Вестник Томского государственного университета. Филология. 2024. № 89. С. 127–151
Tomsk State University Journal of Philology. 2024. 89. pp. 127–151

Научная статья
УДК 81`32
doi: 10.17223/19986645/89/6

Тематическое моделирование художественной прозы: оценка и интерпретируемость результатов (на примере русского рассказа 1900–1930 гг.)

Татьяна Юрьевна Шерстинова[1], Маргарита Александровна Кирина[2], Анна Денисовна Москвина[3]

[1, 2, 3] Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия
[1] tsherstinova@hse.ru
[2] mkirina@hse.ru
[3] admoskvina@hse.ru

Аннотация. В статье рассматривается интерпретируемость результатов тематического моделирования литературных текстов. Цель исследования – определить, насколько тематические распределения отражают содержательные аспекты художественного текста. Описаны эксперименты по оценке тематических моделей, основанных на 3 000 рассказах 927 русских писателей начала XX в. Исследование показало, что 52% рассказов хорошо соответствуют семантически целостным топикам, а 24% им соответствуют частично. Полученные результаты подтверждают целесообразность применения методов тематического моделирования к художественным текстам.

Ключевые слова: русская литература, русский рассказ, малая проза, тематическое моделирование, литературная тема, корпусная лингвистика, компьютерная лингвистика, интерпретация, экспертная оценка

Источник финансирования: 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)» в 2022 г.

Благодарность: авторы выражают благодарность членам Научно-учебной группы междисциплинарных филологических исследований НИУ ВШЭ в Санкт-Петербурге, в особенности А.С. Карышевой, Е.О. Колпащиковой, А.Ю. Москаленко, И.А. Делазари, И.С. Завьяловой, за активное участие и экспертизу на разных этапах исследования.

Для цитирования: Шерстинова Т.Ю., Кирина М.А., Москвина А.Д. Тематическое моделирование художественной прозы: оценка и интерпретируемость результатов (на примере русского рассказа 1900–1930 гг.) // Вестник Томского государственного университета. Филология. 2024. № 89. С. 127–151. doi: 10.17223/19986645/89/6

Original article
doi: 10.17223/19986645/89/6

Topic modeling of prose fiction: Model assessment and interpretability (the case of Russian short stories of the 1900s–1930s)

Tatiana Yu. Sherstinova[1], Margarita A. Kirina[2], Anna D. Moskvina[3]

[1, 2, 3] National Research University Higher School of Economics, St. Petersburg, Russian Federation
[1] tsherstinova@hse.ru
[2] mkirina@hse.ru
[3] admoskvina@hse.ru

Abstract. The article presents two experiments investigating the interpretability of results obtained from automatic topic modeling of literary texts, addressing the broader question of the appropriateness of applying this method to fiction. The relevance of this research is grounded in the successful application of topic modeling to specialized texts, contrasted with the challenges posed by the metaphorical language and thematic complexity of literary works. The study aims to determine how and to what extent the topic distributions produced by the model (story–topic correlations) reflect the thematic aspects of short stories. The research material consisted of 3,000 short stories written by 927 Russian authors, including world-renowned figures such as Nobel laureates I.A. Bunin and M.A. Sholokhov, Russian "classics" like A.P. Chekhov and M. Gorky, as well as lesser-known and nearly forgotten writers. During the study, several samples were generated for each of the three chronological periods: (1) the beginning of the 20th century (1900–1913), (2) the era of wars and revolutions (1914–1922), and (3) the early Soviet period. Each period was represented by three samples consisting of 100, 500, and 1,000 short stories. For each sample, models were constructed using the LDA algorithm with various preprocessing options. The evaluation of topic interpretability was conducted in two stages. The first stage aimed to identify which preprocessing steps yielded the best interpretability of the resulting model. The model trained on the corpus without any POS filtering exhibited the highest interpretability, with 25% of the generated topics deemed interpretable by experts. In the second stage, only those 24 topics that were unanimously considered interpretable by all three experts were further analyzed. During the second experiment, two experts read all 127 texts from the resulting sample and evaluated each topic on a three-point scale: (1) fully corresponds, (2) partially corresponds, (3) does not correspond at all. The experiments revealed that the experts identified a good correspondence between the text and the automatically assigned topic in 52% of the short stories, partial correspondence in 24%, while the remaining 24% of the stories appeared unrelated to the assigned topic. Thus, 76% of all interpretable topics demonstrated a meaningful connection to the content of the stories, beyond being merely statistically significant word clusters within the texts. These results are quite promising and suggest that topic modeling can be effectively applied to fiction, allowing researchers to accurately identify typical themes within a collection of short stories without needing to read them all. However, achieving these results requires a careful preliminary selection of the topics generated by the model, ensuring their semantic coherence, as done in the first stage of the experiment.

Keywords: Russian literature, Russian short story, short fiction, topic modeling, literary theme, corpus linguistics, computational linguistics, interpretability, expert assessment

Acknowledgements: The publication was prepared as a result of the research conducted under Project No. 21-04-053 within the framework of the Scientific Foundation Program of the National Research University Higher School of Economics (HSE) in 2022. The authors express their gratitude to the members of the Scientific and Educational Group of Interdisciplinary Philological Research of the National Research University Higher School of Economics in St. Petersburg, especially to A.S. Karysheva, E.O. Kolpashchikova, A.Yu. Moskalenko, I.A. Delazari, I.S. Zavyalova, for their active participation and expert assessment at different stages of the research.

For citation: Sherstinova, T.Yu., Kirina, M.A. & Moskvina, A.D. (2024) Topic modeling of prose fiction: Model assessment and interpretability (the case of Russian short stories of the 1900s–1930s). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology.* 89. pp. 127–151. (In Russian). doi: 10.17223/19986645/89/6

Тематическое моделирование литературных текстов и проблема интерпретации его результатов

Тема как объект филологических исследований представляет интерес прежде всего своей неоднозначностью. Трудности, возникающие при попытке сформулировать тему произведения, во многом связаны со специфичностью того художественного преобразования, которое претерпевает объективная реальность, становясь в той или иной форме основой сюжета литературного произведения. По сути, текст, как и выбор составляющих его тематических элементов, является результатом осуществляемой автором творческой интерпретации действительности. Отсюда вырастает проблема многозначности литературного текста: как правило, тема произведения не единственна и способна к дроблению, например до подтем, соответствующих отдельным частям произведения. В то же время не в полной мере ясным представляется то, как оптимально сформулировать тему произвольного художественного текста, уже не говоря о ее однозначности. Существуют мнения как о необходимости выделения одной темы, являющейся главной для всего текста, так и нескольких – характеризующих его составляющие. Предполагается, что формальные тематические модели могут стать инструментом, который позволит уравновесить расхождения в плане понимания того, что является темой художественного текста, обнаружив ее через языковые и стилистические особенности текстов, которые могут указывать на ряд сюжетных и тематических отличительных характеристик литературного корпуса.

Тематическое моделирование представляет собой метод машинного обучения, использующийся для категоризации больших неструктурированных текстовых данных и применяющийся главным образом к специальным текстам (научным, публицистическим, новостным и др.). Тематическая модель,

обученная на коллекции текстов, назначает каждому тексту соответствующие ему с определенной вероятностью темы (англ. topics), которые формируются автоматически, а также всем словам – вероятности попадания в определенную тему. Таким образом, темой в тематическом моделировании является группа слов, ее образующих. Далее мы будем называть такие автоматически сформированные темы «топиками» во избежание неоднозначности и для различения их от «тем» литературного произведения, выделяемых экспертным образом [1].

В последние годы рядом авторов предпринимались попытки тематического моделирования художественных текстов на разных языках – английском, русском, испанском, французском, голландском [2–5]. Так, например, в работе О.А. Митрофановой [6], посвященной автоматическому анализу романа М.А. Булгакова «Мастер и Маргарита», обученная в рамках исследования тематическая модель (LDA) смогла противопоставить две сюжетные линии (история Иешуа и Понтия Пилата, история Мастера), а также объединила в отдельные топики имена сюжетно близких персонажей.

Что касается моделирования тематики малой русской прозы, целая серия экспериментов была проведена на материале Корпуса русского рассказа (1900–1930 гг.) [7] – электронного ресурса, разрабатываемого специально для проведения компьютерных исследований языка и стиля художественных текстов [8–10]. Динамическое тематическое моделирование для подкорпуса из 310 текстов, принадлежащих перу 300 разных авторов, позволило выделить «нишевые» топики для разных исторических периодов [11]; далее на этом материале было проведено сравнение полученных топиков с данными ручной тематической разметки [1, 12, 13]. Оказалось, что лучше всего и, что важно, приближенно к экспертной оценке выделяются те тематические элементы, которые составляют фон действия произведения [1]. Также на этом материале было проведено исследование, посвященное теме насилия в русском рассказе, оказавшейся довольно частотной для прозы начала XX в. [14] и эксперименты, направленные на сравнение разных методов тематического моделирования в применении к текстам малой прозы [15].

Однако использование методов тематического моделирования на материале литературных текстов, в особенности поэтических, осложняется богатством, метафоричностью, образностью художественного языка [16], что приводит к тому, что распределения топиков, полученные в результате моделирования художественных текстов, не столь однозначны и убедительны, как топики, построенные на материале специальных текстов. Можно выделить следующие основные проблемы, с которыми сталкиваются исследователи литературных произведений методами тематического моделирования:

1. Сложности интерпретации семантических отношений, связывающих слова в получившихся топиках.

2. Трудности соотношения целей и результатов тематического моделирования: выявляемые топики порой характеризуют не непосредственно темы произведений, а общие сюжетные элементы, мотивные структуры, предметы описания и т.д.

3. Многотемность и «эзопов язык» многих художественных текстов приводят к тому, что полноценное тематическое описание не всегда возможно; как правило, топики выражают только наиболее общие темы, в то время как целые тематические пласты остаются вне фокуса тематической модели.

Кроме того, отдельной проблемой считается сам подход к обработке текста при тематическом моделировании, условно именуемый мешком слов, поскольку он не учитывает позицию слов в предложении – их синтаксические, контекстуальные и семантические характеристики [17. P. 611]. Результат тематического моделирования в таком виде сводится к получению «некоторой информации о содержании корпуса текстов» [18]. Таким образом, часто критике подвергается сама идея полноценного литературного анализа с применением методов тематического моделирования поскольку природа художественных произведений порождает ряд трудностей для интерпретации только на основе получившихся топиков.

Говоря о том, что топики отражают темы, сами исследователи подчеркивают, что оба понятия выполняют своего рода роль посредника, указывающего на такой «тип литературного содержания, который семантически унифицирован и повторяется с некоторой частотностью или регулярностью во всем корпусе» [2. P. 752]. Так или иначе, утверждается, что слова, формирующие кластеры, являются тематическими по своей природе, а определение того, какие из них, действительно, составляют тему, представляет собой задачу, решаемую при интерпретации. И хотя допускается наличие вариаций на этом этапе, человеку свойственно различать тематически более крупные категории, которым данные распределения слов соответствуют [2. P. 752]. В этом смысле интерпретация топиков сводится к выводу из них тематических элементов на некотором «усредненном» уровне абстракции [19. С. 49]. Оценка качества тематической модели через ее интерпретируемость напоминает принцип свертывания и развертывания тем через установление логического соответствия между темой и текстом, хотя последний и подразумевает вывод соответствий после непосредственного прочтения.

Однако в отношении тематического моделирования исследователь сталкивается с тем, что те абстрактные категории, на которые указывают топики, не обязательно отражают тематику произведения – не только потому, что образный язык затрудняет понимание содержания топиков, но и потому, что выявляемые тематические кластеры оказываются более разнообразны. С другой стороны, топики могут включать также информацию о месте и времени действия, различных художественных и стилистических приемах и т.д. По этой причине оценка качества тематических моделей литературных текстов, несмотря на возможность привлечения различных мер качества, должна главным образом осуществляться экспертно [18].

Вопросы интерпретируемости моделей, построенных на «больших» литературных данных, осложняются тем фактом, что у нас в большинстве случаев не может быть получено референтного экспертного мнения о том, насколько структура построенной модели согласуется с исходными данными – хотя бы просто потому, что ни один литературовед не способен за

ограниченное экспериментом время прочитать те тысячи текстов, которые были предложены модели для обработки. В этом случае единственно возможными критериями оценки модели становятся: 1) проверка успешности ее работы для конкретных текстов и 2) оценка полученных результатов с точки зрения некоторых общих законов генеральных совокупностей (подобно тому, как нормальность распределения данных в ряде случаев можно считать одним из критериев качественной выборки). Однако на текущем уровне развития методов тематического моделирования и с учетом относительно небольшого опыта применения их к текстам художественной литературы не существует количественных мер, которые позволяли бы однозначно определять, насколько удачно построенная модель отражает тематическое распределение исходного набора текстов.

Поэтому в данном исследовании мы будем придерживаться первого пути – оценки интерпретируемости модели для отдельных текстов в надежде на то, что полученные нами результаты в конечном итоге будут способствовать лучшему пониманию свойств произвольного набора литературных текстов как исследуемой генеральной совокупности.

При анализе тематических моделей будут использованы как термин «топик», так и термин «тема». Первый термин – «топик» – более технический, отсылает к результатам тематического моделирования, т.е. к тем тематическим кластерам, которые были определены автоматически; второй – непосредственно к темам, определяемым экспертно, если выделение таковых представляется возможным. Как будет показано далее, топик не всегда будет точно описывать тематику литературного произведения; более того, не всегда на основе одного топика выделяется только одна тема – такова особенность материала, поэтому предложенное различение терминов представляется необходимым.

Материал и методика исследования

Данное исследование продолжает серию экспериментов по тематическому моделированию малой русской прозы, выполняемых на материале Корпуса русского рассказа 1900–1930 гг. [6]. Рассказ является наиболее распространенным прозаическим жанром, что позволяет привлекать к исследованию большое количество авторов и их текстов, а тематическое и стилистическое разнообразие рассказов делает их прекрасным материалом для изучения разнообразия тематики литературных произведений [7].

Анализируемый временной период – первые три десятилетия XX в. – представляет для нашей страны эпоху, насыщенную драматическими событиями и социальными преобразованиями. Все входящие в корпус рассказы разделены по дате их написания на три хронологических периода, которые соотносятся со значимыми для страны историческими эпохами [8]: период I – начало XX в. и предвоенные годы (1900–1913 гг.); период II – эпоха войн и революций: Первая мировая война, Февральская и Октябрьская ре-

волюции и последующая за ними Гражданская война (1914–1922 гг.); период III – эпоха становления молодого Советского государства с окончания Гражданской войны до 1930 г. В данной работе мы также будем придерживаться этого разделения.

Другой важной особенностью Корпуса русского рассказа является то, что он создавался не только для проведения разнообразных исследований прозы [7–8], но также и для моделирования «литературно-художественной системы» рассматриваемой эпохи [20–22]. Такой подход подразумевает, что в исследование включают тексты не только всем известных классиков, но и литераторов «второго эшелона», а также малоизвестных и фактически забытых писателей, не отдавая предпочтения отдельным писателям или литературным направлениям [9].

Для проведения этого исследования было решено выйти за пределы аннотированного корпуса, содержащего 310 рассказов 300 разных писателей, существенно (почти в 10 раз) увеличив объем текстового материала. Однако наращение материала было решено осуществлять постепенно – построить модель сначала для 100, потом для 500 и, наконец, для 1 000 текстов для каждого исторического периода, для того чтобы понять, насколько сильно изменяются выделяемые топики и соответствующие им темы при изменении объема выборки. Тем самым косвенным образом решалась задача оценки «сходимости» тематической модели, т.е. ответа на вопрос, возможно ли получить некое обобщенное тематическое распределение, которое уже не будет существенно меняться при увеличении объема выборки.

При отборе текстов для выборок мы старались придерживаться принципа максимальной представительности разных писателей – для выборок в 100 текстов все тексты написаны разными авторами, для выборок большего объема возможны незначительные повторы (по нескольку текстов от одного писателя), что объясняется отсутствием оцифрованных текстов для многих малоизвестных авторов. Материалом для исследования стали тексты, написанные 927 разными авторами, среди которых есть как всемирно известные писатели – нобелевские лауреаты по литературе (И.А. Бунин и М.А. Шолохов), русские «классики» (А.П. Чехов, М. Горький, А. Белый, К.Д. Бальмонт, В.Я. Брюсов, А.А. Блок, Ф.К. Сологуб, И.Э. Бабель, А.С. Серафимович и др.), менее известные современному читателю литераторы (Н.А. Жаринцова, А.С. Гингер, И.Е. Вольнов, В.А. Мазуркевич, Н.А. Лухманова и др.) и почти забытые в наши дни имена (И.М. Вершинин, Д.И. Крутиков, В.Я. Кокосов, О.И. Слицан и др.). Кроме того, мы старались обеспечить относительно равномерное распределение рассказов по году их написания. Конкретные тексты от каждого писателя отбирались для выборок в случайном порядке, безотносительно к тематике и содержанию. Однако в выборки было решено не включать как достаточно крупные рассказы (больше 10 000 словоупотреблений), так и очень краткие (менее 200 слов). Все отобранные тексты подлежали анализу целиком, вне зависимости от их размера.

Было сформировано 9 выборок: по 100, 500 и 1 000 рассказов в трех временных периодах (1900–1913 гг., 1914–1922 гг., 1923–1930 гг.), при этом выборка в 100 рассказов присутствует и в выборке на 500 текстов, а 500, соответственно, в 1 000. Объемы каждой из выборок в текстах и словоупотреблениях представлены в табл. 1.

Таблица 1

Объемы 9 выборок для проведения исследования в текстах и словоупотреблениях

Параметры выборки	Объем выборки								
	I период (1900–1913)			II период (1914–1922)			III период (1923–1930)		
Код	I–100	I–500	I–1000	II–100	II–500	II–1000	III–100	III–500	III–1000
Тексты	100	500	1000	100	500	1000	100	500	1000
Слова	369113	1820446	3564493	321693	1355994	2702207	302682	1516696	2541865

Поскольку предобработка текстов оказывает значительное влияние на результирующую тематическую модель, было приняло решение провести ее в нескольких вариантах, с тем чтобы определить оптимальную. Все рассказы были лемматизированы с помощью библиотеки spacy [23], из текстов были удалены все имена собственные и стоп-слова. Дополнительно были использованы четыре варианта фильтрации: П1 – без частеречных фильтров, П2 – с фильтрами «существительное», «глагол», «прилагательное» (таким образом, при построении моделей были использованы только эти три части речи), П3 – с фильтрами «существительное» и «глагол», П4 – с единственным фильтром «существительное».

Из лемматизированных текстов удалялись стоп-слова по дополнительному словарю. Были собраны в единый токен биграммы, которые встречаются в текстах выборки не менее четырех раз. Из текстов также удалялись слова, встречающиеся менее чем в шести документах и более чем в 80% корпуса. Процедура обучения модели и извлечения списка топиков со словами и распределения рассказов по топикам повторялась для девяти выборок, предобработанных в четырех вариантах. Полученные тематические распределения были сохранены в 36 текстовых документов с топиками и таблицы с распределением рассказов по топикам.

Тематическое моделирование материала. В настоящее время разработано значительное количество как методов тематического моделирования (LSA, pLSA, STM, CTM, NMF и др.), так и их имплементаций (MALLET, BigARTM, Stanford Topic Modelling Toolbox, gensim, tomotopy и др.) [24–26]. Однако наиболее популярным является метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) в имплементации gensim [27].

Метод латентного размещения Дирихле, или LDA, представляет собой мультиноминальную порождающую (генеративную) вероятностную модель

[24, 28]. Согласно данной модели корпус текстов рассматривается как смесь случайных топиков, распределение которых соответствует размещению Дирихле [29. С. 226–228]. Каждый документ в коллекции описывается набором скрытых семантических структур – топиков, складывающихся из слов, составляющих его с некоторой степенью вероятности. На документ может приходиться неограниченное количество топиков, поэтому говорят, что «документы описываются семейством распределений тем» [26. С. 223]. Получается, что одна тема характеризует документ не только с некоторой долей вероятности, но и в пропорциональном соотношении может составлять большую или меньшую часть документа. Количество топиков, которое необходимо найти в текстовой коллекции, предопределено пользователем, с одной стороны, либо высчитано статистически – с другой, и равно k [30].

В нашем исследовании для построения тематических моделей использовался алгоритм LDA, реализованный в библиотеке gensim [27]. Оптимальное количество топиков для каждой выборки определялось автоматически. Количество топиков считалось оптимальным, если при нём когерентность оказывалась наиболее высокой. Для подсчёта метрики использовалась функция CoherenceModel из модуля models библиотеки gensim [31]. Диапазон количества топиков при вычислении когерентности – от 10 до 45; метрика вычислялась при построении моделей в цикле с шагом в 5 топиков. Из построенных моделей выбиралась модель с наибольшим значением когерентности. Рекомендуемое количество топиков для построения тематической модели, и их соответствие выборкам и временным периодам было описано в [32]. Именно с такими настройками извлекались распределение слов по топикам и распределение топиков по документам.

Автоматически выявленные топики представляют собой списки из десяти «ключевых» слов, отсортированных по убыванию вероятности их принадлежности к топику. Например, топик № 7 из модели, построенной для III-1000, представляет собой семантически однородное множество слов, с соответствующими вероятностями: *'корабль', 0.015305743; 'пароход', 0.014277308; 'море', 0.01387377; 'палуба', 0.010964184; 'вода', 0.010195338; 'ветер', 0.009966294; 'волна', 0.009912074; 'капитан', 0.008349794; 'берег', 0.00815364; 'океан', 0.0073074135*.

Абсолютные значения таких вероятностей (и даже порядок этих значений) варьируются от топика к топику, поэтому, к примеру, соотносить все первые слова топиков как в целом наиболее вероятные было бы не вполне корректно. В этой работе мы не даем каждому из топиков специального названия[1], а используем для идентификации лишь его порядковый номер в построенной модели, который, впрочем, должен быть отнесен скорее к номинальной шкале, чем к порядковой, т.е. величина номера топика никак не связана с его качеством и интерпретируемостью.

[1] Назначение меток топикам – это отдельная содержательная задача.

Помимо распределения слов по топикам тематическая модель назначает каждому тексту (в нашем случае, рассказу) и список таких топиков с соответствующими им вероятностями. То есть для каждого рассказа обычно выделяется несколько топиков. Например, для рассказа И.С. Соколова-Микитова «Морской ветер» из выборки III-1000 таких топиков получено три, но наибольшей вероятностью (0,628837) характеризуется только что рассмотренный топик №7. Соответственно, можно получить и список рассказов, в которых этот топик был обнаружен с относительно высокой вероятностью.

Результаты применения тематического моделирования для нашего материала и то, какие тематические модели можно считать наиболее типичными для русского рассказа начала XX в., были описаны в статье [33]. Эти аспекты в данной работе не рассматриваются. Сосредоточим свое внимание на вопросах оценки и интерпретируемости полученных моделей.

Методы оценки и интерпретации моделей. Для оценки адекватности модели представляется целесообразным учитывать два аспекта: во-первых, насколько интерпретируемы получившиеся топики и, во-вторых, как хорошо топики описывают коллекцию документов. С одной стороны, сделать это можно, используя ряд статистических мер, а с другой – привлекая экспертов. Среди статистических мер оценки качества тематических моделей наиболее популярными являются перплексия (*perplexity*) и когерентность (*coherence*) [33].

Среди экспертных подходов к оценке качества тематического моделирования, когда оценка согласованности модели и ее интерпретируемость опираются на суждения людей, можно отметить как непосредственное привлечение экспертом для чтения текстов и сопоставления результатов с представленными топиками, так и метод интрузий. Выделяют словесную интрузию (*word intrusion*) и тематическую (*topic intrusion*) [34]. Эта методика нацелена на поиск экспертом «лишних» слов или «лишних» топиков соответственно. Если данное слово или топик были верно определены экспертом, то это свидетельствует о более высокой связности модели, и наоборот.

Поскольку меры автоматической оценки часто оказываются несоотносимыми с понятием интерпретируемости моделей [18, 35], более важной при анализе художественных текстов выступает экспертная оценка, причем предполагающая непосредственное прочтение текстов. Это связано с тем, что тема в художественном произведении часто задается не в начале, а по мере развертывания сюжетных действий. Как правило, к процедуре подобной оценки привлекаются несколько человек, которым предлагается список критериев для оценки топиков на предмет их интерпретируемости, соотносимости с какой-то, по их мнению, тематической категорией. Для анализа художественных текстов кажется целесообразным использовать именно такую оценку тематической модели, поэтому она была задействована в описываемом нами эксперименте. Метод интрузий к исследованию не привлекался, так как он не позволяет в полной мере оценить соответствие автоматически полученных топиков собственно содержанию исследуемых текстов.

В нашем исследовании экспертная оценка интерпретируемости топиков осуществлялась в два этапа, которые далее будут рассмотрены более подробно. Цель первого этапа состояла в том, чтобы определить, какой из вариантов предобработки текстового материала позволяет получить модель с лучшими интерпретационными свойствами, с тем чтобы именно эту модель использовать на втором этапе эксперимента. Эксперты работали только с конечными топиками, без обращения к самим текстам рассказов, а оптимальным вариантом предобработки был выбран тот, который позволил дать максимальную долю интерпретируемых топиков.

На втором этапе исследования были отобраны топики, которые были сочтены интерпретируемыми всеми тремя экспертами, после чего уже другими экспертами, с привлечением литературоведа, оценивалось их соответствие конкретным рассказам по трехуровневой шкале: 1) полностью соответствует; 2) частично соответствует; 3) абсолютно не соответствует.

Рассмотрим полученные результаты.

Экспертная оценка интерпретируемости топиков

Экспертами на первом этапе эксперимента были студенты-филологи, задачей которых ставилась параллельная независимая оценка интерпретируемости всего множества топиков, полученных для всех 36 построенных моделей. Эксперты просматривали получившиеся темы без обращения к текстам рассказов и должны были оценить каждый из топиков по бинарной шкале: интерпретируемое vs неинтерпретируемое. Параллельно с одним и тем же материалом работали три эксперта, что позволило получить по три независимых оценки для каждого топика. Экспертам предписывалось считать топик интерпретируемым, если он проходит хотя бы по одному из следующих критериев:

1. *Семантическая связность слов (термов), составляющих топик:* если абсолютное большинство слов, образующих топик, относится к одному и тому же семантическому полю или к смежным семантическим полям, то топик считается интерпретируемым. Пример такого топика, относящегося к морской тематике, был приведен выше. При этом допускалось, что 1–2 слова топика могут выбиваться из общего ряда, если в остальном предполагаемая тематика эмпирически узнаваема. По такому критерию, например, можно считать интерпретируемым топик № 2 выборки I-100, в котором явственно прослеживается тема смерти и, возможно, похорон, но при этом присутствует и одно «лишнее» слово – 'игра', которое не кажется его логичной частью: *'гроб', 0.00398, 'склеп', 0.00351, 'игра', 0.00182, 'страшный', 0.00181, 'пожар', 0.00153, 'покойник', 0.00148, 'страх', 0.00127, 'ужас', 0.00126, 'испытать', 0.00122, 'ночь', 0.00123.*

2. *Нарративная связность термов, составляющих топик:* если на основе составляющих топик слов можно построить правдоподобный рассказ или фрагмент рассказа, т.е. выстроить некий нарратив. Например, по этому

критерию один из экспертов оценил топик № 14 выборки II-1000 как интерпретируемый, поскольку из составляющих его слов можно построить вполне связное повествование о семье, которая незадолго до Рождества присутствует на концерте или балете: *'ёлка', 0.04234, 'палец', 0.03037, 'пианист', 0.01558, 'ребёнок', 0.01383, 'ёлочка', 0.01348, 'актриса', 0.01049, 'девочка', 0.00988, 'сочельник', 0.00958, 'мамочка', 0.00954, 'балет', 0.00818.*

3. *Высокая тематическая плотность словарного состава топика:* если слова обладали высокой тематической связностью (т.е. эксперт как опытный читатель мог ассоциировать слова с определенной тематикой художественного текста), то топик принимался как интерпретируемый. При этом из общего семантического ряда допускалось выпадение до 3–4 слов с относительно невысокими вероятностями. Поясним на примере топика № 12 из выборки III-500: *'девушка', 0.007, 'быт', 0.005, 'мат', 0.004, 'ребята', 0.003, 'мода', 0.003, 'советский', 0.003, 'стыд', 0.003, 'губа', 0.003, 'дискуссия', 0.003, 'вопрос', 0.003.* Большинство слов топика можно условно отнести к некой бытовой конфликтной ситуации советского периода, при этом 'губа' и 'мода' несколько выпадают из этой темы. Впрочем, если вероятность таких нерелевантных слов оказывалась выше релевантных, то топик к интерпретируемым не относился.

Выделение дополнительных критериев для оценки интерпретируемости топиков кажется необходимым ввиду специфики материала (как правило, чаще всего используется первый метод – оценка семантической связности термов). При анализе художественных произведений исследователь часто сталкивается с проблемой лексической неоднозначности, с одной стороны, и сюжетности – с другой. Как следует из [15] и [14], топик может объединять произведения на уровне общих сюжетных элементов и литературных приемов. Например, тема смерти может быть выражена как «гаснущее сознание», а революция и постановка спектакля – как игра.

Насколько разные варианты предобработки текстов повлияли на итоговые результаты исследования, было подробно описано в статье [32]: с точки зрения процента наиболее интерпретируемых топиков наиболее удачным (в среднем 25% топиков) оказался вариант предобработки без частеречных фильтров. Результаты именно этой модели стали материалом для следующего эксперимента.

Интерпретируемые топики и их соответствие литературному тексту

Ключевой задачей исследования стало выяснение того, в какой мере топики, признанные всеми тремя экспертами как интерпретируемые, соответствуют тому литературному материалу, который призваны описывать. Тематическая модель с предобработкой без частеречных фильтров дала в совокупности 265 топиков. Из всего этого множества автоматически полученных топиков только 24 (или 9%) были признаны интерпретируемыми по своему лексическому наполнению. Для каждого из этих топиков были отобраны документы (рассказы) с наибольшей степенью вероятности попадания относиться

к теме, выраженной этим топиком[1]. В результате было отобрано 127 документов (122 уникальных рассказа)[2], которые оценивались на степень соответствия топику. В эксперименте приняли участие два эксперта-литературоведа, которые должны были прочитать все 127 текстов[3] и оценить соответствие назначенным им топикам по трехуровневой шкале: 1) полностью соответствует; 2) частично соответствует; 3) абсолютно не соответствует.

При этом оценке подлежало не столько наличие слов топика в самом тексте литературного произведения, сколько смысловое соответствие топика содержанию рассказа. То есть если слова топика ожидаемо предполагают некоторую тему, но суть рассказа ей не соответствует, то этот топик следовало считать не соответствующим рассказу, даже если в нем присутствуют все слова, образующие содержание топика. Например, описанный топик № 7 ассоциируется с каким-то морским плаванием, при этом если в тексте «море» и другие слова топика упоминаются, но не являются значимыми для сюжета или даже противоречат ему, топик оценивался как не соответствующий тексту.

Содержание всех проанализированных 24 топиков и рассказы, к которым эти топики относятся, приведены в табл. 2[4]. Любопытно распределение интерпретируемых топиков по периодам: в самый ранний, довоенный, период наблюдается максимальное количество этих топиков – 12, в военно-революционный – 7, а в раннесоветский – всего 5. Это косвенным образом свидетельствует о том, что тексты, написанные в начале XX в., являются лексически и, соответственно, тематически более разнообразными.

Таблица 2

Интерпретируемые топики и количество предписанных им текстов (N)

Выборка	Интерпретируемые топики		N
	№	Содержание	
I-100	0	село, мужик, народ, поле, ребёнок, старый, гора, дом, вода, церковь	7
	11	озеро, вода, лес, берег, крик, крыло, прелесть, рассвет, безумный, взгляд	1
	14	мир, век, мечта, вера, душа, народ, великий, дух, верить, умирать	1
	15	смерть, страшный, тихий, последний, тело, мужик, сердце, охота, умирать, душа	3

[1] Напомним, что каждому топику – ввиду выполняемой в ходе тематического моделирования «мягкой» кластеризации – соответствуют все рассказы исследуемой коллекции с некоторой степенью вероятности.

[2] Поскольку выборки были вложенными (рассказы из выборок меньшего объема включались в выборки большего объема), одни и те же рассказы могли составить документы топиков, полученных в результате тематического моделирования смежных выборок одного периода.

[3] Ввиду большого объема текстов каждый рассказ был прочитан только одним экспертом.

[4] Для сокращения объема слова приводятся в таблице без соответствующих вероятностей.

Выборка	Интерпретируемые топики		N
	№	Содержание	
I-500	4	лебедь, озеро, вода, лес, лебеди, прелесть, берег, рассвет, крыло	1
	5	красивый, мак, любовь, красный, хмель, вид, песня, старый, звук, полянка	1
	35	шапка, дьякон, арестант, казак, земля, тюрьма, дядя, степь, смотреть, надзиратель	10
I-1000	2	гроб, склеп, игра, страшный, пожар, покойник, страх, ужас, испытать, ночь	1
	5	любить, жена, ребёнок, сидеть, дом, муж, спросить, комната, отец, минута	9
	8	матрос, старший_офицер, доктор, благородие, капитан, боцман, каюта, тюрьма, вашескобродие, смотритель	10
	11	вагон, поезд, станция, полковник, паровоз, платформа, машинист, поручик, роза, телеграфист	10
	12	солдат, офицер, поручик, рота, пуля, японец, полк, выстрел, раненый, благородие	10
II-100	2	карта, жена, клуб, игра, стол, деньга, час, рубль, кабинет, сидеть	1
	33	солнце, весенний, весна, нежный, яркий, душа, память, сверкать, луч, тепло	1
	34	стоять, час, бутылка, минута, улыбаться, чашка, здоровье, приятель, любить, пить	1
II-500	8	господь, рай, земля, угодник, святой, весёлый, бить, спросить, смотреть	2
	37	солдат, немец, пленник, полковник, лес, товарищ, спросить, сторона, рота, начать	10
II-1000	5	товарищ, рабочий, толпа, гражданин, город, улица, совет, партия, владыка, москва	10
	14	солдат, казак, немец, окоп, офицер, генерал, раненый, поручик, бой, винтовка	10
III-100	3	командир, комиссар, фронт, красный, земля, штаб, лететь, вагон, полк, работа	5
	21	кружок, выступление, победа, слава, молодёжь, вечер, последний, клуб, школа, предложить	1
III-500	23	яйцо, мать, потомство, икра, детёныш, жизнь, рыба, развиваться, матка, самка	5
	41	разведчик, дозор, рота, взвод, командир, противник, неизвестный, ребят, позиция, пулемёт	7
III-1000	7	корабль, пароход, море, палуба, вода, ветер, волна, капитан, берег, океан	10
Итого			127

Результаты экспертной работы выявили следующие процентные соотношения документов к общему количеству документов по степени их соответствия заявленной тематике: для 66 документов (52%) обнаруживается полное соответствие топику, для 30 документов (24%) – частичное и для 31 (24%) соответствие отсутствует (табл. 3). Что касается распределения по периодам, хотя бы частичное соотнесение топика и текстов оказалось возможным, вне зависимости от объема выборки, для 78% документов первого периода, 80% документов второго периода и 57% документов третьего периода.

Таблица 3

Распределение количества документов по разным вариантам соответствия с топиками

Выборка: период-объем	Количество анализируемых документов	Соответствие теме					
		Полное	% от количества отобранных документов	Частичное	% от количества отобранных документов	Полное несоответствие	% от количества отобранных документов
I–100	12	9	75	2	17	1	8
I–500	12	3	25	5	42	4	33
I–1000	40	27	68	4	10	9	23
II–100	3	3	100	0	0	0	0
II–500	12	4	33	5	42	3	25
II–1000	20	11	55	5	25	4	20
III–100	6	4	67	2	33	0	0
III–500	12	2	17	2	17	8	67
III–1000	10	3	30	5	50	2	20
Итого	127	66		30		31	

Для большинства рассказов было обнаружено соответствие тематики, выявленной автоматически, и той, что определяется путем непосредственного прочтения текста. Как правило, эти тексты имеют высокую степень вероятности попадания (0,7–0,9), но встречаются также и рассказы, которые, несмотря на относительно низкую вероятность, тоже можно считать совпадающими с выявленными топиками: например, это рассказ В.Д. Хасидовича «Огоньки» с вероятностью 0,3 или произведение А.С. Новикова-Прибоя «Пошутили» (0,4).

Определённые для большинства рассказов топики позволяют определить темы рассказов и отчасти предсказать происходящее в текстах без их прочтения. Один из таких – «Материнство в царстве животных» С.В. Покровского, где подробно описывается процесс рождения рыб, муравьёв, жуков и других животных; соответственно, выделенные для него слова (*'яйцо', 'мать', 'потомство', 'икра', 'детёныш', 'жизнь', 'рыба', 'развиваться', 'матка', 'самка'*) отлично коррелируют с содержанием рассказа и могут использоваться для предварительных выводов о его тематике. Другой пример – рассказ А.П. Матвеева «Красные маки»: по словам *'красивый', 'мак', 'любовь', 'красный', 'хмель', 'вид', 'песня', 'старый', 'звук', 'полянка'* можно понять, о чём будет текст: повествователь, глядя на маки, описывает свои чувства.

Тем не менее такая предсказуемость характерна далеко не для всех рассказов. В некоторых текстах действительно присутствуют все выделенные

методом тематического моделирования слова, но точно проинтерпретировать содержание по ним нельзя. Например, рассказу С.Г. Петрова «В склепе» соответствуют слова *'гроб', 'склеп', 'игра', 'страшный', 'пожар', 'покойник', 'страх', 'ужас', 'испытать', 'ночь'*, которые позволяют составить представление о происходящем в тексте, однако слово «пожар» вводит в заблуждение: можно предположить, что в рассказе кто-то погиб из-за пожара или что пожар произошёл в склепе. На самом деле это слово используется для описания характера одного из героев, который любил пожары, в то время как на сюжет рассказа непосредственно пожар не влияет. Примечательно, что этот топик оказался дискутабельным и на первом этапе эксперимента (см. выше), но тогда экспертами отмечалась нерелевантность слова «игра».

Еще один типичный сложный случай можно представить на примере рассказа В.В. Брусянина «На лыжах», который был отнесен моделью к топику со следующим содержанием: *'товарищ', 'рабочий', 'толпа', 'гражданин', 'город', 'улица', 'совет', 'партия', 'владыка', 'Москва'*. Действительно, все приведённые слова, кроме двух последних, присутствуют в тексте и отражают тематику происходящего. Однако в этом в рассказе говорится о революционных событиях и внутрипартийных отношениях в Финляндии, когда она еще входила в состав Российской империи. Представляется, что для раскрытия сюжета принципиально важно было бы указать место действия рассказа, определить которое из этих слов не невозможно. Более того, возникает ложная гипотеза, что действие рассказа как-то связано с Москвой.

Также интересно посмотреть на рассказ Б.Л. Тагеева «Жемчужный паук». Ему была приписана тема 7 (III-1000): *'корабль', 'пароход', 'море', 'палуба', 'вода', 'ветер', 'волна', 'капитан', 'берег', 'океан'*. Несмотря на то, что они частично совпадают со словами, присутствующими в тексте, рассказ нельзя считать соответствующим топику, так как произведение не связано с морскими путешествиями, как можно было бы ожидать, а относится скорее к производственной тематике: в нём повествуется о японских ныряльщицах, труд которых эксплуатируется капиталистами. Поэтому топик был признан «частично соответствующим» рассказу.

Кроме того, выделенные для некоторых рассказов слова не всегда присутствуют в тексте. Однако это не мешает определению общей тематики произведения. Например, для рассказа И.В. Гриневской «Старушка» топик составляют следующие вероятные слова: *'любить', 'жена', 'ребёнок', 'сидеть', 'дом', 'муж', 'спросить', 'комната', 'отец', 'минута'*. Между тем в тексте произведения присутствуют лишь те из них, которые соответствуют описанию жизни героини с мужем, их быта. Тем не менее этого достаточно для совпадения просматриваемой в топике общей семейной тематики с тем, что определяется как содержание рассказа при знакомстве с его полным текстом.

Помимо удачных соответствий выделенных слов рассказам, встречаются как менее подходящие варианты, так и случаи полного «непопадания» топика, определенного при тематическом моделировании как вероятного, в

содержание рассказа. Даже если вероятность попадания рассказа в тему высока, иногда они никак не коррелируют друг с другом и производят впечатление случайного попадания в топик. Например, рассказ Е. Гуро «В парке» представляет собой бессюжетную зарисовку о дачной прогулке, в то время как слова топика [*'вагон', 'поезд', 'станция', 'полковник', паровоз', 'платформа', 'машинист', 'поручик', 'роза', 'телеграфист'*] абсолютно не соответствуют его содержанию, а значение вероятности высоко (0,9). То же самое можно сказать про рассказ С.С. Юшкевича «Как живет и работает Семен Юшкевич»: согласно модели этот текст с вероятностью 0,9 попадает в топик [*'разведчик', 'дозор', 'рота', 'взвод', 'командир', 'противник', 'неизвестный', 'ребят', 'позиция', 'пулемёт'*], хотя на самом деле он никак не связан с военными действиями – в тексте описывается процесс интервью, которое берут у рассказчика. Можно высказать предположение, что такое яркое несоответствие топика рассказу связано, скорее всего, с его длиной: слишком короткие тексты попадают в топик как вероятные из-за случайного присутствия в них хотя бы одного слова из этого топика и рассчитываются как высоковероятные исключительно благодаря тому, что в рассказе мало других слов.

В качестве примера того, насколько разные по тематике рассказы могут быть приписаны к одному топику, рассмотрим два произведения, отнесенных моделью к топику [*'шапка', 'дьякон', 'арестант', 'казак', 'земля', 'тюрьма', 'дядя', 'степь', 'смотреть', 'надзиратель'*] с вероятностями 99 и 34% соответственно. В рассказе З.Б. Осетрова «Казачий пикет» сюжет построен на повествовании о взаимоотношениях казаков; в центре повествования – *шапка*, которую один казак хочет купить у другого, никаких тюремных мотивов в тексте не просматривается. А рассказ П.Ф. Якубовича «Любимцы каторги», напротив, содержит другую часть слов топика, как раз связанную с темой тюрьмы: речь в нём идет о доброте в тюремной жизни.

В целом результаты проведенного эксперимента можно считать неплохими. Они свидетельствуют о том, что если топик отвечает критериям семантической «цельности», то более чем в половине случаев его можно считать хорошо соответствующим тексту, а в ¾ случав – хотя бы частично соответствующим. Тем не менее около четверти всех полученных тематических кластеров можно считать случайными.

Корреляция между долями хорошо соответствующих топику рассказов и размером выборки не просматривается. Однако можно предположить, что интерпретируемость тематической модели, обученной на наборе текстов, зависит не только от рассмотренных вариантов предобработки, но и от лексической и стилистической однородности анализируемого набора текстов – для более стилистически разнородных коллекций текстов (в нашем случае это проза раннесоветского периода) необходимы большие выборки. Увеличение объема выборки позволяет увеличить и описательную способность модели ввиду соответствующего увеличения тематической дескриптивности текстов: литературные темы склонны повторяться от произведения к произведению (см. «вечные» темы [36]), следовательно, чем больше текстов

включено в выборку, тем больше вероятность уловить более точно их тематическое разнообразие.

Заключение

Главной целью проведенного исследования стало выяснение того, насколько адекватно семантически целостные и интерпретируемые топики, полученные в результате построения тематической модели, отражают какие-либо из содержательных аспектов литературного произведения малой прозаической формы (персонажей, сюжет, фон повествования и др.). С этой точки зрения результаты эксперимента можно считать вполне успешными, так как автоматически выделенные темы для более половины (52%) всех отобранных документов (рассказов) были признаны экспертами соответствующими им в полной мере, а еще 24% – соответствующими частично; и только оставшиеся 24% дают нерелевантный результат. Таким образом, 76% всех интерпретируемых топиков действительно представляют собой не просто набор статистически выделяющихся на данном подмножестве текстов слов, но и имеют непосредственное отношение к содержанию рассказа. Полученный результат можно считать весьма позитивным, поскольку до настоящего времени качественных исследований соответствия топиков литературным текстам не проводилось, более того высказывались сомнения о целесообразности применимости методов тематического моделирования к художественным текстам.

В этой связи несмотря на то, что точное содержание некоторых рассказов не удалось определить, метод тематического моделирования позволил достаточно точно выявить типичные для коллекции рассказов топики без их прочтения, что даёт возможность группировать тексты по тематикам и делать выводы о появлении и развитии той или иной темы во времени. Поэтому эксперименты с тематическим моделированием прозы безусловно стоит продолжать.

При этом нужно иметь в виду, что проведенный эксперимент базировался не на всем множестве построенных моделями топиков на основе выборок разных размеров (которых было получено для оптимального варианта предобработки 265) и не на всем множестве документов (рассказов), а только для 24 топиков (что составляет всего 9%), которые были признаны интерпретируемыми одновременно тремя экспертами с точки зрения семантической цельности составляющих их слов.

Поскольку процесс «отсева» семантически неоднородных групп слов, который был реализован на этапе первого эксперимента, представляет собой значительно более простую задачу, чем экспертное прочтение тысячи текстов, можно предложить использованную методику как рабочую для отбора адекватных топиков на «больших» литературных данных. Фильтрацию семантически неоднородных групп можно проводить или, как в нашем эксперименте, экспертным путем, или подключать специальные модули семантического анализа, базирующиеся на представительных неспециализированных

онтологиях. После этого можно предполагать, что топики, прошедшие критерий семантической целостности, будут хорошо соответствовать приписанным им рассказам с вероятностью 75%. Для автоматической обработки литературного материала это представляется вполне хорошим результатом.

При этом остается нерешенным вопрос, сколько топиков, которые на первом этапе исследования не были признаны интерпретируемыми и поэтому не включались во второй эксперимент, в той или иной мере описывают содержание соответствующих им рассказов. Ответ на этот вопрос еще ждет своего решения – он требует эксперимента, подобного тому, который был описан в последнем разделе статьи.

Результаты экспериментов показали, что на соответствие топика рассказу влияют и некоторые характеристики самих текстов. Так, было замечено, что среди рассказов с нерелевантыми топиками есть значительное количество текстов совсем небольшого размера. Поэтому можно предположить, что размер рассказа может влиять на качество построенных для него моделей. Для изучения этого вопроса необходимо провести специальное исследование. Также можно выдвинуть гипотезу, что топики, построенные для «описательных» рассказов, будут получать лучшие оценки от экспертов, чем топики для динамичных текстов, насыщенных событиями.

Следует отметить, что наше исследование не дает окончательного ответа на вопрос, что в конечном итоге могут описывать полученные топики. Поскольку на каждый из рассказов может быть получено по несколько автоматических топиков, очевидно, собственно содержание рассказа надо искать в совокупности этих элементов, в то время как в данной работе топики рассматривались изолированно, сами по себе. Продолжение этой работы – в будущих исследованиях.

Итак, несмотря на ряд проблем, обусловленных особенностями художественного текста как материала для компьютерного анализа, тематические модели, построенные на литературных корпусах, могут стать инструментом для проведения как филологических, так и лингвистических исследований. Во-первых, как уже было отмечено, художественное произведение представляет собой некоторую компрессию действительности. И тема в этом смысле оказывается в фокусе нашего внимания, поскольку отражает, какие части этой действительности (т.е. образы и характеры, примеры взаимодействия с другими людьми, обществом и природой и т.д.) попадают в поле зрения автора [36]. Однако текст неизбежно содержит и «глубинные», или «подтекстовые», темы, что создаеттрудности при интерпретации, в том числе и традиционной [37].

Тематическая модель же, напротив, должна позволить уравновесить, усреднить эти расхождения, более объективно описать художественную реальность – через языковые и стилистические особенности текстов, которые могут, в свою очередь, указывать на ряд сюжетных и тематических отличительных характеристик литературного корпуса. Деформируется и ее конечная цель, так как, применяя тематическое моделирование к художественным

произведениям, мы стремимся не только описать их содержательную сторону, но и на основе полученных результатов приблизиться к более глобальной идее – моделированию национальной литературы как целостной системы [38].

В заключение стоит отметить, что хотя современные алгоритмы автоматического тематического анализа уступают по качеству экспертному выявлению тем, предполагающему медленное чтение специалистом каждого отдельного рассказа, компьютерные методы необходимо развивать и оптимизировать для решения задач, связанных с обработкой больших данных. В дальнейшем видится важным продолжение разработки проблемы приближения результатов тематического моделирования к тому, что понимается под «темой» художественного произведения. Есть основание предполагать, что более соответствующее тематическое описание текста можно получить, комбинируя тематическое моделирование с другими методами автоматической обработки текстов. Например, излечение именованных сущностей позволит получить схематическое представление о том, где происходит действие рассказа, что может быть существенно для его понимания. А анализ тональности мог бы позволить выявлять наиболее эмоционально насыщенные – «драматические» – фрагменты текста и связанные с ними темы (топики). Таким образом, представляется, что будущее тематического моделирования для обработки больших литературных данных состоит в разработке гибридных моделей с привлечением ряда других методов автоматической обработки текста.

Список источников

1. *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: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. P. 2. 2020. Vol. 12469. P. 134–152.

2. *Jockers M.L., Mimno D.* Significant themes in 19th-century literature // Poetics. 2013. Vol. 41, № 6. P. 750–769.

3. *Jautze K., Cranenburgh A. van, Koolen C.* Topic modeling literary quality // Digital Humanities. Conference Abstracts. 2016. P. 233–237.

4. *Schöch C.* Topic modeling genre: an exploration of French classical and enlightenment drama // Digital Humanities Quarterly. 2017. Vol. 11, № 2. doi: 10.48550/arXiv.2103.13019 (дата обращения: 01.05.2021).

5. *Navarro-Colorado B.* On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry // Frontiers in Digital Humanities. 2018. Vol. 5, № 15. doi: 10.3389/fdigh.2018.00015

6. *Митрофанова О.А.* Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Корпусная лингвистика 2019. СПб., 2019. С. 387–394.

7. *Корпус* русского рассказа 1900–1930. URL: https://russian-short-stories.ru/ (дата обращения: 05.12.2022).

8. *Мартыненко Г.Я., Шерстинова Т.Ю., Мельник А.Г., Попова Т.И.* Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ре-

сурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) // Компьютерная лингвистика и вычислительные онтологии. СПб. : НИУ ИТМО, 2018. № 2. С. 97–102.

9. *Мартыненко Г.Я., Шерстинова Т.Ю., Попова Т.И., Мельник А.Г., Замирайлова Е.В.* О принципах создания корпуса русского рассказа первой трети XX века // Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». Казань, 2018. С. 180–197.

10. *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 International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. 2020. Vol. 2552. P. 105–120. URL: http://ceur-ws.org/Vol-2552/

11. *Zamiraylova E., Mitrofanova O.* Dynamic topic modeling of Russian fiction prose of the first third of the 20th century by means of non-negative matrix factorization // Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). 2020. Vol. 2552. P. 321–339.

12. *Skrebtsova T.G.* Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories // International Conference «Internet and Modern Society» (IMS-2020). CEUR Workshop Proceedings, 2021. P. 265–276.

13. *Sherstinova T., Kirina M.* Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers // Alexandrov D.A. et al. Digital Transformation and Global Society. DTGS 2021. Communications in Computer and Information Science. Vol. 1503. Cham, 2022. P. 332–346. doi: 10.1007/978-3-030-93715-7_24

14. *Gryaznova E., Kirina M.* Defining Types of Violence: Comparing Topic Modeling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories from the 1900s to the 1930s // Proceedings of the International Conference «Internet and Modern Society» 2021. P. 281–290.

15. *Кирина М.А.* Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 2. С. 93–109. doi: 10.25205/1818-7935-2022-202-93-109

16. *Rhody L.M.* Topic Modelling and Figurative Language // Journal of Digital Humanities. 2012.

17. *Da N.Z.* The computational case against computational literary studies // Critical Inquiry. 2019. Vol. 45, № 3. P. 601–639.

18. *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. URL: http://ceur-ws.org/Vol-2723/long7.pdf.

19. *Жолковский А.К., Щеглов Ю.К.* К понятиям «тема» и «поэтический мир» // Щеглов Ю.К. Избранные труды / сост. А.К. Жолковский, В.А. Щеглова. М. : РГГУ, 2013. С. 37–78.

20. *Тынянов Ю.Н.* Архаисты и новаторы. Л. : Прибой, 1929.

21. *Мартыненко Г.Я.* Основы стилеметрии. Л. : Изд-во ЛГУ, 1988.

22. *Мартыненко Г.Я.* Методы математической лингвистики в стилистических исследованиях. М. : Нестор-История, 2019.

23. *Honnibal M., Montani I.* spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. URL: https://spacy.io/models/ru (дата обращения: 01.05.2021).

24. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet allocation // The Journal of machine Learning research. 2003. Vol. 3. P. 993–1022.

25. *Daud A., Li J., Zhou L. et al.* Knowledge discovery through directed probabilistic topic models: a survey // Front. Comput. Sci. China. 2010. T. 4, № 3. P. 280–301. doi: 10.1007/s11704-009-0062-y

26. *Митрофанова О.А.* Моделирование тематики специальных текстов на основе алгоритма LDA // XLII Международная филологическая конференция. 11–16 марта 2013 г. Избранные труды. СПб., 2014.

27. *Řehůřek R., Sojka P.* Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. 2010.

28. *Blei D.M.* Probabilistic topic models // Communications of the ACM. 2012. Vol. 55, № 4. P. 77–84.

29. *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. № 23.

30. *Kherwa P., Bansal P.* Topic modeling: a comprehensive review // EAI Endorsed transactions on scalable information systems. 2020. Vol. 7, № 24.

31. *Röder M., Both A., Hinneburg A.* Exploring the Space of Topic Coherence Measures // Proceedings of the eighth International Conference on Web Search and Data Mining, 2015.

32. *Sherstinova T., Kirina M., Zavyalova I., Karysheva A., Kolpashchikova E., Maksimenko P., Moskalenko A., Moskvina A., Кирина М.А.* Topic Modeling of Literary Texts Using LDA: On the Influence of Linguistic Preprocessing on Model Interpretability // 2022 31st Conference of Open Innovations Association (FRUCT). Vol. 32. IEEE, 2022. P. 305–312.

33. *Шерстинова Т.Ю., Москвина А.Д., Кирина М.А.* Тематическое моделирование русского рассказа 1900–1930: наиболее частотные темы и их динамика // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2022». 2022. Вып. 21. С. 512–526.

34. *Chang J., Gerrish S., Wang C., Boyd-Graber J.L., Blei D.M.* Reading tea leaves: how humans interpret topic models // Adv. Neural Inf. Process. Syst. 2009. Vol. 22. P. 288–296.

35. *Воронцов К.В., Фрей А.И., Апишев М.А., Потапенко А.А.* Тематическое моделирование в BigARTM: теория, алгоритмы, приложения. 2015. URL: http://www.machinelearning.ru/wiki/images/b/bc/Voron-2015-BigARTM.pdf

36. *Томашевский Б.В.* Теория литературы. Поэтика : учеб. пособие. М. : Аспект Пресс, 1996. С. 176–192.

37. *Вершинина Н.Л., Волкова Е.В., Крупчанов Л.М. [и др.]* Введение в литературоведение : учеб. для бакалавров. М. : Юрайт, 2015.

38. *Sherstinova T., Moskvina A., Kirina M.* Towards automatic modelling of thematic domains of a national literature: Technical issues in the case of Russian // 2021 29th Conference of Open Innovations Association (FRUCT). IEEE, 2021. P. 313–323.

References

1. Sherstinova, T. et al. (2020) [Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction]. *Advances in Computational Intelligence*. Proceedings of the 19th Mexican International Conference on Artificial Intelligence MICAI 2020. Part II. Vol. 12469. Mexico City. 12–17 October 2020. Springer. pp. 134–152.

2. Jockers, M.L. & Mimno, D. (2013) Significant themes in 19th-century literature. *Poetics*. 6 (41). pp. 750–769.

3. Jautze, K., van Cranenburgh, A. & Koolen, C. (2016) [Topic modeling literary quality]. *Digital Humanities.* Abstracts of the Conference. Kraków. 11–16 July 2016. Kraków: ADHO. pp. 233–237.

4. Schöch, C. (2017) Topic modeling genre: an exploration of French classical and enlightenment drama. *Digital Humanities Quarterly.* 11 (2). doi: 10.48550/arXiv.2103.13019 (Accessed: 01.05.2021).

5. Navarro-Colorado, B. (2018) On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. *Frontiers in Digital Humanities*. 5. doi: 10.3389/fdigh.2018.00015

6. Mitrofanova, O.A. (2019) [Study of the structural organization of a work of art using thematic modeling: experience of working with the text of the novel The Master and Margarita by M.A. Bulgakov]. *Korpusnaya lingvistika – 2019* [Corpus linguistics – 2019]. Proceedings of the International Conference. Saint Petersburg. 24–28 June 2019. Saint Petersburg: Saint Petersburg State University. pp. 387–394. (In Russian).

7. Korpus russkogo rasskaza 1900–1930 [Corpus of Russian Short Stories 1900–1930]. (n.d.) https://russian-short-stories.ru/ (Accessed: 05.12.2022).

8. Martynenko, G.Ya. et al. (2018) Metodologicheskie problemy sozdaniya Komp'yuternoy antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoy khudozhestvennoy prozy v epokhu revolyutsionnykh peremen (pervoy treti XX veka) [Methodological problems of creating a Computer anthology of Russian short stories as a language resource for studying the language and style of Russian fiction in the era of revolutionary changes (the first third of the 20th century)]. *Komp'yuternaya lingvistika i vychislitel'nye ontologii*. 2. pp. 97–102.

9. Martynenko, G.Ya. et al. (2018) [On the principles of creating a corpus of Russian short stories of the first third of the 20th century]. *TEL 2018*. Proceedings of the 15th International Conference. Kazan. 31 October – 3 November 2018. Kazan: Tatarstan AS. pp. 180–197. (In Russian).

10. Sherstinova, T. & Martynenko, G. (2020) [Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century]. *R. Piotrowski's Readings in Language Engineering and Applied Linguistics*. Proceedings of the 3rd International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). Vol. 2552. Saint Petersburg. 27 November 2019. pp. 105–120. [Online] Available from: http://ceur-ws.org/Vol-2552/

11. Zamiraylova, E. & Mitrofanova, O. (2020) [Dynamic topic modeling of Russian fiction prose of the first third of the 20th century by means of non-negative matrix factorization]. *R. Piotrowski's Readings in Language Engineering and Applied Linguistics*. Proceedings of the 3rd International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). Vol. 2552. Saint Petersburg. 27 November 2019. pp. 321–339. [Online] Available from: http://ceur-ws.org/Vol-2552/

12. Skrebtsova, T.G. (2021) Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories. *IMS-2020. Internet and Modern Society.* Proceedings of the International Conference. Vol. 2813. Saint Petersburg. 17-20 June 2020. pp. 265–276. [Online] Available from: https://ceur-ws.org/Vol-2813/

13. Sherstinova, T. & Kirina, M. (2022) [Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers]. *Digital Transformation and Global Society. DTGS 2021*. Proceedings of the 6th International Conference. Vol. 1503. Saint Petersburg. 23–25 June 2021. Cham. doi: 10.1007/978-3-030-93715-7_24.

14. Gryaznova, E. & Kirina, M. (2021) Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900–1930. *Internet and Modern Society.* Proceedings of the International Conference. pp. 281–290.

15. Kirina, M.A. (2022) Sravnenie tematicheskikh modeley na osnove LDA, STM i NMF dlya kachestvennogo analiza russkoy khudozhestvennoy prozy maloy formy [Comparison of thematic models based on LDA, STM and NMF for the qualitative analysis of Russian short-form fiction]. *Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya*. 2 (20). pp. 93–109. doi: 10.25205/1818-7935-2022-202-93-109

16. Rhody, L.M. (2012) Topic Modelling and Figurative Language. *Journal of Digital Humanities*. 2 (1). pp. 305–312.

17. Da, N.Z. (2019) The computational case against computational literary studies. *Critical Inquiry*. 3 (45). pp. 601–639.

18. Uglanova, I. & Gius, E. (2020) [The Order of Things. A Study on Topic Modelling of Literary Texts]. *CHR 2020*. Proceedings of the Workshop on Computational Humanities Research. Amsterdam. 18–20 November 2020. [Online] Available from: http://ceur-ws.org/Vol-2723/long7.pdf

19. Zholkovskiy, A. K. & Shcheglov, Yu.K. (2013) K ponyatiyam "tema" i "poeticheskiy mir" [On the concepts of "theme" and "poetic world"]. In: Shcheglov, Yu.K. *Izbrannye trudy* [Selected Works]. Moscow: Russian State University for the Humanities. pp. 37–78.

20. Tynyanov, Yu.N. (1929) *Arkhaisty i novatory* [Archaists and Innovators]. Leningrad: Priboy.

21. Martynenko, G.Ya. (1988) *Osnovy stilemetrii* [Fundamentals of Stylometry]. Leningrad: Leningrad State University.

22. Martynenko, G.Ya. (2019) *Metody matematicheskoy lingvistiki v stilisticheskikh issledovaniyakh* [Methods of Mathematical Linguistics in Stylistic Research]. Moscow; Saint Petersburg: Nestor-Istoriya.

23. Honnibal, M. & Montani, I. (2017) *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. [Online] Available from: https://spacy.io/models/ru (Accessed: 01.05.2021).

24. Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet allocation. *The Journal of Machine Learning Research*. 3. pp. 993–1022.

25. Daud, A. et al. (2010) Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of Computer Science in China*. 4. pp. 280–301.

26. Mitrofanova, O.A. (2014) Modelirovanie tematiki spetsial'nykh tekstov na osnove algoritma LDA [Modeling the topics of special texts based on the LDA algorithm]. *Proceedings of the 42nd International Philological Conference*. Saint Petersburg. 11–17 March 2013. Saint Petersburg: Saint Petersburg State University. (In Russian).

27. Řehůřek, R. & Sojka, P. (2010) [Software framework for topic modelling with large corpora]. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta. 22 May 2010. Valletta: ELRA. pp. 45–50.

28. Blei, D.M. (2012) Probabilistic topic models. *Communications of the ACM*. 4 (55). pp. 77–84.

29. Korshunov, A. & Gomzin, A. (2012) Tematicheskoe modelirovanie tekstov na estestvennom yazyke [Thematic modeling of texts in natural language]. *Trudy Instituta sistemnogo programmirovaniya RAN*. 23. pp. 215–242.

30. Kherwa, P. & Bansal, P. (2020) Topic modeling: a comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*. 24 (7). doi: 10.4108/eai.13-7-2018.159623

31. Röder, M., Both, A. & Hinneburg, A. (2015) [Exploring the Space of Topic Coherence Measures]. *WSDM '15*. Proceedings of the 8th ACM International Conference on Web Search and Data Mining. Shanghai. 2–6 February 2015. New York: Association for Computing Machinery. pp. 399–408.

32. Sherstinova, T. et al. (2022) [Topic Modeling of Literary Texts Using LDA: On the Influence of Linguistic Preprocessing on Model Interpretability]. *Proceedings of the 31st Conference of Open Innovations Association (FRUCT).* Vol. 32. Helsinki. 27–29 April 2022. IEEE. pp. 305–312.

33. Sherstinova, T.Yu. et al. (2022) [Thematic modeling of the Russian story 1900–1930: the most frequent themes and their dynamics]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computer Linguistics and Intellectual Technologies]. Proceedings of the International Conference Dialog 2022 [Dialogue 2022]. Vol. 21. Moscow. 11 February 2022. Moscow: Russian State University for the Humanities. pp. 512–526. (In Russian).

34. Chang, J. et al. (2009) [Reading tea leaves: How humans interpret topic models]. *Advances in Neural Information Processing Systems 22*. Proceedings of the 23rd Annual

Conference on Neural Information Processing Systems. Vancouver. 7–10 December 2009. Princeton University. pp. 288–296.

35. Vorontsov, K.V. et al. (2015) *Tematicheskoe modelirovanie v BigARTM: teoriya, algoritmy, prilozheniya* [Thematic Modeling in BigARTM: Theory, algorithms, applications]. http://www.machinelearning.ru/wiki/images/b/bc/Voron-2015-BigARTM.pdf

36. Tomashevskiy, B.V. (1996) *Teoriya literatury. Poetika* [Theory of Literature. Poetics]. Moscow: Aspekt Press. pp. 176–192.

37. Krupchanov, L.M. (ed.) (2015) *Vvedenie v literaturovedenie* [Introduction to Literary Criticism]. Moscow: Izdatel'stvo Yurayt.

38. Sherstinova, T., Moskvina, A. & Kirina, M. (2021) [Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian]. *Proceedings of the 29th Conference of Open Innovations Association FRUCT*. Tampere. 12–14 May 2021. IEEE.

Информация об авторах:

Шерстинова Т.Ю. – канд. филол. наук, доцент департамента филологии Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург, Россия). E-mail: sherstinova@gmail.com

Кирина М.А. – магистрант, приглашенный преподаватель департамента филологии Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург, Россия). E-mail: mkirina2412@gmail.com

Москвина А.Д. – преподаватель департамента филологии Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург, Россия). E-mail: moskvina.anya@gmail.com

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

T.Yu. Sherstinova, Cand. Sci. (Philology), associate professor, National Research University Higher School of Economics (St. Petersburg, Russian Federation). E-mail: sherstinova@gmail.com

M.A. Kirina, master's student, visiting lecturer, National Research University Higher School of Economics (St. Petersburg, Russian Federation). E-mail: mkirina2412@gmail.com

A.D. Moskvina, lecturer, National Research University Higher School of Economics (St. Petersburg, Russian Federation). E-mail: moskvina.anya@gmail.com

The authors declare no conflicts of interests.

Статья поступила в редакцию 10.12.2022;
одобрена после рецензирования 05.06.2023; принята к публикации 27.05.2024.

The article was submitted 10.12.2022;
approved after reviewing 05.06.2023; accepted for publication 27.05.2024.