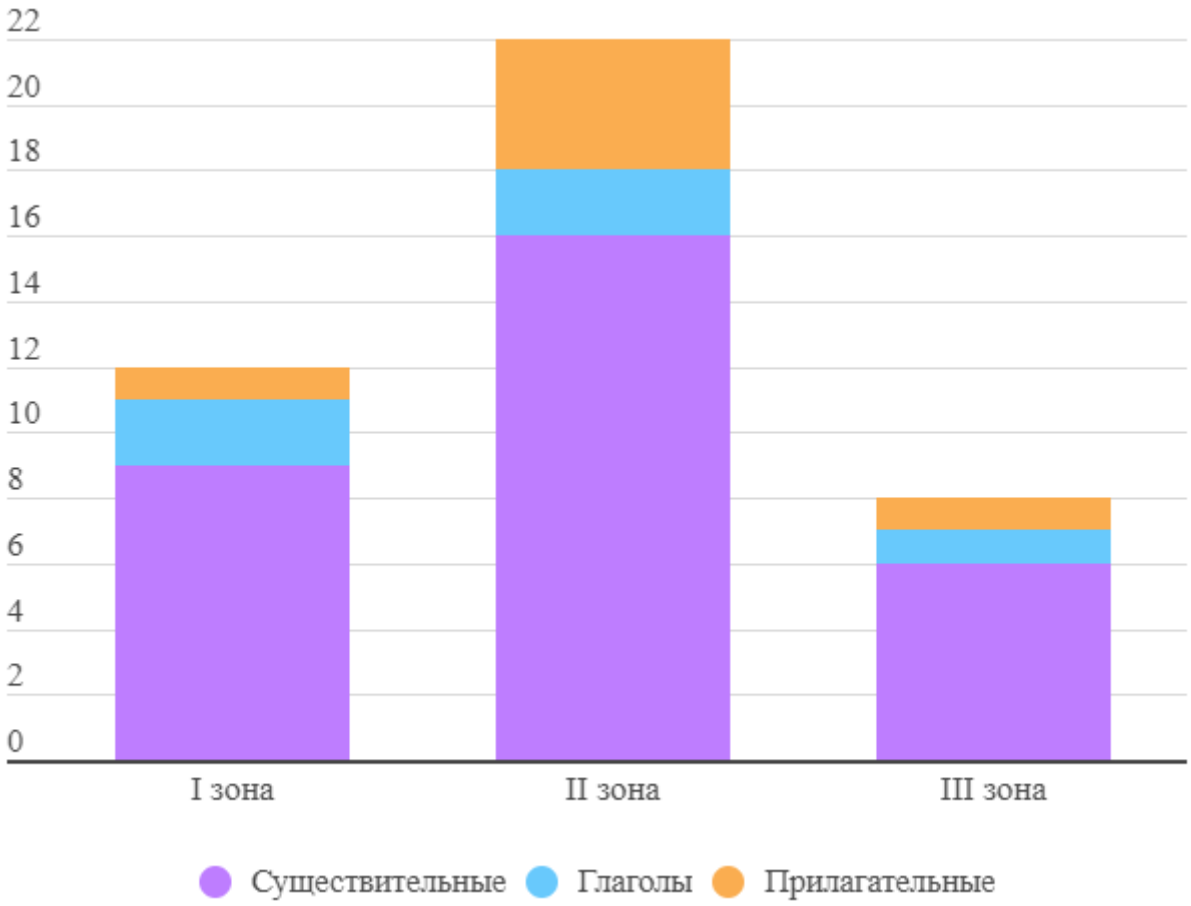

Рисунок 5 – Распределение слов темы «Религия» по трем периодам 1900-1930 гг.

Как видно выше, действительно наблюдается подъем в использовании религиозных слов во втором периоде – *бог* (95 → 61), *церковь* (432 → 262), *грех* (727 → 509), *молиться* (851 → 499), *святой* (821 → 311), но при этом ранги этих же слов разительно меняются в период с 1923 по 1930: *бог* (61 → 225), *церковь* (262 → 758), *грех* (509 → 1005), *молиться* (499 → 1222), *святой* (311→ 1058).

Транспорт

Семантическая группа «Транспорт» (см. Таблицу 8) может показать то, как шло развитие разработки средств для передвижения.

Таблица 8 – «Транспорт» (все зоны)

Транспорт								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
227	ДОРОГА	129	171	ДОРОГА	114	182	ВАГОН	144
241	ДВИЖЕНИЕ	123	266	ПУТЬ	84	265	МАШИНА	112
346	ВАГОН	89	384	ВАГОН	63	271	ПОЕЗД	109
473	ИЗВОЗЧИК	69	465	ПОЕЗД	54	284	ПУТЬ	103
484	ПУТЬ	68	507	ВЕРСТА	49	315	ТЕЛЕГА	95
509	ЯМЩИК	66	578	ТЕЛЕГА	43	340	ВЕРСТА	91
527	ПОЕЗД	64	655	СТАНЦИЯ	39	349	СТАНЦИЯ	89
647	СТАНЦИЯ	53	847	ВОКЗАЛ	31	473	ПАРОХОД	71
734	ЛОДКА	48	921	ЛОДКА	29	502	КОЛЕСО	67
739	ПАРОХОД	48	943	АВТОМОБИЛЬ	28	528	ЛОДКА	63
953	ВЕРСТА	38	991	КОЛЕСО	27	849	ВОКЗАЛ	41
						1027	АВТОМОБИЛЬ	35
						1198	ПАЛУБА	31
						1367	ПАРОВОЗ	28

Например, много новых слов появляется в третьей зоне, которая отмечена периодом развития технологий, в особенности новых способов для перемещения из точки А в точку Б. К тому же именно в начале двадцатого века после Октябрьской революции началось активное строительство железных дорог. Машины стали появляться примерно в это же время, это можно увидеть по тому, что в третьей зоне слово *машина* (265) занимает второе место, уступая только *вагону* (182). *Поезд* же стал гораздо частотнее употребляться, судя по рангу. Так в первом периоде поезд имел 527 ранг, во втором – 465 и наконец в третьем 271. Помимо этого стоит отметить появление таких слов как *пароход* (739 →→ 473) и *паровоз* (→ 1367) в третьей зоне. Очевидно, паровые двигатели получили широкое распространение ближе к началу двадцатого века в России.

Люди

Одной из самых больших групп у имен существительных оказалась семантическая группа «Люди», разделенная на несколько подгрупп: «Люди», «Семья», «Дружба», «Статус» и «Объединения/организации». Наибольший интерес представляют две последние подгруппы (Таблица 9):

Таблица 9 – «Статус» все зоны

Статус								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
156	ГОСПОДИН	161	148	ГОСПОДИН	127	206	БАРИН	131
244	ХОЗЯИН	121	168	ГОСТЬ	115	247	ДОКТОР	115
395	ДАМА	81	210	ПОЭТ	99	291	ХОЗЯИН	101
540	БАРЫШНЯ	62	259	ДАМА	85	310	КОМИССАР	96
702	БАРЫНЯ	49	313	ГУБЕРНАТОР	75	401	ГРАЖДАНИН	82
			327	ХОЗЯИН	73	427	НАЧАЛЬНИК	76
			399	ПАН	62	450	ПРЕДСЕДАТЕЛЬ	74
			401	ПРЕДСЕДАТЕЛЬ	62	538	ГОСТЬ	62
			472	ДОКТОР	53	549	ВРАГ	61
			488	КУПЕЦ	51	581	ПАССАЖИР	57
			510	МАТРОС	49	654	ДАМА	51
			517	БАРЫШНЯ	48	657	СОСЕД	51
			539	НАЧАЛЬНИК	46	683	ВОР	49
			559	ДИРЕКТОР	44	703	ЛАВОЧНИК	48
			639	ИЗВОЗЧИК	39	719	КОММУНИСТ	47
			657	ЧИНОВНИК	39	723	ТОКАРЬ	47
			681	ВОР	37	736	МАТРОС	46
			703	БАРИН	36	776	ГОСПОДИН	44
			713	ГРАЖДАНИН	36	823	БАРЫШНЯ	42
			726	СОСЕД	36	847	БОЛЬШЕВИК	41
			767	КАЗНАЧЕЙ	34	850	КОРОЛЬ	41
			848	ВРАГ	31	871	АГЕНТ	40
			963	РАЗБОЙНИК	28	883	СЕКРЕТАРЬ	40
			1052	НАЧАЛЬСТВО	26	944	КОНТОРЩИК	38
			1056	ОХОТНИК	26	975	КОРОЛЕВА	37
			1078	ХОЗЯЙКА	26	979	ПАН	37
			1101	КОММУНИСТ	25	982	РАБОТНИК	37
			1109	НЕВЕСТА	25	1038	ШАХ	35
			1200	ГУВЕРНАНТКА	23	1099	ИНЖЕНЕР	33
						1190	КОМСОМОЛЕЦ	31
						1257	МИЛИЦИОНЕР	30

Здесь наиболее явная и выделяющаяся тенденция наблюдается в зародыше во второй зоне и полностью раскрывается в третьей. Речь идет о лексике, сопряженной с советским периодом, во время которого, язык претерпевал наиболее активные изменения, избавившись от всего «контрреволюционного» и внеся множество новых слов, являющихся олицетворением советской картины мира. *Коммунист*, имевший в военном периоде ранг 1101, в третьей зоне получил 719-й ранг. Появились такие слова как *комиссар* (310), *большевик* (847), *комсомолец* (1190), *милиционер* (1257). Интересно отметить, что слово *барин* (702 → 206), как можно видеть, переместилось с 702 места по частотности на 206-е. Возможно, произошло это потому, что после революции резкую популярность приобрели рассказы о крестьянском быте, где низшее сословие в лице крестьян зачастую обращается к сословию высшему не иначе как «барин». В совокупности стоит посмотреть и на другую упомянутую семантическую группу (Таблица 10):

Таблица 10 – «Объединения/организации» II и III зоны

Объединения/организации					
Вторая зона			**Третья зона**		
Ранг	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***
99	ТОЛПА	159	173	НАРОД	150
204	НАРОД	100	403	ТОЛПА	82
741	КОМАНДА	35	597	СОБРАНИЕ	55
881	ПАРТИЯ	30	738	ПАРТИЯ	46
914	КОМИТЕТ	29	778	КЛУБ	44
1102	КОМПАНИЯ	25	783	СЕЛЬСОВЕТ	44
			887	ЧЛЕН	40
			1012	ШТАБ	36
			1034	РОД	35
			1070	ПУБЛИКА	34
			1101	КОМАНДА	33
			1267	СОЮЗ	30

Причина, по которой в данной таблице нет первой зоны заключается в том, что данная семантическая группа не была в ней выделена и появилась только во втором периоде. Можно заметить интересные тенденции, также связанные с советским периодом: слово *народ* в третьей зоне оказалось частотнее, чем *толпа*,

которая во второй была более часто употребляемым словом, при этом ранг сильно понизился: *толпа* (99 → 403), а у слова *народ* (204 → 173), наоборот, хоть и немного, но повысился. При этом появились такие слова, как *собрание* (597), *сельсовет* (783), *штаб* (1012) и *союз* (1267), которые явственным образом относятся к лексике советского периода.

Общие тенденции

Несомненно стоит остановиться на группах, которые так или иначе связаны с выражением чувств или эмоций. В данной работе имело место разделение на две группы – положительная и отрицательная коннотации, которые были выявлены в трех частях речи – именах существительных, глаголах и именах прилагательных. Стоит отметить, что в глаголах данные группы получили другое название: «Положительные/Отрицательные эмоции и их выражение», так как в данном случае было важно обозначить то, как с помощью действий можно выражать, что лежит на душе (см. Рисунок 6 и Рисунок 7).

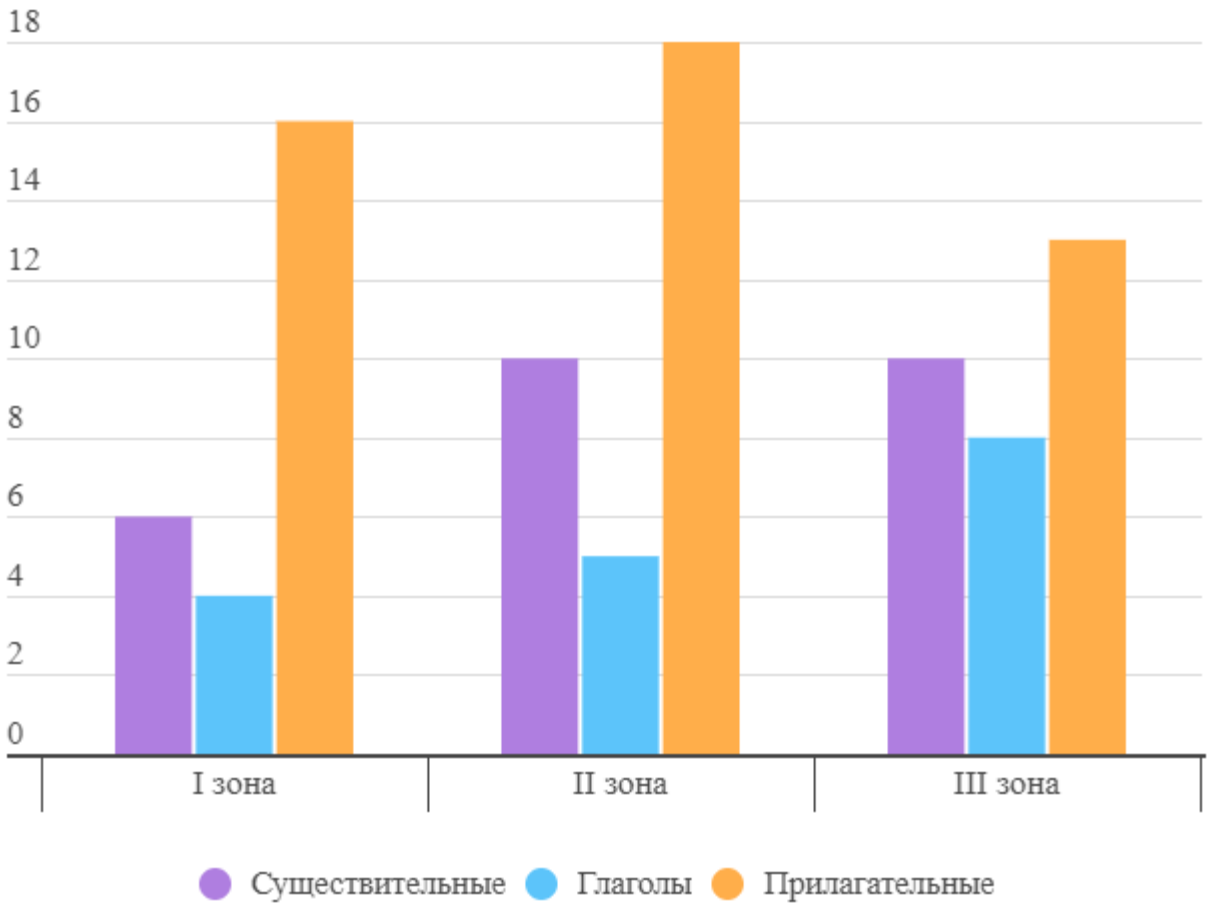

Рисунок 6 – Положительная коннотация существительных, глаголов, прилагательных (все зоны)

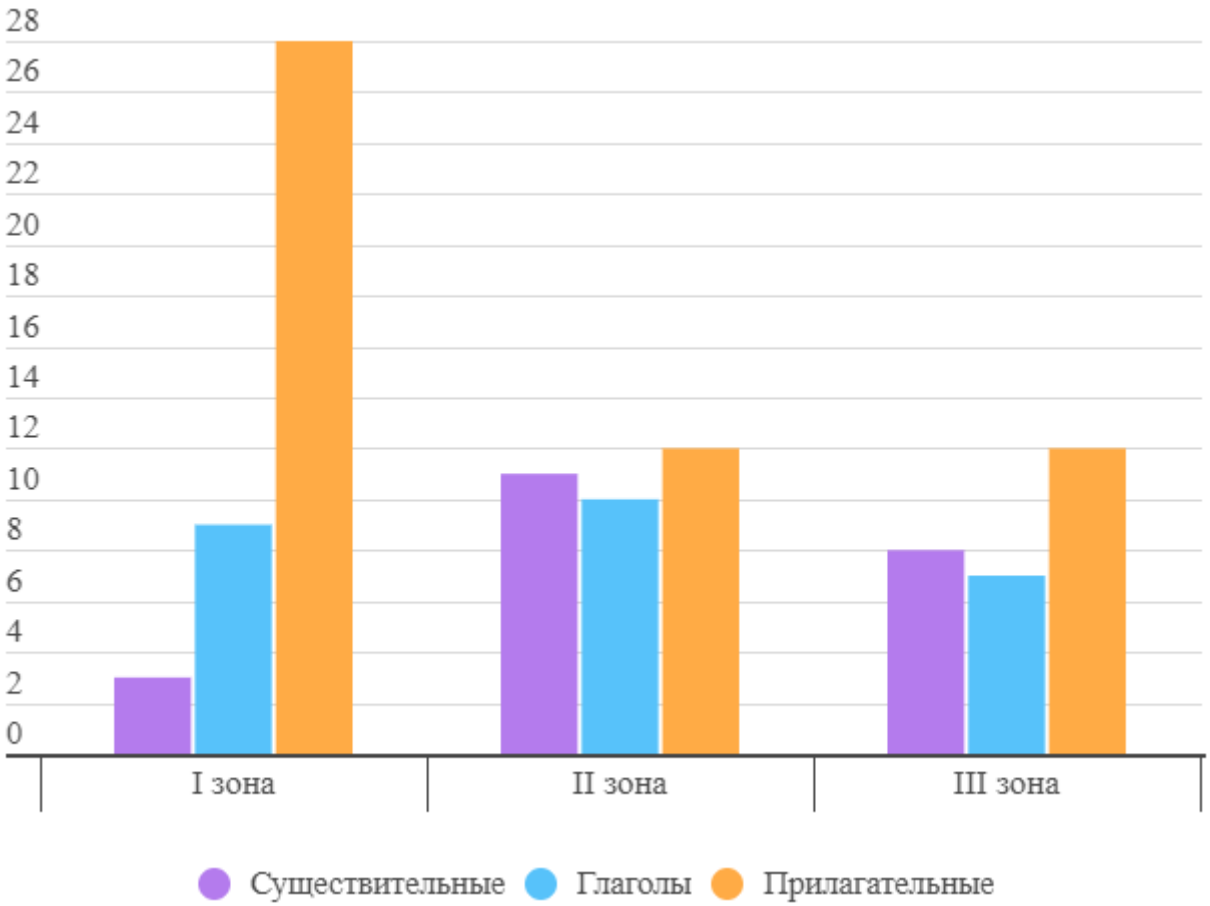

Рисунок 7 – Негативная коннотация существительных, глаголов, прилагательных (все зоны)

Очевидно, что лучше всего с выражением эмоций и чувств справляются прилагательные, что можно отметить на обоих графиках. Можно увидеть, что в первой зоне, содержащей слова, которые выражают нечто отрицательное, прилагательные достигают довольно большого количества – порядка 28, что является наибольшим числом как для положительной коннотации, так и для отрицательной среди всех приведенных на графике частей речи. Возможно, это связано по большей части с настроениями народа, так как известно, что первая революция началась именно в период с 1905 по 1907, то ожидаемо, что рассказы того времени будут пропитаны идеями революции, а с ними, соответственно, и недовольством общества, которое и вылилось в революционные действия против власти. Однако, когда пришло время встречаться с общим врагом во время Первой мировой войны, сплоченность людей стала гораздо выше, отрицательных эмоций все еще оставалось много, но самый их пик пошел на спад, в то время как положительных чувств и эмоций в последствии стало

больше – из-за завершения войны, достижения народом своей цели, – свержения правительства, и появления СССР.

Слова, отвечающие за радостные эмоции, обладают довольно выраженной тенденцией: *радость* (240 → 202 → 204), *счастье* (246 → 542 → 367), *праздник* (296 → 1012 → 620), *смеяться* (141 → 189 → 202), *улыбаться* (185 → 217 → 318), *хороший* (159 → 126 → 180), *веселый* (163 → 263 → 246), *добрый* (293 → 284 → 480), *счастливый* (458 → 599 → 615). Можно увидеть, что до всех военных событий, хоть слов с положительной коннотацией и немного, но все же они занимают более высокий ранг в первой зоне, чем в любой другой. После 1913 года большая часть подобных слов получает довольно низкие ранги по сравнению с предыдущим периодом, хотя есть и такие, которые наоборот становятся более частотными, быть может потому, что именно их не хватает в трудный как физически, так и эмоционально военный период. Некоторые слова, сильно потерявшие в ранге во второй зоне, поднимаются в частотности в третьем периоде, однако результаты все равно гораздо ниже по сравнению с рангами, которые принадлежали данным словам в первой зоне. Можно сказать, что хоть появилось гораздо больше причин для того, чтобы испытывать положительные эмоции после войны, тем не менее чувства для народа в послевоенный, советский период оказались отодвинуты на второй план.

Это видно и по словам с отрицательной коннотацией, например, таким как: *боль* (→ 321 → 470), *страх* (→ 469 → 484), *ужас* (→ 447 → 522), *тоска* (→ 380 → 530), *бояться* (152 → 91 → 156), *плакать* (184 → 159 → 249), *злой* (421 → 333 → 494). Хотя и видно, что многие слова появились в первый раз во второй зоне, а часть оказались более частотны во втором, чем в первом периоде, в третьем они стали использоваться меньше.

Интересно обратить внимание на появившуюся в третьей зоне семантическую группу «Смерть». До третьего периода ее было несостоятельно выделять, так что она находилась в составе группы «Жизнь». Однако после 1922 года она стала более явственной и очевидной для ее отделения (Таблица 11).

Таблица 11 – Семантическая группа «Смерть» (III зона)

Смерть		
Третья зона		
Ранг	***Слово***	***Частота***
237	СМЕРТЬ	120
442	УТОПЛЕННИК	75
953	ТРУП	38
1029	КЛАДБИЩЕ	35
1060	ГРОБ	34
1104	МОГИЛА	33

Во время таких событий, какие происходили за первые три десятка лет начала двадцатого века, человек часто сталкивается со смертью. Причем, это не обязательно должна быть насильственная смерть, она может быть и естественной, однако зачастую естественный уход из жизни на фоне военных действий кажется крайне несвоевременным или даже в какой-то мере неправильным явлением, которое не должно происходить в такой период. На войне и во время революции надо уметь быть готовым к смерти, не страшиться ее и пытаться избегать ее как можно дольше, оттого и не удивительно, что данная тема наконец привлекла больше внимания писателей того времени.

Одни из самых больших семантических групп (то есть тех, которые состоят из 40 и больше слов) приходятся на глаголы. Среди них «Движение», «Речь/общение», «Действия человека», «Действия над кем-то/чем-то» (Таблица 12). В таких группах происходят малые изменения на первых рангах – чаще всего они либо остаются такими же, либо слова порой меняются местами, но первые 10 слов обычно примерно одинаковы во всех зонах. Большее расхождение идет на более низких рангах, причем порой оно может быть довольно критичным.

Таблица 12 – 10 слов самых больших семантических групп глаголов

Движение								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
29	ПОЙТИ	408	7	ИДТИ	569	8	ИДТИ	715
62	УЙТИ	269	27	ПОЙТИ	314	16	ПОЙТИ	475
63	ВЫЙТИ	267	53	ВЫЙТИ	223	45	ВЫЙТИ	316
77	ХОДИТЬ	246	69	ХОДИТЬ	197	63	УЙТИ	269
135	ОСТАНОВИТЬСЯ	180	86	ПРИЙТИ	169	78	ХОДИТЬ	237
147	ПРИЙТИ	174	90	УЙТИ	166	86	БЕЖАТЬ	226
149	ПОДОЙТИ	173	129	ПОДОЙТИ	138	134	ПОДОЙТИ	187
162	ПРОЙТИ	159	133	БЕЖАТЬ	133	139	ОСТАНОВИТЬСЯ	179
167	ВОЙТИ	155	154	УХОДИТЬ	122	161	УХОДИТЬ	159
180	БЕЖАТЬ	151	158	ОСТАНОВИТЬСЯ	119	164	ПРИЙТИ	155
Речь/общение								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
1	СКАЗАТЬ	1222	1	ГОВОРИТЬ	962	3	СКАЗАТЬ	1080
3	ГОВОРИТЬ	1135	2	СКАЗАТЬ	846	4	ГОВОРИТЬ	979
45	СПРОСИТЬ	310	44	СПРОСИТЬ	243	58	СПРОСИТЬ	283
51	МОЛЧАТЬ	294	80	МОЛЧАТЬ	183	61	МОЛЧАТЬ	276
99	ОТВЕТИТЬ	210	120	ОТВЕТИТЬ	144	96	КРИЧАТЬ	216
113	КРИЧАТЬ	199	128	КРИЧАТЬ	138	113	ОТВЕТИТЬ	200
158	ОТВЕЧАТЬ	161	163	ПРОСИТЬ	118	191	КРИКНУТЬ	139
178	КРИКНУТЬ	152	184	ЗВАТЬ	109	201	ПРОСИТЬ	134
202	СПРАШИВАТЬ	140	194	СПРАШИВАТЬ	105	227	ЗВАТЬ	122
301	ПЕТЬ	102	196	ПЕТЬ	103	250	РАССКАЗЫВАТЬ	114
Производить действие над кем-то\чем-то								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***	***Ранг***	***Слово***	***Част.***
70	ВЗЯТЬ	257	47	ВЗЯТЬ	234	38	ВЗЯТЬ	339
233	БРОСИТЬ	126	137	НАЙТИ	132	138	НАЙТИ	179
355	ПОЛУЧИТЬ	88	143	ПОДНЯТЬ	129	147	БРОСИТЬ	171
376	ОТКРЫТЬ	84	231	БРОСИТЬ	92	176	ИМЕТЬ	149
405	СХВАТИТЬ	80	265	ДЕРЖАТЬ	84	192	ПОДНЯТЬ	139
415	ПОЛОЖИТЬ	78	301	ИМЕТЬ	77	242	ДЕРЖАТЬ	117
475	ПОКАЗАТЬ	69	348	ОСТАВИТЬ	68	294	ИСКАТЬ	100
529	ПРИНЯТЬ	64	357	ПОСТАВИТЬ	67	317	ПОЛУЧИТЬ	94
536	САДИТЬ	63	359	СНЯТЬ	67	345	ОСТАВИТЬ	89
557	НАЗЫВАТЬ	61	368	ОТКРЫТЬ	65	346	СНЯТЬ	89

Стоит отметить, что в семантической группе «Действия человека» происходят некоторые интересные изменения и в первых по рангу словах. Так, например, глагол *работать* перешел с 342 ранга второго периода на 110 в третьей зоне. Возможно произошло это потому, что приоритет и направленность идеологии советского времени была на рабочих и, соответственно, работу.

Слова *писать* и *читать* также показывают занимательную тенденцию. Например, с первой зоны *писать* (309 → 187) стало гораздо частотнее, однако затем получило ранг 213. Аналогичное с ним *читать* (291 → 203) с первой по вторую зону, а затем заняло 214-й ранг в третьей. Предполагается, что во время Первой мировой войны многие писали письма на фронт и, конечно же, читали письма с фронта, в том числе новости, таким образом, объясняется такой подъем частотности для этих глаголов в период с 1914 по 1922.

Общее наблюдение семантической группы «Органы чувств» наглядно показало, что в первую очередь человек полагается на свое зрение (так как мы знаем, что именно таким образом воспринимается большая часть информации), а только потом уже на слух.

Таблица 13 – Органы чувств (все зоны)

Органы чувств								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***
20	ВИДЕТЬ	496	26	ВИДЕТЬ	316	19	ВИДЕТЬ	447
23	СМОТРЕТЬ	469	33	СМОТРЕТЬ	275	28	СМОТРЕТЬ	394
90	СЛЫШАТЬ	223	77	ГЛЯДЕТЬ	184	108	УВИДЕТЬ	205
106	ГЛЯДЕТЬ	203	87	УВИДЕТЬ	168	120	СЛУШАТЬ	196
117	СЛУШАТЬ	195	89	СЛУШАТЬ	166	127	ГЛЯДЕТЬ	192

Группы «Классовая характеристика» и «Принадлежность к группе/представителю группы» особенно ясно среди прилагательных показывают изменения социального строя в России за первые тридцать лет двадцатого века. Для слова *рабочий* (288 → 114) ранг стал выше в третий период, советский, несомненно под влиянием нового идеологического направления. Помимо этого в третьем периоде появились такие слова как *советский* (651),

бабий (1020) и *мужицкий* (1353), разграничивающие предреволюционное крестьянское и постреволюционное советское.

Из группы «Состояние предмета» слово *дорого* постепенно повысилось в ранге (148 → 131 → 98). Несомненно, страна была не в лучшем состоянии после Первой мировой войны и революций, а потому многое, что ранее было не столь дорогим, поднялось в цене.

Семантическая группа «Национальность» в прилагательных состоит только из одного слова *русский*. Однако эта группа также показывает определенную любопытную тенденцию: *русский* (342 → 188 → 415). Можно увидеть, что принадлежность к России стала иметь огромное значение во время войны, тем самым влияя на частоту использования слова. При этом в послевоенное время данный фактор перестал играть настолько большую роль, и можно сделать вывод, что патриотизм как никогда раскрывается именно во время военных событий.

В наречиях большую частотность получили слова, относящиеся к теме «Причина» - *зачем* (215 → 179), *почему* (252 → 157), причем к третьему периоду эти вопросы стали использоваться все чаще. Любопытно, что ответ *потому* (→ 102) впервые появился в третьей зоне, возможно от того, что у многих людей не было ответов на многие вопросы, которые могли быть заданы в тот период.

Таблица 14 – Группа «Время» у наречий (все зоны)

Время								
Первая зона			**Вторая зона**			**Третья зона**		
Ранг	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***	***Ранг***	***Слово***	***Частота***
9	ТЕПЕРЬ	683	17	ТЕПЕРЬ	380	23	ТЕПЕРЬ	426
14	ВДРУГ	559	32	ВДРУГ	283	30	ОПЯТЬ	384
64	СЕЙЧАС	267	67	СЕЙЧАС	199	35	ВДРУГ	364
69	ВСЕГДА	261	81	ВСЕГДА	178	41	ТОГДА	326

Здесь можно увидеть, что набор слов для первых двух зон примерно одинаков, в то время как в третьей есть некоторые расхождения, так как появились такие слова как *опять* (→ 30) и *тогда* (→ 41). Обычно эти слова используются, если человек сравнивает один временной период в другим или

сетует на то, что происходило в прошедшем и происходит снова. Таким образом, возможно эти два слова в первых рангах появились потому, что как народ, так и писатели пытались осознать масштаб изменений, привнесенных революцией, сравнивали жизнь прошлую, жизнь нынешнюю и жизнь грядущую.

Изменения в числительных оказались незначительны. Можно увидеть, что их количество постепенно увеличивалось, ровно как и частотность: *два* (24 → 15 → 12), *три* (100 → 94 → 69), *сколько* (304 → 233 → 220).

Рисунок 8 – Количество числительных во всех трех зонах

Выделение верхних зон

Для выделения верхних зон полученных словарей была использована предложенная Г.Я. Мартыненко величина рангового среднего, вычисляемая по формуле:

$$r = \sum r\, f_r / N,$$

где r - ранг, f_r - соответствующая этому рангу частота, N - объем выборки (Sherstinova T., Martynenko G., 2020). Ранговое среднее выступает в качестве меры, отделяющей зону концентрации от зоны рассеивания.

Для полученных частотных словарей значение рангового среднее округленно составило: 1-ый период – 33 (1407 лексем), 2-ой период – 23 (1420 лексем), 3-ий период – 27 (1694 лексемы). Из полученных верхних зон были проанализированы знаменательные части речи (существительные, прилагательные, глаголы, числительные и наречия).

IPM и сравнение с НКРЯ

Последний этап данного исследования – это сравнение лексики первых трех десятилетий начала двадцатого века с лексикой современных рассказов, то есть рассказов первых двух десятилетий начала двадцать первого века. Лексика 1900-1930 гг. взята из Корпуса русских рассказов, точнее из частотных списков, выделенных из Корпуса и разделенных на периоды 1900-1913 гг., 1914-1922 гг., 1923-1930 гг., а для лексики 2000-2022 гг. используется Национальный корпус русского языка (НКРЯ). Объем частотных списков составил 124 081 лексема, 1 077 970 словоформ. Были отобраны лексемы, которые входят в верхнюю зону частотного распределения – около 1000 слов с частеречной разметкой, из которой были выделены существительные, глаголы, прилагательные, наречия и числительные.

Что касается НКРЯ, то с помощью его инструментария был задан подкорпус русских рассказов в период с 2000-2022 год. Подкорпус включал в себя 939 документов, общий объем словоформ оказался 4 016 064 слова.

Так как объемы данных существенно различаются, при сравнении было принято решение использовать параметр относительной частоты – ipm (instances per million words), то есть число употреблений слова на миллион слов.

Из-за того, что лемматизация в некоторых случаях работала не очень успешно (например, лемма *домыть* была интерпретирована как глагол, однако на самом деле это было существительное *домой*), а также наблюдались проблемы, такие как омонимия, данный этап исследования носит выборочный характер, а случаи, подобные приведенным выше, исключены из рассмотрения.

Крестьянский быт и Транспорт

Одной из самых ожидаемых закономерностей являлось снижение частотности у слов, связанных с крестьянским бытом. Как результат идеологий советского периода, а также развития технологий, современный мир давно отошел от крестьянства и сельской жизни. Подобную тенденцию также можно отметить и у транспорта – повышение частотности у слова *машина* почти в два раза, но при этом *лошадь*, которая долгое время использовалась как стандартное и распространенное средство передвижения, стала использоваться гораздо реже, тем более в контексте транспорта (Таблица 15).

Таблица 15 – Лексика крестьянского быта и транспорта

Слово	**1900-1930, IPM**	**2000-2022, IPM**
МУЖИК	701,56	254,48
БАБА	594,63	187,50
СЕЛО	542,47	255,47
ИЗБА	451,19	54,78
ДЕРЕВНЯ	419,89	198,70
ПОЛЕ	297,32	245,02
КРЕСТЬЯНИН	161,70	25,40
ВАГОН	375,56	132,72
МАШИНА	292,10	473,85
ПОЕЗД	284,28	169,82
ПУТЬ	268,63	289,09
ВОКЗАЛ	106,93	86,15
АВТОМОБИЛЬ	91,28	86,65
ЛОШАДЬ	565,94	101,09

Время

Таблица 16 – Лексика, обозначающая время

Слово	1900-1930, IPM	2000-2022, IPM
ДЕНЬ	1734,35	1532,35
ВРЕМЯ	1126,67	1715,11
ЧАС	641,58	592,12
ГОД	625,93	1791,56
ПРОШЛЫЙ	174,74	212,15
НАСТОЯЩИЙ	156,48	285,11

Привлекающую внимание тенденцию можно заметить в контексте времени. Частотность у слова *день* и *час* упали, в то время как *время* и *год* наоборот стали более частотны. Возможно, эта закономерность заключается в том, что сейчас принято узнавать, есть ли у человека свободное время (к тому же, чаще можно услышать вопрос «Сколько времени?» чем «Который час?»), или же вспоминать о событиях, которые произошли в определенный год. Критерии сравнения, такие как *прошлый* и *настоящий*, тоже стали более активно использоваться, помимо этого больше внимания стало уделяться настоящему, а не тому, что имело место в прошлом.

Абстрактные существительные

В данной работе была выделена семантическая группа абстрактных существительных, которая в свою очередь показывает довольно интересные результаты.

Таблица 17 – Лексика, связанная с абстрактными существительными

Слово	**1900-1930, IPM**	**2000-2022, IPM**
СИЛА	576,38	498,75
ДУША	417,29	476,09
СОН	375,56	371,76
ВЛАСТЬ	258,20	113,05
ДУХ	192,99	188,49
ПОРЯДОК	172,13	161,10
ВЕРА	159,09	184,01
ПРОСТРАНСТВО	130,40	118,77
ВОЛЯ	125,19	110,31
ПРАВДА	96,50	458,41
ОТНОШЕНИЕ	91,28	186,00
ПУСТОТА	83,46	79,68
ГОЛОД	73,03	43,33

Такие могущественные явления, как *сила*, *власть*, *порядок*, *воля* стали играть гораздо меньшую роль в рассказах двадцать первого века. Возможно, по сравнению с другими темами, раскрытыми в данной работе, они более частотны, однако по сравнению с постреволюционным периодом важность этих слов для современного человека стала меньше. Тем не менее религиозная тематика, отображенная в словах *душа* и *вера* наоборот стала гораздо популярнее, что неудивительно, учитывая жесткую политику в отношении религии в советский период. Помимо этого гораздо больше в наше время стала цениться *правда*, а такая проблема как *голод* отошла на задний план.

Религия

Выше уже было упомянуто, что религия стала играть более важную роль в наши дни, чем в советский период, что имеет под собой веские основания. Однако стоит отметить, что по полученным в результате исследования данным, это не так однозначно. Несомненно роль *бога* выросла, ровно как и отношение к святости (*святой*) и настолько же сильно ослабла боязнь богохульства, что мы можем видеть по сильно увеличенной частотности использования слова *черт* (скорее всего как ругательства). К тому же роль *церкви* и его атрибутов (*икона*, *крест*), а также роль *греха* судя по всему уменьшилось, а также и важная часть большинства религиозных обрядов – процесс молитвы (*молиться*). Таким образом можно сделать вывод, что хотя роль религии в целом возросла по сравнению с советским периодом, однако отчасти возможно именно из-за него произошло отступление от религиозных обрядов и уменьшение их значимости.

Таблица 18 – Лексика религии и веры

Слово	**1900-1930, IPM**	**2000-2022, IPM**
БОГ	323,40	485,30
ЦЕРКОВЬ	117,36	107,57
ГРЕХ	93,89	70,97
ИКОНА	88,67	50,80
ЧЕРТ	83,46	221,61
КРЕСТ	80,85	70,22
СВЯТОЙ	88,67	97,11
МОЛИТЬСЯ	78,24	41,58

Чувства/выражение чувств и положительная/отрицательная коннотация

Таблица 19 – Лексика чувств и положительной/отрицательной коннотации

Слово	**1900-1930, IPM**	**2000-2022, IPM**
СЛЕЗА	362,52	227,59
СМЕХ	299,92	116,78
УЛЫБКА	258,20	196,46
ЛЮБОВЬ	245,16	392,18
ЧУВСТВО	151,27	264,69
ОБИДА	99,11	60,76
ЖЕЛАНИЕ	83,46	135,21
РАДОСТЬ	344,26	173,55
СМЕЯТЬСЯ	346,87	238,54
СЧАСТЬЕ	224,29	217,38
ХОРОШИЙ	380,77	1480,80
ВЕСЕЛЫЙ	299,92	220,12
МИЛЫЙ	213,86	161,10
ДОБРЫЙ	182,56	225,10
СЧАСТЛИВЫЙ	138,23	206,42
ИГРА	182,56	177,54
ПРАЗДНИК	138,23	117,78
ЗДОРОВЬЕ	88,67	67,23
БОЛЬ	185,17	177,29
СТРАХ	182,56	180,28
УЖАС	166,91	147,91
ТОСКА	164,31	85,16
БОЛЕЗНЬ	161,70	78,44
ТРЕВОГА	109,54	46,31
ЗЛОБА	104,32	25,15
БЕДА	101,71	78,93
БОЯТЬСЯ	425,11	383,71
ПЛАКАТЬ	297,32	250,99
РУГАТЬСЯ	117,36	41,83
СТРАШНЫЙ	297,32	375,74
СТРАННЫЙ	182,56	374,50
ЗЛОЙ	177,35	116,53
ПОШЛЫЙ	161,70	174,80
ХОТЕТЬ	1137,10	1984,78
ЛЮБИТЬ	670,27	883,95
ЧУВСТВОВАТЬ	339,04	341,88
ВЕСЕЛО	169,52	100,10
СПОКОЙНО	169,52	156,62
СТРАШНО	166,91	191,23
РАДОСТНО	125,19	68,72

Можно увидеть, что выражение эмоций (*слеза, смех, улыбка, смеяться, бояться, плакать, ругаться*) стало менее свободно использоваться. Если раньше

проявлять свои эмоции было силой, то теперь многие могут посчитать это за слабость и таким образом теперь люди стараются держать дистанцию и минимально проявлять какие-либо эмоции и чувства. Хотя при этом можно увидеть, что потребность в *любви, чувствах* стала гораздо сильнее. Люди хотят видеть вокруг себя больше *хороших, добрых* и *счастливых* людей, так необходимость в том, чтобы человек был *хорошим*, возросла даже больше, чем в десять раз в отличие от двадцатых годов двадцатого века. *Праздники* и *здоровье* стали менее обсуждаемы и быть может менее важны. В то же время существительные с отрицательной коннотацией также стали менее злободневны. Однако боязнь *странного* и *страшного* и в том числе *пошлого* наоборот выросла, а веселье (*весело*) и радость (*радостно*) уменьшилась. Помимо этого роль *желания*, того, что человек что-то *хочет*, также очень возросла.

Бытовые аспекты жизни

Таблица 19 – Лексика движения

Слово	1900-1930, IPM	2000-2022, IPM
ИДТИ	1864,75	1255,21
ПОЙТИ	1238,82	926,28
БЕЖАТЬ	589,42	260,70
ОСТАНОВИТЬСЯ	466,84	197,96

Стоит обратить внимание на глаголы движения (Таблица 19). Занимательная закономерность прослеживается у глаголов, отвечающих за перемещение пешком в любых его проявлениях, будь то бег (*бежать*) или простая ходьба (*идти, пойти*). Снижение частотности таких глаголов тесно связано с развитием транспорта – возможностей покрывать как большие, так и малые дистанции за меньшее количество времени с использованием определенного средства передвижения становится гораздо больше.

Таблица 20 – Лексика различных напитков

Слово	1900-1930, IPM	2000-2022, IPM
ВОДА	774,59	750,74
ЧАЙ	106,93	218,37
КОФЕ	78,24	139,19
ПИВО	73,03	118,52

Можно отметить, что хоть доступность *воды* сделала ее менее частотной, при этом такие напитки как *чай*, *кофе* и *пиво* стали гораздо более популярны, а вместе с тем стали чаще использоваться в рассказах, равно как и повседневной жизни.

Таблица 21 – Лексика денежных средств

Слово	1900-1930, IPM	2000-2022, IPM
ДЕНЬГА	292,10	1,74
ДЕНЬГИ	96,50	530,62

Деньги в свою очередь стали играть большую роль в современном мире. При этом слово *деньга* совсем уже не используется (скорее потому, что данного номинала больше не встретишь в обиходе).

Работа, являющаяся в нынешние дни одним из самых необходимых аспектов бытовой жизни, также показывает некоторые занимательные тенденции:

Таблица 22 – Лексика работы и места работы

Слово	1900-1930, IPM	2000-2022, IPM
ДЕЛО	1251,86	1161,34
РАБОТА	795,45	602,83
БИБЛИОТЕКА	213,86	52,79
ТРУД	211,25	166,33
ФАБРИКА	208,64	37,85
ЗАВОД	174,74	78,19
РАБОЧИЙ	519,00	115,04

Так, хоть *работа* и является одной из важной составляющей жизни, можно увидеть, что ее значимость тем не менее уменьшилась. Скорее всего, произошло это из-за того, что сменилась идеология – в советском периоде ядром счастья и равенства была работа, быть рабочим было важно, в то время как в двадцать первом веке работа для многих – это лишь способ заработка. От ранее почетной, а теперь в какой-то степени «зазорной» работе на *заводах* и *фабриках* стараются уклониться, предпочитая офисы, фриланс и другие места работы. К тому же по классовой характеристике (*рабочий*) можно также отметить спад частотности. Физический *труд* стал оцениваться меньше, чем умственный, а *библиотеки* с появлением и развитием интернета и онлайн-библиотек стали еще менее популярны.

Таблица 23 – Лексика умственной деятельности

Слово	1900-1930, IPM	2000-2022, IPM
ЗНАТЬ	1877,79	2172,77
ДУМАТЬ	1197,09	1060,74
ПОНЯТЬ	555,51	820,21
ПОДУМАТЬ	503,35	527,63
ПОНИМАТЬ	492,92	697,95
ВСПОМНИТЬ	320,79	329,43

Помимо всего прочего с развитием технологий, появлением компьютеров и, конечно же, интернета, бумажные носители информации такие как *книги*, *газеты*, *письма* и т.д. стали гораздо меньше использоваться в бытовой жизни:

Таблица 24 – Лексика, касающаяся бумажных носителей

Слово	1900-1930, IPM	2000-2022, IPM
КНИГА	326,00	315,48
ПИСЬМО	255,59	215,14
ГАЗЕТА	192,99	156,12

Однако интернет и электронные девайсы повлияли не только на бумажные носители, но также и на общение в целом:

Таблица 25 – Лексика общения и речи

Слово	1900-1930, IPM	2000-2022, IPM
СКАЗАТЬ	2816,68	3191,68
ГОВОРИТЬ	2553,27	2302,01
СПРОСИТЬ	738,07	826,43
МОЛЧАТЬ	719,82	383,21
КРИЧАТЬ	563,34	278,38
ОТВЕТИТЬ	521,61	471,61
КРИКНУТЬ	362,52	117,78
ПРОСИТЬ	349,48	237,05
ЗВАТЬ	318,18	303,28
РАССКАЗЫВАТЬ	297,32	284,61
СПРАШИВАТЬ	271,24	326,44
ПЕТЬ	266,02	207,92
ЗАКРИЧАТЬ	258,20	106,32
ОТВЕЧАТЬ	252,98	303,28
РАССКАЗАТЬ	224,29	257,22
РАЗГОВАРИВАТЬ	161,70	125,75
НАЗЫВАТЬ	153,87	222,36
РАЗГОВОР	326,00	285,60
ВОПРОС	234,72	381,47
ОТВЕТ	190,39	242,03

Очевидно, что глаголы связанные с устным общением стали использоваться реже. Частотность повысилась и у глаголов, и у существительных только в контексте вопроса-ответа, то есть люди стали задаваться вопросами (*спросить, спрашивать, вопрос*) и искать на них ответы (*ответ, отвечать*) в большей мере, однако все остальное вербальное взаимодействие в расцвет онлайн-переписок перестало быть настолько необходимым, как раньше.

В семантической группе семьи наблюдается традиционный уклад семьи в России, сохранившийся до сих пор:

Таблица 26 – Лексика семьи

Слово	1900-1930, IPM	2000-2022, IPM
ОТЕЦ	623,32	715,88
МАТЬ	552,90	346,11
ДЕД	516,39	244,02
БРАТ	503,35	271,91
СЫН	495,53	400,14
ЖЕНА	448,58	692,97
МУЖ	297,32	507,21

Особое внимание стоит обратить на слова *отец*, *мать*, *жена* и *муж*. Роль *отца* стала гораздо значимее, чем роль *матери*, которая напротив стала менее частотной. При этом выросла роль *жены*, так как про *жену* можно говорить лишь в контексте мужчины (*жена Павла, жена начальника*). Примерно такой же частотный рост можно видеть и у слова *муж*. При этом можно отметить общую тенденцию повышенного внимания к браку как таковому (рост частотности слов *муж* и *жена*) и при этом сомнительное желание заводить не то, что детей, а даже одного ребенка (падения частотности у слова *сын*, малая частотность слова *брат* и *дед*).

Еще одна из занимательных тенденций, казалось бы, звучит парадоксально, но при этом наглядно олицетворяет современный русский менталитет. На мировой арене граждане России крайне сплочены, существует четкое разделение России и русских как страны и нации и других. Однако в самой стране нет такого единения в небольших социальных группах:

Таблица 27 – Лексика объединения/организаций и национальности

Слово	1900-1930, IPM	2000-2022, IPM
НАРОД	391,21	255,72
ТОЛПА	213,86	117,03
СОБРАНИЕ	143,44	41,83
ПАРТИЯ	119,97	63,00
РОССИЯ	88,67	164,84
СТРАНА	80,85	218,12
РУССКИЙ	206,03	384,95

Последняя закономерность, которая явно выделяется на фоне остальных – это отношение к природе. С периода активной урбанизации народ из сел, деревень и различных отдаленных мест стремился переехать в город. До сих пор

из маленьких провинций люди хотят ехать в большие города в поисках лучшей жизни. Так русский народ начал все сильнее отдаляться от природы, предпочитая ей город.

Таблица 28 – Лексика природы и неба

Слово	**1900-1930, IPM**	**2000-2022, IPM**
ЛЕСА	638,97	305,52
ВЕТЕР	516,39	272,16
ВОЗДУХ	451,19	361,05
БЕРЕГ	430,33	277,39
КАМЕНЬ	315,57	215,14
РЕКА	310,36	225,35
ДЕРЕВО	307,75	313,49
ТРАВА	294,71	221,11
ГОРА	292,10	246,01
КУСТ	284,28	117,78
МОРЕ	255,59	211,65
БОЛОТО	130,40	36,60
БЕРЕЗА	96,50	35,86
ЗЕМЛЯ	1097,98	622,75
СОЛНЦЕ	610,28	359,31
НЕБО	511,18	402,88
ЗВЕЗДА	247,76	148,90

Заключение

Данное исследование было направлено на изучение связи художественного языка в жанре рассказов с историческими событиями, которые могли тем или иным образом влиять на картину мира обывателя. Можно с уверенностью заключить, что такая связь несомненно существует, это было наглядно показано в данной работе. Помимо того, что удалось увидеть, как менялось отношение к таким темам, как религия, война, общество, и экзистенциальным вопросом жизни, времени и смерти, получилось сравнить лексику, использовавшуюся примерно сто лет назад, с лексикой современного русского языка и выявить определенные закономерности, являющиеся либо результатом прошлого, либо настоящего. Например, развитие различных технологий оказало огромное влияние на мир. Хоть это и довольно очевидный факт, порой из вида упускается глобальность некоторых явлений – то есть то, что они влияют не только на отдельные аспекты жизни, в которой появились, но и на общие, например, язык.

В ходе данной работы были выполнены следующие задачи:

1. На базе выборки из «Корпуса русского рассказа 1900-1930 гг.» были построены частотные словари его подкорпусов: довоенный (1900-1913), военно-революционный (1914-1922) и советский (1923-1930).
2. Статистически выделены и проанализированы верхние зоны частотного распределения лексем для данных подкорпусов.
3. В полученных верхних зонах проанализирована семантика знаменательных частей речи: существительных, прилагательных, глаголов, наречий, числительных.
4. На основании семантического анализа верхних зон частотных распределений лексем были выделены основные темы, которые главенствовали в обществе в данные периоды времени и нашли свое отражение в анализируемых текстах.

5. Были проведено сравнение частотностей входящих в выделенные тематические группы лексем с частотностями аналогичных лексем в современных русских рассказах с использованием материала НКРЯ.
6. Выявлены тенденции в изменении частотности отдельных слов и лексических групп от одного исторического отрезка к другому.
7. На основе полученных результатов были сделаны предположения о причинах появления подобных изменений и тенденций.

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 позволяет существенно дополнить анализ динамики тем произведений.

Список литературы

1. Аношкина Ж.Г. Подготовка частотных словарей и конкордансов на компьютере. // – М.: МГУ. – 1995.
2. Апресян Ю.Д. Лексическая семантика. // Синонимические средства языка. – Москва. – 1974. – С. 175-217.
3. Болдырев Н. Н. Когнитивная семантика. Введение в когнитивную лингвистику. – 2014.
4. Герд А. С. и др. Лексикография русского языка: учебник для высших учебных заведений Российской Федерации/Герд А. С., Ивашко Л. А., Лутовинова И. С. и др.; под ред. Поцепни Д. М./Учебно-методический комплекс по курсу «Лексикография русского языка». – 2013.
5. Гребенников А. О., Мартыненко Г. Я. Частотный словарь рассказов А.И. Куприна. – 2006.
6. Гребенников А. О., Мартыненко Г. Я. Частотный словарь рассказов И. А. Бунина. // Издательство Санкт-Петербургского университета. – 2012.
7. Гребенников А. О., Мартыненко Г. Я. Частотный словарь рассказов Л. Н. Андреева. – 2003.
8. Гребенников А. О. и др. База русского рассказа XIX-XX веков. Модели аппроксимации //Корпусная лингвистика-2019. – 2019. – С. 379-386.
9. Гребенников А. О. Индивидуально-авторский характер различных зон распределения в частотных словарях языка писателя // Структурная и прикладная лингвистика. – 2015. – №. 11. – С. 100-110.
10. Гребенников А. О., Марусенко Н. М. Корпус русского рассказа начала XX века. Пример лингвостатистического анализа // Компьютерная лингвистика и вычислительные онтологии. – 2020. – №. 4. – С. 21-28.
11. Гребенников А. О. О стилеразличительных возможностях частотных словарей языка писателя // Русский язык и литература в пространстве мировой культуры. – 2015. – С. 93-96.
12. Гребенников А. О., Скребцова Т. Г. Корпус русских рассказов (1900-1930). Устойчивость лингвостатистических характеристик. – 2021.

13. Гребенников А. О., Скребцова Т. Г. Языковая картина мира в русском рассказе начала XX века // Философия и гуманитарные науки в информационном обществе. – 2019. – №. 3. – С. 82-92.
14. Гребенников А. О. Частотный словарь и образ мира писателя //Словоупотребление и стиль писателя: межвуз. сб. СПб. – 2006. – №. 3.
15. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. – 2011.
16. Кобозева И. М. Лингвистическая семантика. – УРСС. – 2004.
17. Кузнецов О. П. Когнитивная семантика и искусственный интеллект //Искусственный интеллект и принятие решений. – 2012. – № 4. – С. 32-42.
18. Кукушкина О. В. и др. Частотный грамматико-семантический словарь языка художественных произведений А.П. Чехова с электронным приложением //М.: МАКС Пресс. – 2012.
19. Кураш С. Б. Лингвометафорология и корпусная лингвистика: зоны общих интересов. – 2015.
20. Мартыненко Г. Я. и др. 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. – 2018. – №. 2. – С. 97-102.
21. Мартыненко Г. Я. и др. О принципах создания корпуса русского рассказа первой трети XX века // ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ. – 2018. – С. 180-197.
22. Митрофанова О. А. Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика-2015». – Санкт-Петербург. – 2015. – С. 332-343.
23. Национальный корпус русского языка [https://ruscorpora.ru/] (дата обращения 24.05.2022)

24. Семантический словарь под общей редакцией Н.Ю. Шведовой [http://slovari.ru/default.aspx?s=0&p=235] (дата обращения 03.05.2022)
25. Толковый словарь русского языка под редакцией Д.Н. Ушакова [https://ushakovdictionary.ru/] (дата обращения 24.04.2022)
26. Толковый словарь С.И. Ожегова и Н.Ю. Шведовой [http://slovari.ru/default.aspx?s=0&p=244] (дата обращения 03.05.2022)
27. Ханинова Р. М. Антропологическая поэтика русской повести и рассказа 1900-1930-х гг. – Федеральное государственное бюджетное образовательное учреждение высшего образования "Калмыцкий государственный университет имени ББ Городовикова", 2013.
28. Brookes G., McEnery A. Corpus linguistics for indexing //The Indexer: The International Journal of Indexing. – 2019. – Vol. 37. – № 2. – P. 105-124.
29. Furimsky-Lackova M. The Russian short story and the development of the Slovak short story. – University of Ottawa (Canada). – 1981.
30. Goddard C., Wierzbicka A. (ed.). Semantic and lexical universals: Theory and empirical findings. – 1994.
31. Gries S. T. Useful statistics for corpus linguistics // A mosaic of corpus linguistics: Selected approaches. – 2010. – Vol. 66. – P. 269-291.
32. Grondelaers S., Speelman D., Geeraerts D. Lexical variation and change //The Oxford handbook of cognitive linguistics. – 2007.
33. Kovalev I. V., Karaseva M. V., Voroshilova A. A. Automated approach to building the multilingual frequency dictionary on system analysis and computer technologies //IOP Conference Series: Materials Science and Engineering. – IOP Publishing. – 2020. – Vol. 862. – № 4. – P. 042054.
34. Lyashevskaya O. Frequency dictionary of inflectional paradigms: core Russian vocabulary //Higher School of Economics Research Paper No. WP BRP. – 2013. – Vol. 35.
35. Mitrofanova O. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology //International Workshop on Language, Music, and Computing. – Springer, Cham. – 2015. – P. 69-76.

36. Palmer F. R., Frank Robert P. Semantics. – Cambridge university press. – 1981.
37. Sherstinova T. Y. et al. Frequency word lists and their variability (the case of Russian fiction in 1900-1930) // Conference of Open Innovations Association, FRUCT. – FRUCT Oy. – 2020. – № 27. – P. 366-373.
38. Sherstinova T., Martynenko G. 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 Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019). CEUR Workshop Proceedings. – 2020. – Vol. 2552, P. 105–120
39. Sherstinova T. Y., Skrebtsova T. Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900-1930 // IMS. – 2020. – P. 117-128.
40. Sinclair J., Carter R. Trust the text: Language, corpus and discourse. – Routledge. – 2004.
41. Stubbs M. Words and phrases: Corpus studies of lexical semantics. – Oxford: Blackwell publishers. – 2001. – P. 1-267.
42. Tribble C., Jones G. Concordances in the classroom: A resource guide for teachers. – Athelstan. – 1997.
43. Wang S. F. et al. Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in Two Conversational Corpora //International Journal of Computational Linguistics & Chinese Language Processing. – Vol. 17, № 2. – 2012.