Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2022»

Москва, 15–18 июня 2022 г.

Topic modeling of the Russian short stories of 1900–1930s: the most frequent topics and their dynamics

Tatiana Sherstinova, Anna Moskvina, Margarita Kirina, Asya Karysheva, and Evgenia Kolpashchikova
National Research University Higher School of Economics, Saint Petersburg
121 Kanala Griboedova Embankment, 190068, Saint Petersburg, Russia
{tsherstinova, admoskvina, mkirina}@hse.ru
{askarysheva, eokolpaschikova}@edu.hse.ru

Abstract

The article describes the results of an experiment on topic modeling of Russian short stories for three successive historical periods of the early 20th century: 1) the beginning of the 20th century until 1913, 2) the war-revolutionary period (1914–1922), and 3) the early Soviet period (1923-1930). Using the Latent Dirichlet Allocation (LDA) algorithm, 9 models were built — 3 samples of different sizes (100, 500, and 1000 stories) for each of the periods. It turned out that in every model there are very frequent "themes" (topics) that characterize with a high probability a fairly significant share of texts in each sample. Moreover, one can also observe a meaningful dynamics of these frequent topics over different time periods, which allows us to consider them as thematic and stylistic markers of the analyzed text collections along with the more traditional quantitative measures of text analysis. The variety of frequent topics turned out to be higher in the second and third periods, which can be explained by the greater lexical and stylistic diversity of the prose of the "era of change".

Keywords: quantitative text analysis; topic modeling; fiction; Russian short story; topic variety; dynamics of language and style

DOI: 10.28995/2075-7182-2022-21-XX-XX

Тематическое моделирование русского рассказа 1900–1930: наиболее частотные темы и их динамика

Шерстинова Т. Ю., Москвина А. Д., Кирина М. А., Карышева А. С., Колпащикова Е. О.
Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург
Россия, 190068, Санкт-Петербург, наб. канала Грибоедова, 121
{tsherstinova, admoskvina, mkirina}@hse.ru
{askarysheva, eokolpaschikova}@edu.hse.ru

Аннотация

В статье описаны результаты эксперимента по построению тематических моделей малой русской прозы (русского рассказа) трех последовательных исторических периодов начала XX века: 1) начала XX века до 1913 г. включительно, 2) военно-революционного периода (1914–1922) и 3) раннесоветского периода (1923–1930). С помощью алгоритма латентного размещения Дирихле (LDA), построено 9 моделей (по 3 выборки разного размера для каждого из периодов – по 100, 500 и 1000 рассказов). Оказалось, что в каждой из моделей присутствуют весьма частотные «темы» (топики), характеризующие довольно существенную долю текстов каждой выборки с высокой вероятностью, а также наблюдается содержательная динамика этих частотных тем по разным временным периодам, что позволяет считать их тематико-стилистическим маркерами анализируемых коллекций текстов наряду с более традиционными квантитативными мерами анализа текстов. Разнообразие частотных топиков оказалось выше во втором и третьем периоде (для

выборок в 500 и 1000 рассказов), что можно объяснить большим лексико-стилистическим разнообразием прозы «эпохи перемен».

Ключевые слова: квантитативный анализ лексики; тематическое моделирование; художественная проза; русский рассказ; тематическое разнообразие; динамика языка и стиля

1 Введение

Построение тематических моделей для коллекции текстовых документов — активно развивающееся направление автоматической обработки текста [5; 18; 19; 25; 26]. Этот метод машинного обучения позволяет выявлять из корпуса текстов скрытые семантические структуры — темы. Под «темой» понимается «набор ключевых слов, характеризующий отдельный документ или набор документов» [19, с. 221]. Каждый текст в анализируемом корпусе представляется как набор тем, т. е. один текст описывается несколькими темами одновременно. Стоит отметить, что зачастую при тематическом моделировании реализуется подход, называемый bag-of-words, т. е. не учитываются грамматические и синтаксические характеристики слов. Любой текст осмысляется как «случайная выборка слов, порожденная неким множеством тем» [там же]. Важно также и то, что при тематическом моделировании в результате бикластеризации происходит объединение в семантически схожие группы не только слов, но и текстов [45, с. 216].

Изначально методы тематического моделирования разрабатывались для анализа специальных (научных, технических, новостных и т. п.) документов. В последние годы появляется все больше примеров, когда эти методы применяются для интеллектуального анализа и кластеризации художественной прозы [8; 20; 21; 22; 24]. Однако результаты, которые получаются при обработке литературных текстов, принципиальным образом отличаются от результатов обработки текстов специальных [3; 28; 31; 42]. Метафоричность языка художественных произведений, образность повествования, применение разнообразных стилистических приемов, а во многих случаях и отсутствие ярко выраженной «темы повествования» приводят к тому, что математические подсчеты совместного употребления слов, лежащие в основе методов тематического моделирования, дают не совсем те результаты, которые исследователь ожидает получить от применения этих методов. С другой стороны, в отличие от традиционного анализа художественного текста, результатом построения такой тематической модели в большинстве случаев является не информация «в чем главная идея/содержание рассматриваемых текстов», сколько информация о тематико-стилистическом разнообразии анализируемой текстовой коллекции, в которой каждая из выделенных тем-топиков состоит из семантически связанных слов, формирующих некоторую интерпретируемую лексико-семантическую группу (или тематическое ядро).

Данное исследование продолжает серию работ по изучению языка, стиля и тематического разнообразия малой русской прозы первых трех десятилетий XX века, начатого в [6; 14; 15; 33; 35; 37; 38; 44], осуществляемых главным образом на материале Корпуса русского рассказа 1900-1930 гг. [16; 17; 34], и ставит своей целью построение тематических моделей для трех последовательных исторических периодов, каждый из которых соотносится со значимой исторической эпохой: I период (1900-1913 гг.) — начало XX века до Первой мировой войны, II период (1914-1922 гг.) — эпоха острых социальных катаклизмов, войн и революций (Первая мировая войны, Февральская и Октябрьская революция, Гражданская война), III период (1923–1930 гг.) — становление молодого советского государства. Вслед за [38; 39] мы полагаем, что исторический фон эпохи, в котором создаются литературные тексты, так или иначе будет проявляться в художественном творчестве современных ему писателей и оказывать влияние не только на язык, но и на содержание и тематику литературных произведений, косвенным отражением которых можно считать темы/топики, полученные в результате тематического моделирования.

В статье описаны результаты эксперимента по тематическому моделированию малой русской прозы (русского рассказа) начала XX века. С помощью алгоритма латентного размещения Дирихле (LDA) построено 9 тематических моделей, из которых выделены и описаны наиболее частотные топики, а также продемонстрировано изменение состава частотных топиков в зависимости от объема выборки.

2 Материал и методика

2.1 Отбор литературных текстов

Анализ динамики тематических моделей художественной прозы проводится на материале русского рассказа. Выбор рассказа как жанра для проведения исследования определяется тем, что рассказ является наиболее распространенным жанром прозы, охватывающим все литературные направления — рассказы присутствуют в творчестве практически всех прозаиков (и даже многих поэтов!), что позволяет вовлечь в исследование тексты максимального количества авторов, получить тем самым наиболее статистически достоверный «литературный портрет» эпохи и оценить все его тематическое разнообразие. Малый текстовый объем рассказа способствует тому, что рассказы значительно быстрее, чем повести и романы, проходят издательский цикл, достаточно большая их доля публикуется в литературных журналах, которые также вовлекаются в единый художественно-литературный процесс эпохи [17]. Можно утверждать, что рассказ, как особый жанр, выполняет «разведочную» функцию и даже работает на опережение, чутко улавливая и реагируя на изменения в общественном сознании и культуре общества» [там же], является «диагностом социальных процессов» [16]. Наконец, *«в русской литературе рассказ традиционно был сильным жанром. Пожалуй, лишь американская литература приближается в этом отношении к нашей»* [46].

Приведенные аргументы стали причиной того, что для задачи моделирования языка русской прозы были начаты именно для жанра рассказа. С этой целью был создан и продолжает развиваться Корпус русского рассказа[1] [13; 16; 17; 34].

По сравнению с предыдущими исследованиями, посвященными тематическому моделированию русского рассказа [6; 14; 15; 33; 35; 37; 38; 44], для получения более достоверных результатов мы существенно (в 10 раз) расширяем объем исследовательского материала — до 1000 рассказов в каждом из изучаемых периодов. Более того, нам показалось целесообразным посмотреть, насколько меняются результаты тематического моделирования при последовательном расширении исследовательской выборки — для 100, 500 и 1000 текстов, тем самым рассмотреть зависимость тематической модели от объема выборки. Ограниченный объем публикации вынуждает нас остановиться на рассмотрении наиболее частотных тем (топиков) художественных текстов для каждого временного среза, которые, однако, представляются достаточно показательными для русской литературы изучаемого периода.

Чтобы оценить изменение «тематического разнообразия» в динамике, было решено подготовить 9 выборок: по 100, 500 и 1000 рассказов для трех последовательных временных срезов — довоенного (1900-1913), военно-революционного (1914-1922) и раннесоветского (1923-1930). Поскольку конечной целью проводимых исследований является моделирование национального литературного процесса [36; 41], при формировании выборки основной акцент делался на включение текстов максимального количества русских писателей, работавших в жанре рассказа [16; 17], не только «известных» и хрестоматийных, но и малоизвестных и даже фактически забытых.

Основными источниками для формирования выборки стали два открытых литературных ресурса — библиотека Lib.ru Максима Машкова [9], являющаяся одним из старейших и представительных ресурсов русского литературного мира и уже упомянутый Корпус русского рассказа 1900-1930 гг. [13; 16; 17; 34], содержащий большой объем редких текстов, специально оцифрованных для этого корпуса. Однако для поставленных задач исследования этих двух текстовых коллекций оказалось недостаточно, и нам пришлось обращаться к другим открытым интернет-ресурсам (в частности, [10; 11; 12; 30; 43]), для пополнения текстовых коллекций раннесоветского периода.

При формировании выборок ставилась две задачи: обеспечение относительно равномерного распределения текстов по году написания/первой публикации рассказа и обеспечение максимальной представительности разных авторов внутри как периода в целом, так и для каждого отдельного года. Конкретные рассказы отбирались в случайном порядке, вне зависимости от тематики и содержания. В выборки не включались как достаточно крупные

[1] В настоящее время работа идет над периодом первых трех десятилетий XX в.

рассказы (больше 10000 словоупотреблений), так и очень краткие (менее 200 слов). Отобранные тексты анализировались целиком, вне зависимости от размера. Выборки меньшего объема входили подмножеством в выборки большего объема (то есть тексты выборки из 100 рассказов входят в выборку 500 и 1000).

Всего для исследования отобраны были отобраны 3000 текстов: 74% из них взяты Lib.ru, 24 % — из Корпуса русского рассказа и 2 % — из других открытых источников. В табл. 1 приведены данные по объему итоговых выборок в словоупотреблениях.

Период	Объем выборки			Количество разных писателей		
	100	500	1000	100	500	1000
1 (1900-1913)	369113	1820446	3564493	100	320	690
2 (1914-1922)	321693	1355994	2702207	100	254	366
3 (1923-1930)	302682	1516696	2541865	100	307	313

Таблица 1: Объем исследовательских выборок в словоупотреблениях.

Данные табл. 1 показывает, что суммарный объем текстов первого периода превышает выборки второго и третьего, а количество писателей, вовлеченных в литературный процесс в довоенный период, по нашим данным, также оказалось больше, чем в последующие годы. Это еще раз подтверждает выводы о том, что писательская активность в начале века была выше, чем в периоды острых социальных конфликтов и преобразований [17].

Далее по тексту выборки обозначаются в виде кода из двух чисел, первое из которых обозначает период, а второе — объем выборки в рассказах (напр., 2-500 — выборка текстов из второго периода в 500 рассказов).

2.2 Методика обработки данных

Тексты были лемматизированы с помощью библиотеки spaCy [7]. Из текстов удалялись все имена собственные и стоп-слова, служебные части речи и другие стоп-слова.

Для построения тематических моделей использовался алгоритм латентного размещения Дирихле (LDA), реализованный в библиотеке gensim [1; 2]. Оптимальное количество тем (топиков) для каждой выборки определялось автоматически на основе меры когерентности [29]: выбиралось то количество топиков, при котором достигалось наибольшее ее значение. Для подсчёта метрики использовалась функция CoherenceModel из модуля models библиотеки gensim [27]. Диапазон количества топиков при вычислении когерентности был определен в интервале от 10 до 45; метрика вычислялась при построении моделей в цикле с шагом в 5 топиков.

Графики когерентности для определения оптимального количества тем приведены на рис. 1, а итоговая статистика о количестве автоматических полученных тем для каждого из периодов представлена в табл. 2. Неожиданно, для всех трех периодов максимальный «тематический» разброс показала выборка среднего размера в 500 текстов. Для первого и второго периода количество рекомендуемых тем для крупной выборки в 1000 текстов меньше, чем для малых выборок, для раннесоветской прозы, напротив, это число несколько возрастает.

Период	Объем выборки		
	100	500	1000
1 (1900-1913)	20	40	15
2 (1914-1922)	35	40	15
3 (1923-1930)	25	45	30

Таблица 2: Количество тем (топиков), выявленных для каждой выборки.

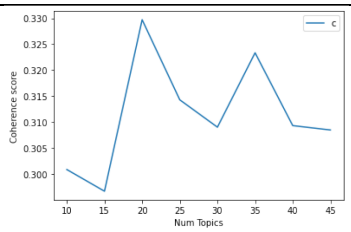
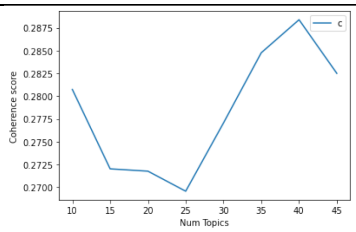
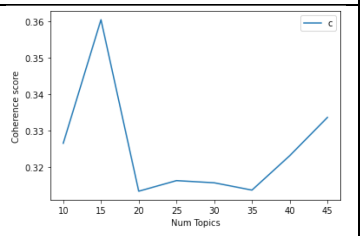
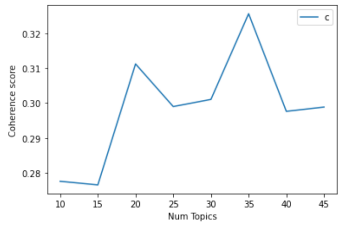
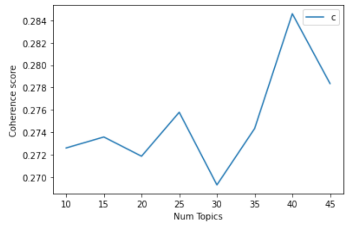
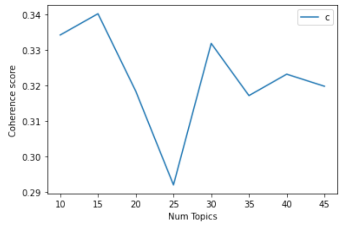
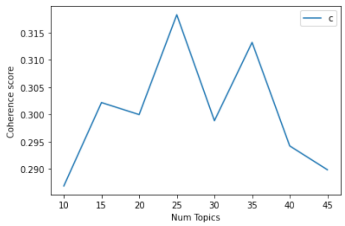
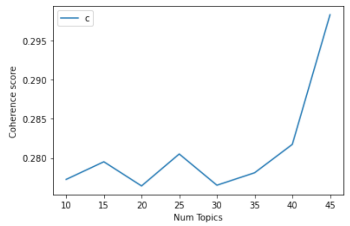
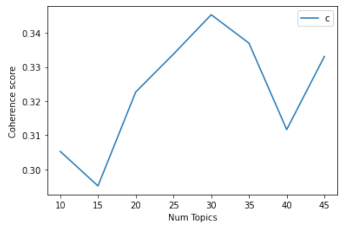
Период	Объем выборки		
	100	500	1000
1 (1900-1913)			
2 (1914-1922)			
3 (1923-1930)			

Рисунок 1: Графики когерентности для исследуемых выборок.

3 Результаты

Итогом работы алгоритмов тематического моделирования являются: 1) собственно список тем (топиков), представленный в виде набора ключевых слов, каждый из которых встречается в тексте с определенной вероятностью, 2) вероятности отнесения к каждой конкретной теме для каждого документа текстовой коллекции. Таким образом, имеет место «мягкая кластеризация» [19] исследуемой выборки. В центре нашего исследования находятся именно полученные темы, а также их частотность (распространенность) для выборки того или иного среза.

Ввиду больших объёмов выборки и большого количества выявленных тем, в рамках данного исследования было решено ограничиться наиболее частотными темами, характеризующими максимальное количество анализируемых текстов.

Для этого среди всех полученных топиков, были отобраны темы, которые показали встречаемость для наибольшего количества документов. В качестве нижнего порогового значения было выбрано 25%, то есть ниже рассматриваются только те темы, которые превышают установленное моделью пороговое значение для более чем четверти документов выборки. Таких тем оказалось следующее количество (см. табл. 3).

Период	Объем выборки		
	100	500	1000
1 (1900-1913)	2	3	4
2 (1914-1922)	2	4	6
3 (1923-1930)	2	4	6

Таблица 3: Количество наиболее частотных тем (топиков) для каждой выборки.

Однако детальное рассмотрение наполнения каждой из этих тем показало, что многие рассказы входят туда с небольшими вероятностями. Поэтому было решено взять еще два среза данных — подсчитать количество документов, входящих в каждую из тем: 1) с вероятностью не

менее 75%, 2) с вероятностью не менее 95%. В табл. I-III Приложения[2] представлены эти самые частотные темы для довоенного, военно-революционного и раннесоветского периодов соответственно. Темы представлены их номером (столбец № 2), с точки зрения статистики являющимся номинальной переменной [4], содержание топика с вероятностями ключевых слов, округленных до 10000-тысячных, представлено в правом столбце, и также приводится относительное количество рассказов, содержащих соответствующую тему с высокой вероятностью.

Рассмотрим полученные результаты.

Характеризуя слова, образующие топики, мы решили использовать понятие «мотив», поскольку для описания литературных текстов это кажется более привычным, чем «ключевые слова» или термины.

При выборке в 100 текстов во всех трех периодах по предложенной методике выделяются всего по 2 темы, все они достаточно общего плана. Для первого периода это топики 17(1-100) и 18(1-100). Темы представляют собой оппозицию статики и динамики (*стоять – пойти*, *прийти)*, молодости-старости (*молодой – старик*), личного и общего (*душа – народ*), одиночества – семьи (*женщина – отец, мать*), духовного и материального (*душа – деньги*). Есть в темах и существенные пересечения — это «голос», «спросить» и «земля». Мотив «голоса» дважды появляется в частотных темах выборки 500, для выборки в 1000 он уже уходит из верхней зоны тем, «земля» появляется в двух темах больших выборок, а слово «спросить» является абсолютным лидером довоенного периода, характеризуя 7 из 9 частотных тем.

Оценить соответствие выделенных топиков «тематике» текста можно на примере отдельных рассказов. Так, продажа/покупка становится центральным мотивом рассказа П. И. Астрова «Из жизни человека» (1907), с вероятностью 0,91 относящегося к теме 18(1-100)[3]: ('отец', 0.0095), ('пойти', 0.0069), ('старик', 0.0065), ('мать', 0.0057), ('деньга', 0.0046), ('прийти', 0.0041), ('земля', 0.0039), ('голос', 0.0039), ('народ', 0.0038), ('спросить', 0.0037). Главный герой — «*старик*» Алейка звонким «*голосом*» вопрошает на улице: «Нет ли чего продавать?», на что человек «с глухим сдавленным *голосом*» высовывается из окна своей комнаты и просит Алейку зайти к нему. Умирающий барин предлагает Алейке свои последние вещи; Алейка долго не соглашается на сделку, повторяя, что барин назначил слишком большую сумму. Однако все-таки Алейка покупает вещи барина, отдав за них меньше денег, чем изначально просил «продавец»… В рассказе много глаголов движения «*пойти*»/«*прийти*», во время торгов старика с барином неоднократно встречается «*спросить*» и «*деньги*». «*Народ*» и «*земля*» появляется в самом конце рассказа, в котором говорится о похоронах умершего барина. Два слова из топика («*отец*» и «*мать*») не встретились в рассказе ни разу[4].

При расширении выборки до 500 текстов, «динамичная» тема 18(1-100) модифицируется в тему 3(1-500), появляется мотив «мужика», значительная доля ее образующих слов — это глаголы, причем не только динамичные, но и статичные (*смотреть, стоять*). А тема первого периода 17(1-100), которую можно условно назвать «молодая женщина в комнате», при выборке в 500 текстов распадается на две — тему «любви» 2(1-500) и тему «семьи» 22(1-500).

С высокой вероятностью (0,999) относится к теме 2(1-500) рассказ В. Березовского «Утро» (1901): Иванов идёт через лес к своей возлюбленной. Он наслаждается природой и своей любовью. В то же время его возлюбленная не спит всю ночь, ждёт его у окна. Наконец, с приходом утра, Иванов приходит к дому. Возлюбленная вспоминает, что забыла отпереть дверь, и ещё несколько минут не опирает, предвкушая встречу со счастьем. Наконец встреча происходит. Как видно, содержание рассказа действительно хорошо коррелирует с автоматически определенной для него темой.

На выборке в 1000 текстов мы уже имеем 4 частотные темы. Тему 6(1-1000) можно считать «преемницей» тем 18(1-100) и 3(1-500), но возникает новая окраска — «ночь», «лес». Стоит

[2] Содержание частотных тем вынесено в приложение в конце статьи, чтобы не превышать требуемого объема.

[3] См. Табл. I Приложения.

[4] Чем больше вероятность вхождения рассказа в топик, тем больше вероятность встретить в тексте рассказа ключевые слова топика, причем неоднократно.

обратить внимание на высокую частоту этого топика: 10% или 100 рассказов из 1000 содержат в себе эту тему с вероятностью более 95%. Примером хорошего соответствия топику 6(1-1000) с вероятностью 0,999 можно считать рассказ Б. А. Верхоустинского «Лесное озеро» (1912) — действие происходит в лесу, на озере, потом дома, присутствует и дед-старик. Две следующие по частоте темы 5(1-1000) и 14(1-1000), которые можно условно назвать «любовь и семья» и «любовь и душа», восходят к темам 2(1-500) и 22(1-500), и ранее — к теме 17(1-100). Наконец, появляется тема 3(1-1000), которую можно условно назвать «Деньги – дом – служба». Это единственная из всех частотных тем с ориентацией на финансовое состояние и понятие «служба». При этом сохранятся мотив семьи («жена», «дом»).

Частотные топики второго, военно-революционного, периода во многом наследуют тематику рассказов начала века. Но предсказуемо появляются и новые мотивы. Так, на малой выборке в 100 рассказов максимальную частоту показали две темы 16(2-100) и 22(2-100), обе из которых можно считать «преемницами» темы «молодая женщина в комнате» первого периода. Между темами есть существенное пересечение в виде глаголов «пойти», «спросить», «сидеть». Время сдвигается в сторону «вечера» и «ночи», акцент с нейтрального «дом» в теме 16(2-100) смещается на «дверь» и «окно», которые можно считать метафорами *расставания/встречи* и *ожидания*. В теме 22(2-100) возникает мотив «письма», весьма частотный для русской прозы того периода [35], но особенно значимый во время войны и вынужденной разлуки, а также мотив *мысли*, наполняющий человеческое сознание в тревожные времена. Примером рассказа, относящегося и к 16, и к 22 топикам, могут служить «Ситцевые колокольчики» Ю. Л. Слезкина (1922). Один из героев рассказывает другому о том, как в молодости гостил в семье, где было 5 дочерей-красавиц, в одну из которых он был влюблен, но мог лишь позволить любоваться ею со стороны, поскольку она рано вышла замуж по любви. Переехав жить в другое место, герой ждет писем от своей старой знакомой и получает их, что коррелирует с мотивами ожидания и письма. Другой рассказ, относящийся к этим двум топикам — «Тень счастья» Н. Д. Телешова (1921), в котором главного героя разыгрывают коллеги, посылая любовные письма от прекрасной незнакомки. Герой никак не может встретиться с ней и страдает, особенно когда коллегам надоедает эта игра и они «убивают» его возлюбленную, прислав ему письмо с новостями о смерти девушки.

Новым мотивом тем этого периода является понятие «жизнь», которое никак не проявилось в частотных темах малой выборки из 100 рассказов, но присутствует с высокими вероятностями во всех четырех топиках выборки из 500 рассказов и в половине частотных тем выборки из 1000 рассказов. *Ценность жизни* наиболее остро проявляется во времена социальных катастроф, поэтому появление этого мотива вполне закономерно. Надо отметить, что мотив «жизни» сохранится и в прозе третьего, раннесоветского, периода. Самая частотная тема выборки 500 20(2-500) наиболее неоднородна: помимо «жизни на земле» и «голоса сердца» она окрашена оппозициями: «ночь – солнце», «белый – черный». Тема 24(2-500) формируется вокруг «дома», содержит много глаголов (*пойти, сидеть, спросить, смотреть*), субъект исчезает, но появляется мотив «ночи». Тема 25(2-500) — это традиционная тема женской любви, но с акцентом на мотивы «жизни» и «мысли». Наконец, тема 29(2-500) — первая в этом блоке, в котором угадывается революционный дух эпохи: действие переносится в «город», на «улицу», действующим лицом становится «толпа», но сквозь призму семейных (*жена, отец*) и духовных (*душа*) ценностей. К этому топику был отнесен, например, рассказ Н.Н. Никандрова «Катаклизма» (1917), показывающий митинги глазами простых городских жителей, которые даже не знают слова «митинг», а их протест напрямую связан со страхом за семью и близких.

Из 6 наиболее частотных тем выборки 1000 две — 0(2-1000) и 10(2-1000) — относятся к традиционной тематике «женщина и любовь», при этом 0(2-1000) восходит к типичной довоенной теме «женщина в комнате» (14(1-1000), 5(1-1000)), но с новым мотивом «нужности». Тема 10(2-1000) является наиболее светлой во всем этом блоке, ее темообразующие слова легко преобразуются в вполне связный текст, напр., «*Жизнь. Душа. Любовь. Женщина любит сердцем. Счастье для прекрасной девушки — письмо*». Тема 8(2-1000) весьма близка к теме 6(1-1000) первого периода (*земля, лес, отец, старый*), но окрашена оппозициями *белый-черный*, а также *земля-небо*, а тема 4(2-1000) имеет много пересечений с темой 3(1-500) — *пойти, отец, дом, смотреть, сидеть, мужик* (причем порядок слов, отражающий вероятности, тоже во многом совпадает). Новыми словами темы становятся

«мать» и отмеченные нами мотивы, общие для всего второго периода — «жизнь» и «окно». Уникальными в этом периоде является «созерцательный» топик 1(2-1000): «смотреть в окно в ночи» на «воду, лес и небо», но с оттенком оппозиции «белый-черный» и единственный «военный» топик 3(2-1000), который можно описать как «между жизнью и смертью», «мысли солдата/капитана в последнюю минуту». К «военному» топику был отнесен, к примеру, рассказ Н. Уклеина «Поезд мертвых» (1915), в котором на празднично подготовленный перрон приходит поезд, полный вражеских трупов. В последней теме кажется несколько неожиданной появление такого персонажа как «князь», впрочем, его вероятность явно меньше других темообразующих слов этого периода.

Ожидаемо, частотные топики третьего, раннесоветского, периода содержательно ближе к военно-революционному периоду, чем к рассказам начала века. Следует отметить, что у большинства частотных тем этого периода темообразующим становится слово «лицо» (появляются и другие слова, связанные описанием внешности героев — «голова», «глаз»), становится меньше глаголов движения, максимум для которых наблюдался во втором периоде, сохраняется актуальность понятия «жизнь», повышается значимость (частота) «земли» и «воды». На малой выборке в 100 рассказов выделяются две частотные темы, первая 9(3-100) из которых напоминает темы 8(2-1000), 6(1-1000), 1(2-1000) — *земля, вода, ночь, старик, белый-черный*, а вторая 12(3-100) содержательно представляет собой синтез уже знакомых с начала века мотивов (*голос, спросить, комната*) и слов, отражающих реалии нового времени — *тетка, работа, товарищ*. Так, образ *женщины*, частотный в 1-2 периодах, замещается образом *тетки-товарища*. Можно предположить, что две частотные темы первой выборки символизируют противопоставление старого и нового мира. Так, среди героев рассказа В. Инбер «Квартира № 32» (1924), принадлежащего к топику 12, есть Эсфирь Абрамовна, почти карикатурная обитательница коммунальной квартиры; слово «товарищ», впрочем, в тексте относится только к мужчинам.

Для выборки в 500 рассказов третьего периода определены 4 частотные темы, три из которых достаточно традиционны и встречались с небольшими изменениями ранее. Выделяется на их фоне тема 40(3-500), которая ассоциируется с состоянием *болезни*. Одним из рассказов, принадлежащих к этому топику, стало «Общежитие» В. Зазубрина (1923), в котором описываются соседи по общежитию; среди них есть доктор, пишущий научный труд, и упоминаются проблемы со здоровьем других соседей. Неожиданно, на этом срезе не проявилось ни одной частотной темы, маркированной советскими реалиями.

Наконец, на большой выборке в 1000 рассказов, мы имеем 6 наиболее частотных топиков, тематика которых распределяется следующим образом: 20(3-1000) имеет основным лейтмотивом «мужик/товарищ на земле», 26(3-1000) можно было бы назвать «собрание граждан/товарищей/рабочих» (кстати, здесь присутствует и более свойственные первому периоду «голос» и «спросить», но в этом контексте они воспринимаются совсем иначе). Примером рассказа о собрании товарищей являются «Именины» М.Я. Козырева (1925). В этом тексте переосмысляется ритуал празднования (и другие части жизни) с учетом наступления новой эпохи. Герои празднуют вместе со своим начальником и обсуждают актуальные перемены в обществе. Интересна по составу тема 10(3-1000): связующими элементами являются объекты природы (*вода, лес, земля, снег*), при этом отсутствует явный субъект, хотя описывается его/ее *глаза, лицо, голова*. Тема 29(3-1000) представляет собой возвращение к традиционной «семейной жизни в любви» (см. 5(1-1000) в первом периоде). Последние две темы 23(3-1000) и 2(3-1000) можно условно отнести к развитию науки и техники и освоению новых территорий.

Обобщая полученные данные, можно сделать следующие выводы:

1) Наблюдается содержательная динамика частотных тем по разным временным периодам. Поэтому хотя тематические модели далеко не всегда являются отражением собственно «тематики» литературного текста, тем не менее их можно рассматривать как тематико-стилистический маркер анализируемой коллекции текстов.
2) Частотные темы, построенные на небольших выборках, трансформируются или перераспределяются на выборках большего объема, некоторые темы можно считать универсальными (сквозными) для всех трех рассмотренных периодов. Топики на

больших выборках в среднем выглядят более конкретными с содержательной точки зрения.

3) Разнообразие частотных топиков выше во втором и третьем периоде — для выборок в 500 и 1000 рассказов. Очевидно, это определяется большим лексико-стилистическим разнообразием прозы «эпохи перемен». И это несмотря на то, что для первого периода выборка содержит максимальное количество разных авторов.
4) В случае работы с художественным текстом оценка адекватности модели посредством оценки интерпретируемости порожденных тем является непростой задачей. Традиционные подходы, предполагающие понимание экспертом значения «удачного» топика, в некоторых случаях оказываются затруднительными, поскольку тема может представлять собой как группу семантически близких слов, так и свертку сюжета, а также подвергаться влиянию образности художественного языка и имплицитных смыслов произведения. Поэтому вопрос правильного подхода к оценке интерпретируемости результатов работы модели на материале художественной прозы должен стать темой отдельного исследования.
5) Даже в том случае, когда конкретно выделенный топик является хорошо интерпретируемым и рассказ относится к нему с большой вероятностью, ключевые слова одного единственного топика как правило не исчерпывают основные тематические категории рассматриваемого текста.

4 Заключение

Проведенное исследование позволило выявить наиболее частотные темы рассказов для трех последовательных исторических периодов, полученные автоматически в результате построения тематической модели. Под частотными здесь понимаются темы, отнесенные к большему количеству проанализированных документов текстовой коллекции с высокой вероятностью. Разумеется, тематическое моделирование в большинстве случаев приводят к результатам, отличным от того, что понимается под темой литературного произведения при экспертной оценке. Тем не менее, построенные модели кажутся во многом осмысленными, позволяют посмотреть на литературные тексты с неожиданной стороны и выявить отличительные особенности больших массивов текстов, прочитать и переосмыслить которые не в состоянии ни один эксперт. Поэтому эксперименты с тематическим моделированием литературных произведений безусловно имеет смысл продолжать.

В нашем случае, как и в большинстве других компьютерных исследований литературного материала, встает вопрос об интерпретируемости полученных тем. Разумеется, полученные темы не являются (и не могут быть) отражением «темы рассказа» в литературоведческом понимании. В целом, на выборках большего объема рассмотренные частотные топики выглядят более «содержательными» и «интерпретируемыми», чем на меньших выборках. Тем не менее, полученные данные показывают, что содержание топиков между тремя исследуемыми периодами отличается, что позволяет считать их тематико-стилистическим маркером анализируемой коллекции текстов наряду с более традиционными квантитативными мерами анализа текстов.

Задача интерпретации автоматически полученных тем является непростой задачей, поскольку традиционные подходы, используемые при работе со специальными текстами, при анализе художественной литературы оказываются недостаточными вследствие содержательных и стилистических ее особенностей. Предполагается, что разработка правильного подхода к оценке интерпретируемости результатов работы модели на материале художественной прозы должна стать темой отдельного исследования.

При этом следует иметь в виду, что на сегодняшний день существует большое количество алгоритмов тематического моделирования, и результаты их применения к одной и той же текстовой коллекции могут несколько отличаться. На построенные модели оказывают влияние особенности предобработки текстов, а также и сама методология проведения исследований. Поэтому полученные частотные топики, описанные в данной статье, не могут рассматриваться как единственно возможная и, тем более, оптимальная схема. Выявление оптимальной схемы работы алгоритмов тематического моделирования на художественных текстах — задача

будущих исследований, весьма важная для решения многих гуманитарных задач, связанных с *distant reading* [23].

Проведенный эксперимент стоит считать пилотным, а его результаты — предварительными. Тем не менее сравнение выявленных тем по периодам отчасти согласуется с информацией о тематике русских рассказов, полученных при экспертной литературоведческой оценке для тех же временных срезов [37]. Согласно этим исследованиям, в частности, в довоенный период по сравнению с последующими максимально преобладают семейные ценности, романтическая любовь, интерес к финансовому благополучию. Эти темы хорошо видны на частотных топиках построенной модели. Выявленные экспертом особенности малой прозы раннесоветского периода, в частности — высокая доля рассказов о жизни на селе, а также интерес к науке и техническому прогрессу, — также нашли свое отражение в полученных частотных топиках.

Однако у построенных моделей есть и недостатки. В первую очередь, таковыми кажутся повторения темообразующих слов в нескольких частотных темах (*спросить, дом, комната* и др. для 1-го периода, *пойти, жизнь, спросить* и др. для 2-го периода, *лицо, жизнь, земля* и др. для 3-го периода). С одной стороны, эти слова можно считать «ключевыми» для прозы соответствующих эпох, но с другой стороны, такие повторы «размывают» содержательные отличия между топиками, что снижает общее качество модели. Другим недостатком моделей можно считать «потерю» военно-революционной тематики, ожидаемо частотной для рассказов 2-го периода [38], а также темы насилия, довольно частотной для русской прозы изучаемого периода [6]. Объяснения этому явлению могут быть разные. Во-первых, возможно, указанные топики просто не попали в ограниченный список частотных, рассмотренных в данной статье. Во-вторых, это может быть связано с тем, что при построении моделей была отфильтрована частотная лексика, встречающаяся более чем в 80% текстов для каждой выборки. Тем самым могло быть утрачено «военное» своеобразие второго периода. Отсюда можно сделать вывод, что при анализе художественных текстов, в отличие от специальных текстов, нужно очень осторожно подходить к «отбрасываемой» лексике и имеет смысл повторить расчеты для более мягких условий построения модели. В-третьих, возможно, что статистика, полученная в предыдущих исследованиях [37; 44] на материале 310 текстов не является достаточно представительной, а как раз новые данные, выполненные на выборке в 3000 рассказов, точнее показывают реальную дистрибуцию тем.

Кроме того, стоит отметить, что без первичного фильтра в 25% рассказов от общего объема выборки количество частотных тем могло быть существенно выше и расширило бы перечень общих частотных тем. Будем считать это задачей уже следующего этапа исследования. Поскольку рассмотрены результаты тематической модели, построенной без применения частеречных фильтров, стоит попробовать другие алгоритмы и методы — например, рассмотреть только существительные. Кажется целесообразным сравнить полученные результаты с частотными словарями художественной прозы и со списками ключевых слов [32; 40], а также провести эксперименты, предназначенные для оценки интерпретируемости полученных тематических моделей.

Благодарности

Публикация подготовлена в результате проведения исследования по проекту № 21-04-053 «Методы искусственного интеллекта для филологических исследований» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)».

References

[1] Blei D. M. Probabilistic topic models, Communications of the ACM. — 2012. — Vol. 55(4) — pp. 77–84.
[2] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation, The Journal of machine Learning research. — 2003. — Vol. 3. — pp. 993–1022.
[3] Da N. Z. The computational case against computational literary studies, Critical Inquiry. — 2019. — Vol. 45(3). — pp. 601-639.
[4] Glass V, Stanley J. Statistical Methods in Education and Psychology. — Englewood Cliffs, NJ: Prentice-Hall, 1972.

[5] Greene D., Cross J.P. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis, Proceedings of the ACM Web Science Conference (WebSci'15), Oxford, UK. — 2015.
[6] Gryaznova E., Kirina M. Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900–1930, 2021 International Conference "Internet and Modern Society", IMS 2021, CEUR Workshop Proceedings, 2021, Vol. 3090, pp. 281–290.
[7] Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing — 2017.
[8] Jockers M. L., Mimno D. Significant themes in 19th-century literature, Poetics. — 2013. — Vol. 41(6). — pp. 750–769.
[9] Lib.ru: "Classics" (Maxim Moshkov's Library), Web: http://az.lib.ru.
[10] Library CoolLib. 2012 – 2022, Web: https://coollib.net/
[11] Library LitMir, Web: https://www.litmir.me/
[12] Library RoyalLib.Com, 2010-2022, Web: https://royallib.com/
[13] Corpus of Russian Short Stories of 1913-1930s: https://russian-short-stories.ru/
[14] Martynenko G. Stylized syntactic triads in Russian short story of the first third of the 20th century [Stilizovannyye sintaksicheskiye triady v russkom rasskaze pervoy treti XX veka], Proceedings of the Int. Conf. 'Corpus Linguistics – 2019', St. Petersburg State University, St. Petersburg — 2019. — pp. 395–404.
[15] Martynenko G., Sherstinova T. Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture, Digital Transformation and Global Society. DTGS 2018. *Communications in Computer and Information Science*, Springer, Switzerland. — Vol. 859 — 2018. — pp. 299–309.
[16] Martynenko G.Ya., Sherstinova T.Yu., Melnik A.G., Popova T.I. Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century), Proc. of the XXI Int. United Conference 'The Internet and Modern Society', IMS-2018, Computational linguistics and computational ontologies. ITMO University, St. Petersburg — 2018. — Iss. 2 — pp. 99–104.
[17] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamirajlova E.V. On the Principles of Creation of the Russian Short Stories Corpus of the First Third of the XX Century [O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka] // Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics 'TEL 2018'. Kazan Federal University, Kazan. — 2018. — pp. 180–197.
[18] McFarland D. A. et al. Differentiating language usage through topic models // Poetics. — Vol. 41(6). — 2013. — pp. 607–625.
[19] Mitrofanova O. Topic modeling of special texts based on LDA algorithm [Modelirovanije tematiki special'nyh tekstov na osnove algoritma LDA] // XLII International philological conference [XLII Mezhdunarodnaya filologicheskaya konferencija]. — 2014.
[20] Mitrofanova O.A. Analysis of Fiction Text Structure by means of Topic Modelling: Case Study of "Master and Margarita" Novel by M. A. Bulgakov [Issledovanie strukturnoj organizacii hudozhestvennogo proizvedenija s pomoshh'ju tematicheskogo modelirovanija (opyt raboty s tekstom romana «Master i Margarita» M.A. Bulgakova)], Korpusnaya lingvistika-2019. — 2019. — pp. 387–394.
[21] Mitrofanova O.A. Topic modelling of the Corpus of 'Russian folk tales by A. N. Afanasiev', Structural and applied linguistics [Strukturnaya i prikladnaya linguistika]. — Vol. 11. — 2015. — pp. 146–154.
[22] Mitrofanova O.A., Sedova A.G. Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // Information Technology and Computational Linguistics (ITCL 2017), ICPS Proceedings, IMS2017: Proceedings of the International Conference IMS-2017 — 2017.— pp. 175–180.
[23] Moretti F., Distant Reading, London: Verso, 2013.
[24] Navarro-Colorado B. On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry //Frontiers in Digital Humanities. — 2018. — Vol. 5.
[25] Nikolenko S., Koltcov S., Koltsova O. Topic modelling for qualitative studies, Journal of Information Science. — Vol. 43(1). — 2017. — pp. 88–102.
[26] Panicheva P., Litvinova O., Litvinova T. Author Clustering with and Without Topical Features, Speech and Computer, Proceedings of the 21st Int. Conf., SPECOM 2019, LNAI 11658, Springer, Cham — 2019. — pp. 348–358.
[27] Rehurek R., Sojka P. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. — Vol. 3(2). — 2011.
[28] Rhody L. M. Topic Modelling and Figurative Language, Journal of Digital Humanities. — 2012.
[29] Röder M., Both A., and Hinneburg A.: Exploring the Space of Topic Coherence Measures, In Proceedings of the eighth International Conference on Web Search and Data Mining. — 2015.

[30] Ruthenia.ru https://www.ruthenia.ru/
[31] Schöch C. Topic modeling genre: an exploration of french classical and enlightenment drama, arXiv preprint arXiv:2103.13019. — 2021.
[32] Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina, M. Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930), 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy. — 2020. — pp. 366–373.
[33] Sherstinova T., Kirina M. Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers, 6th International Conference on Digital Transformation and Global Society, DTGS 2021, *Communications in Computer and Information Science* (CCIS). — Vol. 1503, — pp. 332–346.
[34] 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), St. Petersburg, Russia, CEUR Workshop Proceedings — Vol. 2552. — 2020. — pp. 105–120.
[35] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction, Advances in Computational Intelligence, MICAI 2020, Lecture Notes in Computer Science. — Vol. 12469. — 2020. — pp. 134–151.
[36] Sherstinova T., Moskvina A., Kirina M. Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian, 29th Conference of Open Innovations Association FRUCT, FRUCT 2021. — pp. 313-323.
[37] Sherstinova T., 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, Proceedings of the International Conference "Internet and Modern Society" IMS-2020. CEUR Workshop Proceedings. — pp. 117-128.
[38] Skrebtsova T. G. Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories, Proceedings of the International Conference “Internet and Modern Society” IMS-2020. CEUR Workshop Proceedings. — Vol. 2813. — 2020. — pp. 265-276.
[39] Skrebtsova T. Narrative structure of the Russian short story in the early XX century [Struktura narrativa v russkom rasskaze nachala XX veka], Proc. of the Int. Conf. Corpus Linguistics-2019, St. Petersburg. — 2019. — pp. 426–431.
[40] Skrebtsova T., Grebennikov A., Sherstinova T. The Dynamics of Vocabulary in Russian Prose (Based on Frequency Dictionaries of the Corpus of Russian Short Stories 1900-1930), 21st Annual International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2021 [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii]. — 2021. — pp. 646-659.
[41] Tynyanov Yu. Archaists and Innovators [Arkhaisty i novatory]. Priboi Publ., Leningrad, 1929.
[42] Uglanova I., Gius E. The Order of Things. A Study on Topic Modelling of Literary Texts, Proc. of the CHR 2020: Workshop on Computational Humanities Research, CEUR Workshop Proceedings. — 2020. — Access mode: http://ceur-ws.org/Vol-2723/long7.pdf.
[43] Wikisource: https://ru.wikisource.org/
[44] Zamiraylova E., Mitrofanova O. Dynamic topic modelling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization, 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*. — Vol. 2552. — 2020. — pp. 321–339.
[45] Korshunov A., Gomzin A. Tematicheskoye modelirovaniye tekstov na yestestvennom yazyke [Thematic modeling of natural language texts]. In: Trudy Instituta sistemnogo programmirovaniya RAN [Proceedings of the Institute for System Programming RAS], 2012, no. 23, https://ispranproceedings.elpub.ru/jour/article/view/982?locale=ru_RU.
[46] Nagibin Yu. M. Antologiya russkogo sovetskogo rasskaza. Predisloviye [Anthology of Russian Soviet short stories. Foreword], Bibliotechka «Knizhnogo obozreniya». Moscow: Knizhnoe obozrenie, 1987.

Приложение

Выборка	Тема	Кол-во текстов		Содержание топика
		P>75%	P>95%	
1-100	17	29%	17%	('голос', 0.0040), ('стоять', 0.0034), ('спросить', 0.0034), ('минута', 0.0030), ('душа', 0.0030), ('комната', 0.0028), ('женщина', 0.0028), ('молодой', 0.0027), ('дом', 0.0027), ('земля', 0.0027)
	18	9%	5%	('отец', 0.0095), ('пойти', 0.0069), ('старик', 0.0065), ('мать', 0.0057), ('деньга', 0.0046), ('прийти', 0.0041), ('земля', 0.0039), ('голос', 0.0039), ('народ', 0.0038), ('спросить', 0.0037)
1-500	2	13,8%	5,6%	('смотреть', 0.0037), ('любить', 0.0035), ('окно', 0.0031), ('душа', 0.0028), ('сидеть', 0.0027), ('комната', 0.0027), ('голос', 0.0026), ('белый', 0.0025), ('ночь', 0.0025), ('тёмный', 0.0025)
	3	13,4%	6%	('пойти', 0.0041), ('отец', 0.0037), ('дом', 0.0033), ('смотреть', 0.0032), ('спросить', 0.0028), ('земля', 0.0027), ('стоять', 0.0027), ('мужик', 0.0026), ('сидеть', 0.0025), ('прийти', 0.0024),
	22	7,4%	3%	('жена', 0.0033), ('спросить', 0.0028), ('комната', 0.0026), ('голос', 0.0025), ('дом', 0.0024), ('женщина', 0.0024), ('сидеть', 0.0023), ('час', 0.0023), ('нужный', 0.0022), ('минута', 0.0022)
1-1000	6	23%	10%	('земля', 0.0035), ('пойти', 0.0031), ('отец', 0.0028), ('стоять', 0.0026), ('старик', 0.0026), ('старый', 0.0026), ('дом', 0.0024), ('ночь', 0.0024), ('сидеть', 0.0023), ('лес', 0.0023)
	5	8,4%	2,5%	('любить', 0.0030), ('жена', 0.0028), ('ребёнок', 0.0028), ('сидеть', 0.0026), ('дом', 0.0026), ('муж', 0.0025), ('спросить', 0.0024), ('комната', 0.0024), ('отец', 0.0024), ('минута', 0.0023)
	14	8%	2,3%	('любить', 0.0034), ('душа', 0.0034), ('комната', 0.0029), ('любовь', 0.0028), ('ночь', 0.0028), ('мысль', 0.0028), ('спросить', 0.0028), ('женщина', 0.0028), ('странный', 0.0027), ('сердце', 0.0026)
	3	0,9%	0,3%	('деньга', 0.0055), ('комната', 0.0046), ('жена', 0.0044), ('рубль', 0.0043), ('спросить', 0.0039), ('квартира', 0.0038), ('час', 0.0035), ('пойти', 0.0031), ('дом', 0.0031), ('служба', 0.0029)

Таблица I: Наиболее распространенные топики периода 1900–1913 гг.
(для объемов выборки в 100, 500 и 1000 рассказов)

Выборка	Тема	Кол-во текстов		Содержание топика
		P>75%	P>95%	
2-100	16	10%	9%	('женщина', 0.0067), ('пойти', 0.006323), ('сидеть', 0.0056), ('ночь', 0.0045), ('дверь', 0.0041), ('стоять', 0.0038), ('белый', 0.0038), ('спросить', 0.0036), ('голос', 0.0034), ('окно', 0.0034)
	22	10%	6%	('любить', 0.0061), ('письмо', 0.0051), ('сидеть', 0.0041), ('душа', 0.0040), ('мысль', 0.0040), ('дом', 0.0038), ('пойти', 0.0037), ('спросить', 0.0037), ('молодой', 0.0036), ('вечер', 0.0035)
2-500	20	9%	3,60%	('жизнь', 0.0049), ('земля', 0.0033), ('ночь', 0.0032), ('солнце', 0.0030), ('пойти', 0.0029), ('белый', 0.0029), ('голос', 0.0027), ('смотреть', 0.0026), ('сердце', 0.0026), ('чёрный', 0.0024)
	24	6,60%	2%	('пойти', 0.0030), ('дверь', 0.0029), ('смотреть', 0.0027), ('сидеть', 0.0027), ('спросить', 0.0027), ('комната', 0.0027), ('жизнь', 0.0027), ('ночь', 0.0026), ('голова', 0.0025), ('дом', 0.0025)
	25	4,60%	2,20%	('жизнь', 0.0046), ('душа', 0.0036), ('женщина', 0.0032), ('мысль', 0.0031), ('любить', 0.0029), ('сидеть', 0.0029), ('спросить', 0.0027), ('минута', 0.0027), ('пойти', 0.0025), ('прийти', 0.0024)
	29	3,60%	2,60%	('жизнь', 0.0044), ('душа', 0.0036), ('город', 0.0036), ('жена', 0.0034), ('смотреть', 0.0033), ('дом', 0.0033), ('толпа', 0.0031), ('спросить', 0.0028), ('улица', 0.0026), ('отец', 0.0026)
2-1000	0	17%	6,5%	('жизнь', 0.0041), ('спросить', 0.0028), ('женщина', 0.0027), ('комната', 0.0027), ('любить', 0.0026), ('дом', 0.0025), ('смотреть', 0.0024), ('сидеть', 0.0024), ('нужный', 0.0024), ('пойти', 0.0021)
	8	7,7%	2,9%	('земля', 0.0053), ('ночь', 0.0036), ('белый', 0.0031), ('отец', 0.0029), ('душа', 0.0027), ('пойти', 0.0027), ('лес', 0.0027), ('старый', 0.0027), ('чёрный', 0.0026), ('нёбо', 0.0026)
	1	2%	0,9%	('вода', 0.0041), ('чёрный', 0.0039), ('ночь', 0.0037), ('белый', 0.0036), ('лес', 0.0032), ('окно', 0.0028), ('пойти', 0.0027), ('смотреть', 0.0027), ('нёбо', 0.0026), ('стоять', 0.0025)
	4	2,4%	0,6%	('пойти', 0.0058), ('отец', 0.0046), ('дом', 0.0041), ('смотреть', 0.0041), ('сидеть', 0.0040), ('мать', 0.0033), ('спросить', 0.0031), ('мужик', 0.0031), ('окно', 0.0030), ('жизнь', 0.0029)
	3	1,2%	0,5%	('капитан', 0.0042), ('жизнь', 0.0034), ('смерть', 0.0031), ('минута', 0.0027), ('последний', 0.0025), ('мысль', 0.0022), ('солдат', 0.0022), ('стоять', 0.0020), ('князь', 0.0020), ('сторона', 0.0020)
	10	1,1%	0,3%	('жизнь', 0.0128), ('душа', 0.0093), ('любовь', 0.0082), ('женщина', 0.0077), ('любить', 0.0056), ('сердце', 0.0044), ('счастие', 0.0041), ('прекрасный', 0.0038), ('письмо', 0.0032), ('девушка', 0.0033)

Таблица II: Наиболее распространенные топики периода 1914–1922
(для объемов выборки в 100, 500 и 1000 рассказов)

Выборка	Тема	Кол-во текстов		Содержание топика
		P>75%	P>95%	
3-100	9	25%	11%	('лицо', 0.0052), ('земля', 0.0049), ('белый', 0.0046), ('старик', 0.0038), ('вода', 0.0037), ('ночь', 0.0036), ('жизнь', 0.0034), ('мать', 0.0034), ('сидеть', 0.0034), ('чёрный', 0.0034)
	12	13%	8%	('отец', 0.0078), ('голос', 0.0062), ('тётка', 0.0057), ('спросить', 0.0056), ('работа', 0.0054), ('товарищ', 0.0051), ('старик', 0.0050), ('лицо', 0.0049), ('комната', 0.0045), ('жизнь', 0.0044)
3-500	36	11%	4,2%	('пойти', 0.0041), ('лицо', 0.0038), ('голова', 0.0032), ('земля', 0.0030), ('вода', 0.0029), ('лес', 0.0029), ('ночь', 0.0027), ('старик', 0.0027), ('мужик', 0.0027), ('белый', 0.0026)
	0	7,2%	2,4%	('лицо', 0.0043), ('дом', 0.0039), ('земля', 0.0037), ('жизнь', 0.0033), ('пойти', 0.0032), ('старик', 0.0030), ('ночь', 0.0029), ('отец', 0.0029), ('сидеть', 0.0028), ('смотреть', 0.0028)
	40	3,6%	1,6%	('жизнь', 0.0037), ('комната', 0.0035), ('белый', 0.0028), ('книга', 0.0027), ('лицо', 0.0027), ('нужный', 0.0025), ('женщина', 0.0025), ('доктор', 0.0024), ('дверь', 0.0023), ('последний', 0.0023)
	34	3%	2%	('жизнь', 0.0034), ('земля', 0.0033), ('вода', 0.0032), ('лицо', 0.0030), ('город', 0.0030), ('чёрный', 0.0027), ('море', 0.0026), ('ветер', 0.0025), ('белый', 0.0025), ('ночь', 0.0025)
3-1000	20	11,5%	2,5%	('глаз', 0.0068), ('пойти', 0.0042), ('лицо', 0.0039), ('сидеть', 0.0031), ('мужик', 0.0029), ('спросить', 0.0028), ('голос', 0.0028), ('голова', 0.0026), ('товарищ', 0.0026), ('земля', 0.0026)
	26	3,1%	1%	('товарищ', 0.0061), ('рабочий', 0.0038), ('город', 0.0029), ('председатель', 0.0029), ('глаз', 0.0028), ('лицо', 0.0027), ('ответить', 0.0025), ('голос', 0.0024), ('спросить', 0.0022), ('гражданин', 0.0022)
	10	2,9%	0,8%	('глаз', 0.0054), ('вода', 0.0052), ('лес', 0.0047), ('ночь', 0.0041), ('пойти', 0.0032), ('земля', 0.0032), ('лицо', 0.0030), ('белый', 0.0030), ('голова', 0.0029), ('снег', 0.0028)
	29	1,7%	0,1%	('любить', 0.0059), ('жена', 0.0059), ('жизнь', 0.0057), ('лицо', 0.0048), ('женщина', 0.0047), ('глаз', 0.0046), ('любовь', 0.0043), ('смотреть', 0.0041), ('нужный', 0.0040), ('муж', 0.0040)
	23	1,4%	0,2%	('жизнь', 0.0047), ('глаз', 0.0040), ('земля', 0.0040), ('мир', 0.0036), ('профессор', 0.0030), ('тело', 0.0029), ('дом', 0.0028), ('ко3мната', 0.0027), ('мысль', 0.0026), ('стена', 0.0024)
	2	1,2%	0,2%	('жизнь', 0.0037), ('книга', 0.0033), ('город', 0.0027), ('последний', 0.0027), ('русский', 0.0024), ('остров', 0.0023), ('нужный', 0.0020), ('вода', 0.0020), ('лицо', 0.0018), ('ответить', 0.0018)

Таблица III: Наиболее распространенные топики периода 1923–1930
(для объемов выборки в 100, 500 и 1000 рассказов)