Вестник Томского государственного университета. Филология. 2023. № 86. С. 84–101
Tomsk State University Journal of Philology. 2023. 86. pp. 84–101

Научная статья
УДК 81'33
doi: 10.17223/19986645/86/6

Дискурсивные варианты тематического моделирования пандемии Covid-19 (новостной медиадискурс VS социальные сети)

Зоя Ивановна Резанова[1], Андрей Александрович Степаненко[2]

[1, 2] Национальный исследовательский Томский государственный университет, Томск, Россия
[1] rezanovazi@mail.ru
[2] stepanenkone@mail.ru

Аннотация. Представлены результаты сравнительного исследования репрезентации тем новой коронавирусной инфекции в новостных текстах официальных СМИ и социальной сети «Твиттер», проведенного на основе совмещения методов дискурс-анализа и математического автоматического анализа текста (Латентное размещения Дирихле (LDA) в сочетании с методом выявления ключевых слов TF-IDF. Характеризуются темы, общие и различающие два дискурса, а также особенности концептуального моделирования в общих и различных темах.

Ключевые слова: коронавирус, Covid-19, пандемия, инфодемия, новостной дискурс, социальные сети, «Твиттер», автоматическое тематическое моделирование

Благодарности: исследование выполнено за счет гранта Российского научного фонда (проект № 23-28-01001).

Для цитирования: Резанова З.И., Степаненко А.А. Дискурсивные варианты тематического моделирования пандемии Covid-19 (новостной медиадискурс VS социальные сети) // Вестник Томского государственного университета. Филология. 2023. № 86. С. 84–101. doi: 10.17223/19986645/86/6

Original article
doi: 10.17223/19986645/86/6

Discursive variants of thematic modeling of COVID-19 (news media discourse VS social networks)

Zoya I. Rezanova[1], Andrei A. Stepanenko[2]

[1, 2] National Research Tomsk State University, Tomsk, Russian Federation
[1] rezanovazi@mail.ru
[2] stepanenkone@mail.ru

Abstract. The article presents the results of a comparative research of the representation of the topic related to the novel coronavirus infection in the news texts of official

Russian state media and the social network Twitter. We compared the COVID-19 pandemic topic modeling in two different types of discourse and detected the similarities and differences that show in common themes, in the nature of the variability of similar topics' conceptual modeling in the discourses. The analysis was based on the combination of discourse analysis and text mining. The discourse analysis method was applied at the initial stage of the project to justify the choice of the material (types of texts and their genre varieties in the two discourses) and at the final stage of discourse analysis to interpret the results of the automatic analysis of topic modeling. The text mining method included Latent Dirichlet allocation (LDA), which was applied in combination with the TF-IDF keyword detection method. The study was carried out based on the text material of RIA, TASS and Lenta.ru information agencies (10,499,390 tokens) and Twitter (1,163,511 tokens) published in 2020. This period is characterized as the beginning of infection, the first wave and the peak of COVID-19 in the Russian Federation. Thus, the unprecedented global scale of the new coronavirus infection, the severe course of the disease, the high mortality rate, the resulting restrictions, including restrictions on global travel, the need for strict quarantine measures within the country, and the treatment of the ill in medical institutions are the topics that united the state-owned media and the "independent" Internet discourse in the country in the first year of the pandemic. What should also be noted is the high level of unity in the lexical representation of the topics in the two discourses. However, the same topics are parts of different thematic units in the discourses. Topics on Twitter are combined and coordinated with themes that reveal a personal projection of the course and experience of the pandemic: the residents' tracking of the daily statistics on the new infection, deaths and recoveries, which caused active discussions; testing procedures, experiences of the course of the disease, and changes in private life associated with restrictions. The difference between the topic of quarantine restrictions, common for both the state media and the Internet discourse, and the Twitter-specific topic of life in self-isolation represented by lexemes that actualize the topic of personal life, is illustrative in this respect.

Keywords: coronavirus, COVID-19, pandemic, infodemic, news discourse, social networks, Twitter, automated topic modeling

Acknowledgements: The study was supported by the Russian Science Foundation, Project No. 23-28-01001.

For citation: Rezanova, Z.I. & Stepanenko, A.A. (2023) Discursive variants of thematic modeling of COVID-19 (news media discourse VS social networks). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology.* 86. pp. 84–101. (In Russian). doi: 10.17223/19986645/86/6

Пандемия новой коронавирусной инфекции Covid-19 затронула практически все аспекты жизни человека. Одним из непосредственных социальных последствий пандемии явилась инфодемия, которая представляет собой переизбыток как онлайновой, так и оффлайновой информации, характеризующейся наличием противоположных мнений, оценок, слухов. Социальная значимость феномена инфодемии была осознана мировым сообществом, что отразилось в ряде документов ВОЗ, ОНН и ЮНЕСКО, в которых дано ее определение и содержится призыв ко всем государствам «принимать меры по противодействию ложным слухам и дезинформации... и содействовать предоставлению научно обоснованных данных в распоряжение общественности» [1]. Большие кризисы дают весьма репрезентативный материал

для исследования направленности информационных потоков, степени их согласованности и вариативности. Феномен инфодемии был осознан как социально значимый в период распространения атипичной пневмонии SARS в 2003 г. [2. С. 941], но в исследовательский фокус ряда гуманитарных наук, в первую очередь социологии и журналистики [3], попал в период распространения пандемии Covid-19.

Проблемное поле инфодемии в пределах лингвистических исследований разрешается прежде всего в логике дискурс-анализа. При этом, на наш взгляд, ключевым является вопрос о соотношении информации, распространяемой государственными институтами, в частности государственными медиа, и персональных, личностных «ответов» на социальные потрясения, транслируемых в социальных сетях (СС). Новостной медиадискурс и персональные интернет-дискурсы занимают ключевое положение в интерпретации ситуации пандемии новой коронавирусной инфекции и вместе с тем характеризуются разнонаправленностью как общих институционально, так и личностно обусловленных интересов их субъектов.

Значимыми при изучении информационных потоков в период ситуаций социальных рисков станорвятся определение доминирующих тем обсуждения в новостном информационном пространстве и социальных сетях и актуализация ключевых концептов в их рамках, определение степени их совпадения и различий, основных зон несовпадений информационных фокусировок событий.

Уже в работах Т. ван Дейка был убедительно обоснован принцип структурирования событийного ряда в новостном дискурсе по принципу релевантности, обусловленному процессами институционального и социального производства и потребления новости [4. С. 125, 137]. Современная медиалингвистика обращена не только к отражательно-информационной, но и интерпретационной функции новостей. Все больше осознается необходимость исследования тематических фокусировок в новостном потоке как способов объективации точек зрения, «фокусов» эмпатии официальных акторов медиадискурса. Как отмечает Т.Г. Добросклонская, в медиалингвистике при анализе новостей решающее значение имеет анализ того, как в новости реализуется интерпретация события и каков их диапазон в разных медиа [5. С. 19].

Современные медийные коммуникации переживают существенные трансформации в связи с широким распространением социальных сетей, которые вывели в пространство медиа в качестве активных акторов человека, транслирующего не только и не столько институционально обусловленную, сколько личностную позицию. При этом одну из лидирующих площадок в данном коммуникационном пространстве современных социальных сетей занимает «Твиттер», что отмечается в ряде зарубежных и российских исследований (см. обзор в [6]). С одной стороны, «Твиттер» стал площадкой представления институциональных позиций, он широко используется представителями разных социальных институтов, прежде всего в политической и бизнес-коммуникации (см., например, анализ в [7–9]). С другой стороны, эта социальная сеть, обеспечивая возможность адресанта «самостоятельно

определять тему, содержание и форму твита», создает условия для развития межличностной, групповой и массовой коммуникации, нередко реализуемой «без учета национальных, культурных, государственных, экономических и политических границ», что делает его привлекательным «для абсолютно любых пользователей, независимо от возраста, пола, национальности, социального положения и политических взглядов» [9. С. 32].

Проведенный нами анализ англоязычной литературы, посвященной информационным аспектам Covid-19, по данным сайта https://arxiv.org, расположившего препринты статей всех тематик, связанных с пандемией, выявил, что значительно преобладают исследования того, как пандемия интерпретируется в социальных сетях. При этом число обращений к новостному дискурсу значительно уступает исследованиям проблем распространения информации в «Твиттер»; отмечена одна работа, в которой анализируется степень смысловой близости текстов новостей и текстов в «Твиттер» (T. Matthew Osborne et al.), доминирующей темой является распространение фейков (Wilton O. Júnior et al., Gullal S. Cheema et al., Hui Yin et al., Jan Philip Wahle et al., Elnaz Zafarani-Moattar et al., Mrinal Rawat et al., Dongwoo Lim et al.). Представленные на сайте исследования выполнялись на основе применения математических методов анализа текстов социальных сетей и СМИ.

Результаты анализа русскоязычного сегмента исследований, посвящённых информационному аспекту пандемии Covid-19, свидетельствуют о незначительном числе работ, раскрывающих направления интерпретации пандемии в массмедиа. Как правило, это работы социологической направленности, обобщенно рассматривающие информационные процессы в официальных СМИ и СС [2, 3, 13]. Отметим следующие значимые аспекты, отраженные в публикациях: инфодемийный характер информационного сопровождения пандемии [2, 3], проблемы информационных рисков и фейков [2, 10], «страхогенности» СМИ и социальных сетей [11], тематическое представление пандемии в новостных медиа и социальных сетях [12, 13].

Ряд работ выполнен в логике дискурс-анализа с применением интерпретативных интроспективных методов и методов концептуального моделирования текстов СМИ и социальных сетей разных жанров. В результате авторы исследований представляют типовые контексты актуальных тем, формируемые в текстах концепты [14, 15]. Однако полагаем, что на современном этапе исследований информационных потоков необходимо расширение лингвистической методологии, дополнение качественных собственно лингвистических текстологических методов, интроспективного анализа количественными методами автоматического математического анализа текстов.

В данной статье мы представляем результаты сравнительного исследования репрезентации тем в новостных текстах официальных СМИ и твитах, проведенного на основе совмещения методов дискурс-анализа и методов математического автоматического анализа текста. Метод дискурс-анализа был применен на начальном этапе проекта, при обосновании выбора материала – типов текстов, фиксирующих вариантные жанровые типы двух дис-

курсов, и на заключительном этапе – при интерпретации результатов анализа текстов с использованием методов автоматического анализа. Из спектра математических методов тематического анализа текста были применены метод латентного размещения Дирихле (LDA, Latent Dirichlet allocation) в сочетании с методом выявления ключевых слов TF-IDF.

Цель проведенного анализа – сравнить направленность тематического моделирования ситуации пандемии Covid-19 в двух типах дискурсов, выявить степень сходств и различий, проявляющуюся в наличии общих тем, а также в характере вариативности концептуального моделирования одних и тех же тем в сравниваемых дискурсах.

Исследование проведено на материале текстов новостей информационных агентств РИА, ТАСС и Лента.ру (10 499 390 токенов) и твитов СС Твиттер (1 163 511 токенов за период с 01.01.2020 по 31.12.2020), представленных в следующих временных интервалах: а) РИА – со 2 января по 28 декабря 2022; б) Лента – с 1 августа по 28 декабря 2020 г.; в) ТАСС – со 2 января по 30 декабря 2020 г. Данный период является первой волной, началом и пиком заболеваемости и распространения Covid 19 в РФ. Сбор текстового массива данных новостных агентств осуществлен при помощи автоматического скрипта (веб скрапер), написанного на языке программирования R 4.0.2 и пакета rvest 1.0.0; скрапер текстов СС Твиттер был осуществлен на базе языка программирования Python 3.

Необходимая при математическом анализе предобработка текстов включала токенизацию, лемматезацию, приведение вех слов в единый (нижний) регистр, удаление малоинформативных лексических единиц, удаление «стоп-слов», удаление высокочастотных и низкочастотных лексических единиц в соответствии с абсолютной частотой их встречаемости. Последний тип удаления был обусловлен тем, что подобные лексические единицы редко являются тематическими. Препроцессинг и последующий анализ был осуществлен при помощи языка программирования R 4.0.2 и библиотеки quanteda. Лемматезация была осуществлена при помощи программы Mystem 3.0.

При решении задачи автоматического тематического моделирования были использованы следующие частные методы. Нормализация матрицы частот на основе относительной нормализации лексических единиц в документе ($TF_{t,d} = m/n$), где m – абсолютная частота лексемы; n – количество слов в документе, была обусловлена различием в объеме текстов как самих новостных агентств, так и текстов «Твитера». Таким способом исключалось влияние объема текста на результат исследования. Второй вариант нормализации был основан на применении классической формулы *TF–IDF*: $TF\text{–}IDF_{t,d} = TF_{t,d}\, IDF_t$, где *TF* – относительная частота слова в документе, а *IDF* – логарифм количества документов, в которых встретился термин (слово), деленный на количество всех документов. В качестве удельного веса (Score measure) каждого вектора оценки документа d применялось суммирование TF–IDF в d (документе):

$$Score(q,d) = \sum_{t \in q} \text{TF–IDF}_{t,d}$$

Высокий удельный вес лексической единицы демонстрирует «важность» текущей темы в коллекции документов.

Далее были составлены частотные матрицы лексических единиц двух групп тексты новостных агентств и тексты социальной сети «Твиттер» со следующими колонками: текст (text), препроцессинг (stem), источник (source: риа, лента, тасс), тип (type: новость, твиттер), краткое псевдоназвание текста (MetaDoc). Колонка «текст» включает исходный текст новости или пост «Твиттера»; «препроцессинг» – вывод результата алгоритмов препроцессинга; колонка «источник» демонстрирует ресурс, из которого был взят текст: риа, лента, тасс, твиттер; колонка «тип» включает два варианта разметки принадлежности текста к жанру: новость или твит; колонка «краткое псевдоназвание текста» – иератор, включающий название источника (например, тексты информационного агентсва ТАСС размечены как tass и номер текста – 1: tass1)

Далее был применен алгоритм Латентного размещения Дирихле (LDA, Latent Dirichlet allocation) [16. P. 56]. Основной принцип работы LDA заключается в создании матрицы, которая отображает комбинацию тем (абстрактных тематических категорий, которые скрыты в документах) и слов в документах. Сначала каждый документ представляется как набор слов и их частот. Затем модель LDA рассчитывает вероятности наличия тем в каждом документе и вероятности наличия определенных слов в каждой теме. Сама модель LDA была ограничена гиперпараметрами, которые значительно влияют на качество работы модели. Список гиперпараметров был ограничен: а) количеством тем; б) значениями α и β, где α – плотность тем в документе (чем выше его значение, тем больше тем содержится в одном документе), β – плотность терминов в теме (чем выше коэффициент, тем больше количество терминов в теме. Гиперпараметры подбирались эмпирически, по соответствию параметров α и β уровню < 0,95 при наиболее когерентном или полном выводе лексем в каждой теме. Ограничение количества тем основывалось на метрике R. Deveaud [17. P. 61–84]. Данный тип метрики позволил выявить 12 оптимальным тем, которые представлены в таблице. Таким образом, была найдена оптимальная модель для данной коллекции документов с гиперпараметрами и темами для всего корпуса текстов.

В результате применения алгоритма были получены списки тематически связанных слов, из которых для визуализации были взяты пять наиболее вероятных слов. В качестве примера отобразим полученный результат выделенных тем и распределения в них лексем (рис. 1).

Гистограмма, представленная на рис. 1, демонстрирует распределение вероятностей встретить каждую лексему в каждой из выявленных двенадцати тем.

Рис. 1. Гистограмма распределения пяти наиболее вероятных лексем в 12 темах

На втором шаге были объединены методы LDA и меры выявления ключевых слов *TF–IDF*; на этой основе выявлены ключевые темы для того или иного класса текстов (СМИ и социальной сети «Твиттер»). В данном случае суммируется средневзвешенное значение *TF–IDF* всех слов, входящих в соответствующую тему.

В результате также были получены группы тематически связанных слов, из которых для дальнейшего анализа были взяты по 20 слов, характеризующиеся наибольшим уровнем совокупной частотности[1]. Далее был проведен квантитативный анализ единиц каждого подкорпуса («Твиттер» и СМИ). В результате анализа был присвоен класс подкорпуса в таблице выявленных тем. Если слова были распределены равномерно, то присваивался класс «Твиттер»+СМИ (колонка «Корпус» в таблице).

[1] С развернутым списком лексем и тем можно ознакомиться по ссылке: URL: https://docs.google.com/document/d/1OU56LZ867XQQnAn5i0_8mazCumyf-hyJ9wluBHOSIJU/edit?usp=sharing

Примеры тем, выделенных методом LDA
и методом выявления ключевых слов TF–IDF

ID	Лексические репрезентанты	Название темы	TF–IDF	Корпус
1	Карантин, Украина, коронавирус, закрывать, дома, Казахстан, самоизоляция, отменять, корона, паника, Киев, Беларусь, чемпионат, матч, команда, почему, футбол, побеждать, Алматы, выходить, клуб, украинец, Зеленский, читать, интересно, страшно	Эмоциональные реакции на пандемию и спорт	0.103507100046159	«Твиттер»
2	Случай, новый, Китай, заражение, число, здравоохранение, covid, страна, последний, скончаться, вспышка, заболевание, данные, воз, сутки, провинция, выявлять, мир, зафиксировать, всемирный, общий, ухань, власть, организация, пневмония, заражать	Начало распространения коронавирусной инфекции в мире	0.103419162819095	СМИ
3	Коронавирус, Италия, эпидемия, вирус, Москва, пандемия, заражать, умирать, смерть, заболевать, Испания, распространение, подробно, борьба, Франция, заражаться, Германия, Европа, Иран, тест, здоровье, статистика, итальянский, Петербург, Великобритания, Грузия	Распространение пандемии в мире	0.102927835638728	«Твиттер»
4	Режим, рейс, работа, вводить, Москва, ограничение, гражданин, школа, мера, транспорт, самоизоляция, турист, аэропорт, приостанавливать, общественный, россиянин, пассажир, граница, закрывать, дистанционный, власть, обучение, мероприятие, разрешать, территория, авиакомпания	Государственные ограничительные меры в пандемию	0.101949517675341	«Твиттер»
5	Маска, рука, просто, друг, нужно, магазин, понимать, делать, дом, помогать, никто,	Частная жизнь в условиях пандемии	0.101525484917406	«Твиттер»

ID	Лексические репрезентанты	Название темы	TF–IDF	Корпус
	выходить, продукт, бумага, очень, жить, ходить, туалетный, гречка, вообще, носить, мыть, купить, вода, находить, бояться			
6	Край, пациент, инфекция, коронавирусный, медицинский, больница, штаб, республика, умирать, оперативный, правительство, сообщение, помощь, федеральный, новосибирский, центр, случай, зарегистрировать, выздоравливать, говориться, врач, район, находиться, лечение, медик, округ	Госпитализация, лечение Covid-19 в РФ	0.101318170336452	СМИ
7	Компания, экономика, доллар, рынок, цена, поддержка, бизнес, рост, нефть, миллиард, процент, кризис, снижение, спрос, банк, уровень, развитие, объем, млрд, мера, программа, проект, экономический, эксперт, доход,	Экономика в пандемию	0.0988424006205264	СМИ + «Твиттер»
8	Вирус, вакцина, врач, исследование, ученый, препарат, заболевание, грипп, университет, лечение, животное, риск, институт, испытание, лекарство, болезнь, система, отмечать, иммунитет, первый, клинический, инфекция, лаборатория, наука, мир, центр	Исследования вируса и разработка вакцины	0.0969448467714175	СМИ + «Твиттер»
9	США, Трамп, президент, заявлять, американский, дело, информация, штат, СМИ, суд, происходить, отношение, борьба, несколько, Белый дом, издание, называть, представитель, Китай, пытаться, власть, давать, писать, заявление, мир	Политика в США в период пандемии	0.0962885262239954	СМИ
10	Страна, президент, Путин, глава, военный, голосование, совет, государство, депутат, принимать, российский, ЕС,	Политика РФ в период пандемии	0.0955722832614734	СМИ

ID	Лексические репрезентанты	Название темы	TF–IDF	Корпус
	ООН, выборы, поправка, конституция, сила, участие, право, комиссия, министр, правительство, заседание, партия, участок			
11	Онлайн, театр, фильм, показывать, становиться, проект, русский, ребенок, программа, концерт, история, хороший, зритель, музей, спектакль, фестиваль, проходить, самый, главный, культура, видео, песня, московский, роль, артист, участник	Культура в пандемию	0.0951839257944702	СМИ + «Твиттер»
12	Covid, мир, данные, коронавирус, последний, ситуация, портал, представить, страна, случай, сообщить, число, новый, заявить, здравоохранение, актуальный, скончаться, стать, отметить, организация, объявить, заражение, коронавирусный, инфекция, фиксировать, выявить	Статистика заболеваемости коронавирусом	0.093674010682404	СМИ

Так как алгоритмы на выходе выдают только списки слов с обозначением частотности и их взаимной встречаемости, на следующем этапе анализа было проведено маркирование класса тем. Существует ряд подходов, позволяющих решить данную задачу, их можно разделить на два типа. К первому относятся варианты автоматической разметки за счет готовых алгоритмов нейронных сетей [18], глубокого обучения [19], векторов [20], онтологий [21]. К преимуществам вариантов решения задачи первого типа можно отнести скорость работы, а к недостаткам – зависимость от корпуса, на котором была обучена модель. Если тексты, в которых осуществляется разметка, не совпадают с обученной моделью, то вероятность ошибки автоматического маркирования темы значительно увеличивается. Так как тексты новостей и «Твиттер» очень зависят от текущей информационной повестки, значения векторов и сам словарь могут меняться.

Подобным подходам противопоставляются экспертный анализ групп слов и обозначение тем. При этом исследователи при выборе имени темы либо используют метод «тренированной интроспекции», представляя авторскую интерпретацию семантической связи слов в сочетании с методиками лингвистического или литературоведческого анализа [22. С. 227], например, мотивного анализа художественного текста [23], либо расширяют список

экспертов, являющихся носителями языка и обладающих разным объемом специальных знаний. Есть мнение, что последний способ позволяет точнее определить эмпирическую значимость маркера темы, а в тематическом моделировании художественного текста экспертная оценка необходима [24]. В нашем исследовании мы опирались на экспертные оценки. В качестве экспертов выступили филологи, сотрудники лаборатории лингвистической антропологии ТГУ, а также студенты бакалавриата и магистратуры направления подготовки Фундаментальной и прикладной лингвистики, владеющие на разном уровне методами автоматического тематического моделирования. Всего было привлечено 12 экспертов. Эксперты работали, не контактируя друг с другом, время выполнения задания не ограничивалось, эксперты были свободны и в способе формулирования темы, имели возможность отказаться от формулировки темы, если она им казалась не определяемой. Отказов от определения темы не было, что свидетельствует о весьма высоком уровне интуитивно осознаваемой семантической общности лексем.

Наибольшей вариативностью отличалась *форма* представления семантического единства: эксперты выбирали однословные номинативные обозначения (*самоизоляция / локдаун / искусство / экономика*), словосочетания (*экономика в пандемию / начало пандемии / коронавирус в Китае / жизнь искусства в ковид / искусство в ковид*). Семантические варианты формулировок интерпретировались следующим образом: синоним-дублеты и квазисинонимы приводились к одному варианту, например: *ковид / Covid-19 / коронавирус / коронавирусная инфекция / новая коронавирусная инфекция* были обобщены *в «пандемия» искусство* и *культура* – в «*культура»*. Из вариантов, различающихся степенью и направлениями конкретизации номинаций единых концептов, при принятии результативного маркера темы избирались те, которые были представлены в большем количестве в оценках экспертов.

Отметим темы, формулировки разных экспертов которых обнаружили наибольшую согласованность. Это тема «экономика в *пандемию» (экономический кризис в пандемию / мировая экономика в пандемию / экономика в условиях коронавируса / экономическая обстановка / экономическая ситуация / меры экономической поддержки бизнеса в условиях пандемии / кризис/ развитие экономики в условиях пандемии / экономика / экономика во время пандемии / экономика в пандемию / экономика).* Как видим, три эксперта конкретизировали тему экономики в пандемию*: кризис, поддержка, мировой масштаб,* другие же варианты отличались формальным представлением единого концепта. Тема «статистика заболеваемости коронавирусом» была выделена но основании преобладания слова *статистика* в формулировках маркеров темы (были формулировки семантически связанные – *информация о распространении, заболеваемость*). Тема «культура в пандемию» в трех экспертных оценках, глубинно соотносясь с большинством оценок, интерпретировалась «через потребителя» (маркеры: *досуг в пандемию, сфера развлечений*).

Также значительным семантическим единством экспертные решения характеризовались при определении тем «политика в США в период пандемии», «политика РФ в период пандемии», «исследования вируса и разработка вакцины». При формулировании четвертой и пятой тем было принято решение принять более обобщенные формулировки «частная жизнь в условиях пандемии» и «государственные ограничительные меры в пандемию» при наличии единичных конкретизирующих тему вариантов «паника в ковид», «нехватка товаров».

Наибольшей вариативностью отличались маркеры первого тематического объединения: экспертами были отмечены темы спорта в пандемию и эмоциональных реакций на пандемию. Было решено объединить эти два аспекта в общей формулировке. В таблице представлены первые 26 лексических единиц, маркеры тем, их TF–IDF меры и название корпусов текстов. С более полными данными можно ознакомиться по ссылке[2], отметим, что эксперты оценивали полные списки слов, которые невозможно представить в статье из-за ограничения ее объема.

Далее был проведен содержательный сравнительный анализ: были выявлены направления совпадения и различия тематического моделирования событийного ряда в новостном дискурсе и дискурсе социальной сети в соответствии с принципом релевантности, по Т. ван Дейку [4. С. 125, 137].

Сравнение выявило как прогнозируемые аспекты несовпадения в тематическом представлении ситуации распространения коронавирусной инфекции в личностной коммуникации в Твиттере и институционально ориентированных СМИ, так и некоторую долю пересечения, что проявилось на уровне отдельных тем и в их интерпретации через ключевые слова и их связи.

Основное различие тематического моделирования ситуации пандемии заключается в преобладании в твитах личностной проекции пандемии и более выраженной в новостном потоке СМИ, в соответствии с принципом институциональной релевантности, темы государственных мер, направленных на профилактику и лечение заболевания, информированности населения о развитии пандемии, отражения политической жизни в стране и мире в данный период с фокусировкой на актуальном событийном ряде.

В текстах СМИ тема госпитализации, лечения Covid-19 репрезентируется в номинациях государственных органов (*правительство, федеральный центр, оперативный штаб*), административных локусов (*республика, район, новосибирский, округ, край*), медицинских учреждений (*больница, медицинский, госпиталь*), процесса госпитализации и лечения и их основных участников, медиков и пациентов (*пациент, врач, лечение, инфекция, медик, умирать, коронавирусный, врач, помощь, случай, зарегистрировать, выздоравливать, говорить, находиться*).

Тема «Статистика заболеваемости коронавирусом» отражает представление на страницах СМИ статистических данных о развитии пандемии в мире, России; репрезентирована в качестве наиболее частных словами *Covid, мир, коронавирус, коронавирусный, страна*), а также номинациями

действий, связанных с фиксацией и обнародованием статистических данных (*данные, портал, представить, сообщить, заявить, отметить, объявить, фиксировать, выявить*), номинаций самих данных (*последний, ситуация, случай, число, актуальный, скончаться, заражение, инфекция*).

Тема коронавирусной инфекции непосредственно соседствовала на страницах анализируемых СМИ с отражением других аспектов социальной и политической жизни РФ, мировой политической повестки. Данные проведенного анализа показывают актуальные темы политической жизни в стране и мире, которые делили рейтинги актуальности с темой пандемии. Как видно из лексических рядов темы «политика РФ в период пандемии», в фокусе политической активности РФ в данный период в находились события как международные, репрезентированные номинациями *ЕС, ООН*, так и внутрироссийские, репрезентируемые через номинацию событий, их акторов и действий: *страна, выборы, голосование, военный, совет, поправка, конституция, российский, участок. комиссия*, *президент, Путин, глава, министр, правительство, депутат, партия принимать, участие, заседание*.

Результаты анализа свидетельствуют также о тематической выделенности в новостном потоке в этот период американской политической жизни, репрезентированной в частотных лексемах, называющих органы власти, политических деятелей и действий, с ними связанных: *США, Белый дом, Вашингтон, американский, Трамп, президент, штат, заявлять, дело, информация, суд, происходить, отношение, борьб*а и др.

Сравнение второй и третьей тем, выделенных по наибольшим совокупным весам **в «Твиттере» и СМИ**, показывают, что, во-первых, наблюдается значительное пересечение лексических рядов в текстах сравниваемых дискурсов: это лексемы, называющие распространение пандемии, ее релевантные для общества и личности признаки («Твиттер» – *заражать, умирать, смерть, заболевать, распространение*; СМИ – *заражение, число, здравоохранение, последний, скончаться, заболевание, организация, пневмония, заражать*). Во-вторых, группы отличаются по фокусировке этой более общей темы. В текстах СМИ тема распространения пандемии связывается прежде всего с китайским локусом, а следовательно, и началом пандемии: *случай, новый, Китай, провинция, Ухань, КНР, Корея, Хубэй*, в тематическом объединении Твиттера обсуждается распространение пандемии по миру с фокусировкой наиболее пострадавших стран и регионов: *коронавирус, Италия, Испания, Франция, Германия, Европа, итальянский, Петербург, Великобритания, Грузия, Иран.*

Как представляется, одним из выраженных направлений дифференциации тематической фокусировки является противопоставление личностной («Твиттер») и социальной (новости) проекция пандемии.

В «**Твиттере»** три темы с наибольшим весом в тематической фокусировке пандемии объединяются общей направленностью ее отражения, которую мы определили как «частный человек перед лицом пандемии». В трех темах представлены разные аспекты переживания человеком не самой бо-

лезни, но вынужденных изменений всего привычного уклада жизни. В четвертой теме фокусируются государственные регулирующие меры в пандемию (*ограничение, режим*), связанные с ограничениями межгосударственных и внутрироссийских передвижений (*рейс, транспорт, турист, аэропорт пассажир, граница, закрывать, авиакомпания, приостанавливать, вводить*), организаций профессиональной деятельности и системы образования (*дистанционный, власть, работа, школа, обучение, мероприятие*). В пятой теме фокусируются изменения сугубо частной жизни, эмоциональная личностная реакция, вызванная ограничительными мерами, ожиданиями финансовой нестабильности и под.: *магазин, дом, дома, продукт, бумага, туалетный, гречка, вода, жить, носить, мыть, купить, находить, помогать, бояться очень, понимать.*

Как мы отмечали ранее, первая тема, характеризующаяся наибольшей совокупностью относительных частот, вызвала наибольшую трудность в поиске обобщающей номинации, однако своеобразие лексических репрезентаций личностной эмоциональной реакции на пандемию и связанные с ней ограничениями (*самоизоляция, отменять, корона, паника, паниковать, интересно, страшно, шутка, зараза, юмор*), фокусируются прежде всего относительно спорта: *отменять, чемпионат, матч, команда, футбол, побеждать, выходить, клуб, спорт, игра, турнир, лига, соревнование, олимпийский.*

Данные автоматического тематического моделирования выявили три темы, которые характеризуются общностью совокупных весов лексических репрезентантов в двух типах текстов, относимых к разным дискурсам, **СМИ** и **«Твиттер»**. Замечательно, что одна тема объединяет единицы, связанные с самой пандемией, – это тема изучения нового вируса и разработки вакцины, две другие – с экономическими и социальными аспектами пандемии, особыми условиями реализации культурных практик в условиях локдаунов.

Тема «Исследования вируса и разработка вакцины», характеризующаяся значительной плотностью семантически близких единиц, называющих объект изучения, результат разработки, субъектов процесса и институциональные органы в качестве наиболее частотных включает лексемы *вирус, вакцина, врач, исследование, ученый, препарат.*

Перестройка экономики в период пандемии была объектом регулирующих мероприятий со стороны государства и в то же время затронула самым непосредственным образом частного человека, что нашло отражение в практически равной частотности представленности темы в институциональных дискурсах СМИ и личностном дискурсе, репрезентированном твитами.

Тема «культура в пандемию», представляющая перестройку искусства в ситуации ограничений на публичные мероприятия (*онлайн, интернет, трансляция, запись*), в качестве ключевых наиболее частотных слов репрезентирована лексемами, называющими виды искусства, типы событий и их участников (*театр, спектакль, фестиваль, фильм, постановка, артист, концерт, режиссер, музей, актер* и др.).

Таким образом, беспрецедентный по размаху, общемировой масштаб новой коронавирусной инфекции, тяжелое течение болезни, высокий уровень смертности, обусловленные этим ограничения практически во всех сферах социальной жизни, в том числе ограничения передвижения в мире, необходимость строгих карантинных мероприятий внутри страны, лечение заболевших в медучреждениях – темы, объединившие новостной медийный и персональный интернет-дискурс «Твиттера» в стране в первый год пандемии.

Однако эти темы включаются в разные тематические единства в двух сравниваемых дискурсах. В «Твиттере» они объединяются, координируются с темами, раскрывающими личностную проекцию течения и переживания пандемии. Принципиальное отличие новостного контента в тематическом моделировании пандемии – ее выведение за пределы частной жизни, сочетание с представлением прежде всего социальных аспектов развития пандемии и аспектами политической жизни в стране и мире, непосредственно с пандемией не связанных.

Список источников

1. *Борьба* с инфодемией на фоне пандемии COVID-19: поощрение ответственного поведения и уменьшение пагубного воздействия ложных сведений и дезинформации: Совместное заявление ВОЗ, ООН, ЮНИСЕФ, ПРООН, ЮНЕСКО, ЮНЭЙДС, МСЭ, инициативы ООН «Глобальный пульс» и МФКК. URL: https://www.who.int/ru/news/item/23-09-2020-managing-thecovid-19-infodemic-promoting-healthy-behaviours-andmitigating-the-harm-from-misinformation-and-disinformation (дата обращения: 20.05.2021).

2. *Серегина Т.Н., Сухова С.К.* Информационные риски в условиях пандемии // Манускрипт. 2021. № 5. С. 940–944.

3. Инфодемия: существующие подходы к анализу паник, фобий, слухов, фейков во время эпидемий и предложения по борьбе с ними. URL: https://www.ranepa.ru/documents/monitoring/120-infodemiya.pdf (дата обращения: 01.04.2022).

4. *Дейк Т.А. ван.* Анализ новостей как дискурса // Т.А. ван Дейк. Язык. Познание. Коммуникация. Казань, 2000. С. 111–160.

5. *Добросклонская Т.Г.* Новостной дискурс как объект медиалингвистического анализа // Дискурс современных масс-медиа в перспективе теории, социальной практики и образования. Белгород, 2016. С. 13–22.

6. *Горошко Е.И.* «Чирикающий» жанр 2.0 Твиттер, или Что нового появилось в виртуальном жанроведении // Вестник Тверского государственного университета. 2011. № 3. С. 11–21.

7. *Горошко Е.И., Полякова Т.Л.* Политический твиттинг как новый жанр интернет-коммуникации // Вопросы психолингвистики. 2014. Вып. 1 (19). С. 92–103.

8. *Копцева В.А.* Жанр твиттинга в политическом дискурсе Г.А. Зюганова // Сибирский филологический журнал. 2016. № 1. С. 144–154.

9. *Гончарова Е.А.* Языковые характеристики англоязычного бизнес-твиттер как инструмента профессиональной коммуникации : дис. … канд. филол. наук. Пятигорск, 2021. 237 с.

10. *Садыков Д.И., Ахметьянова Н.А.* Распространение фейковых новостей во время пандемии COVID-19 // PHILOLOGICAL SCIENCES / «Colloquium-journal». 2020. № 8 (60). С. 78–79.

11. *Баринов Д.Н.* Медиавирус страха: особенности репрезентации российскими СМИ пандемии коронавирусной инфекции (COVID-19) в период первой волны (январь–июнь 2020 года) // Социодинамика. 2021. № 2. С. 73–86. doi: 10.25136/2409-7144.2021.2.35066

12. *Пестова М.Е., Сафонов Е.А.* Пандемия нового десятилетия: освещение темы коронавируса в СМИ // Медиасреда. 2020. № 17. С. 166–172.

13. *Мартыненко И.В., Стогова Е.С.* Коронавирус в повестке дня информационных агентств РИА «Новости» и Reuters // Вопросы теории и практики журналистики. 2021. Т. 10, № 2. С. 338–350.

14. *Карасик В.И.* Эпидемия в зеркале медийного дискурса: факты, оценки, позиции // Политическая лингвистика. 2020. № 2 (80). С. 25–34. doi: 10.26170/pl20-02-02

15. *Ерофеева И.В., Толстокулакова Ю.В., Муравьёв А.В.* Пандемия коронавируса в концептуальной сфере медиадискурса России и Китая: стратегия выживания // Вопросы теории и практики журналистики. 2021. Т. 10, № 1. С. 78–93. doi: 10.17150/2308-6203.2021.10(1)

16. *Charu C.* Aggarwal Machine Learning for Text. Springer, Mohegan Lake NY, USA, 2018. 565 p. doi: 10.1007/978-3-030-96623-2

17. *Deveaud R., SanJuan E., Bellot P.* Accurate and effective Latent Concept Modeling for ad hoc information retrieval // Document Numerique. 2014. Vol. 17, № 1. P. 61–84.

18. *Alokaili A., Aletras N., Stevenson M.* Automatic Generation of Topic Labels. 2020. URL: https://doi.org/10.1145/3397271.3401185 (дата обращения: 20.09.2023).

19. *Amparo C.B., Xu R.* Automatic Labelling of Topic Models Learned from Twitter by Summarisation // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. P. 618–624. doi: 10.3115/v1/P14-2101

20. *Kou W., Fang L., Baldwin T.* Automatic labelling of topic models using word vectors and letter trigram vectors // In Proceedings of the 11th Asian Information Retrieval Societies Conference (AIRS 2015). 2015. P. 229–240.

21. *Allahyari M., Kochut K.* Automatic Topic Labeling using Ontology-based Topic Models. URL: http://linkeddata.org/ (дата обращения: 20.09.2023).

22. *Митрофанова О.А.* Моделирование тематики специальных текстов на основе алгоритма LDA // Избранные труды XLII Международной филологической конференции. СПб. : Филологический факультет СПбГУ, 2014. С. 220–233.

23. *Шерстинова Т.Ю., Москвина А.Д., Кирина М.А., Карышева А.С., Колпащикова Е.О.* Тематическое моделирование русского рассказа 1900–1930: наиболее частотные темы и их динамика // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2022». Москва, 15–18 июня 2022 г. URL: https://www.dialog-21.ru/media/5790/sherstinovatyplusetal042.pdf/

24. *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

References

1. World Health Organisation. (2020) *Bor'ba s infodemiey na fone pandemii COVID-19: pooshchrenie otvetstvennogo povedeniya i umen'shenie pagubnogo vozdeystviya lozhnykh svedeniy i dezinformatsii* [Addressing the COVID-19 infodemic: promoting responsible behavior and reducing the harmful impact of misinformation and disinformation]. [Online] Available from: https://www.who.int/ru/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation (Accessed: 20.05.2021).

2. Seregina, T.N. & Sukhova, S.K. (2021) Informatsionnye riski v usloviyakh pandemii [Information risks in a pandemic]. *Manuskript*. 5. pp. 940–944.

3. RANEPA. (n.d.) *Infodemiya: sushchestvuyushchie podkhody k analizu panik, fobiy, slukhov, feykov vo vremya epidemiy i predlozheniya po bor'be s nimi* [Infodemic: existing approaches to the analysis of panics, phobias, rumors, fakes during epidemics and proposals for combating them]. [Online] Available from: https://www.ranepa.ru/documents/monitoring/120-infodemiya.pdf (Accessed: 01.04.2022).

4. van Dijk, T.A. (2000) *Yazyk. Poznanie. Kommunikatsiya* [Language. Cognition. Communication]. Translated from English. Blagoveshchensk: BGK im. I.A. Boduena de Kurtene. pp. 111–160.

5. Dobrosklonskaya, T.G. (2016) [News discourse as an object of medialinguistic analysis]. *Diskurs sovremennykh mass-media v perspektive teorii, sotsial'noy praktiki i obrazovaniya* [Discourse of Modern Mass Media in the Perspective of Theory, Social Practice and Education]. Proceedings of the 2nd International Seminar. Belgorod. 5–7 October 2016. Belgorod: Belgorod; Belgorod State University. pp. 13–22. (In Russian).

6. Goroshko, E.I. (2011) "Chirikayushchiy" zhanr 2.0 Tvitter ili chto novogo poyavilos' v virtual'nom zhanrovedenii ["Tweeting" genre 2.0 Twitter or what's new in virtual genre studies]. *Vestnik Tverskogo gosudarstvennogo universiteta*. 3. pp. 11–21.

7. Goroshko, E.I. & Polyakova, T.L. (2014) Politicheskiy tvitting kak novyy zhanr internet-kommunikatsii [Political tweeting as a new genre of Internet communication]. *Voprosy psikholingvistiki*. 1 (19). pp. 92–103.

8. Koptseva, V.A. (2016) Zhanr tvittinga v politicheskom diskurse G.A. Zyuganova [The genre of tweeting in political discourse G.A. Zyuganov]. *Sibirskiy filologicheskiy zhurnal*. 1. pp. 144–154.

9. Goncharova, E.A. (2021) *Yazykovye kharakteristiki angloyazychnogo biznes-tvitter kak instrumenta professional'noy kommunikatsii* [Linguistic characteristics of English-speaking business Twitter as a tool of professional communication]. Philology Cand. Diss. Pyatigorsk.

10. Sadykov, D.I. & Akhmet'yanova, N.A. (2020) Rasprostranenie feykovykh novostey vo vremya pandemii COVID-19 [Spread of fake news during the COVID-19 pandemic]. *PHILOLOGICAL SCIENCES / Colloquium-journal*. 8 (60) pp. 78–79.

11. Barinov, D.N. (2021) Mediavirus strakha: osobennosti reprezentatsii rossiyskimi SMI pandemii koronavirusnoy infektsii (COVID-19) v period pervoy volny (yanvar'-iyun' 2020 goda) [Media virus of fear: features of the Russian media's representation of the coronavirus infection (COVID-19) pandemic during the first wave (January-June 2020)]. *Sotsiodinamika*. 2. pp. 73–86. [Online] Available from: https://nbpublish.com/library_read_article.php?id=35066. doi: 10.25136/2409-7144.2021.2.35066

12. Pestova, M.E. & Safonov, E.A. (2020) Pandemiya novjgo desyatiletiya: osveshchenie temy koronavirusa v SMI [Pandemic of the new decade: coverage of the coronavirus topic in the media]. *Mediasreda*. pp. 166–172.

13. Martynenko, I.V. & Stogova, E.S. (2021) Koronavirus v povestke dnya informatsionnykh agentstv RIA "Novosti" i Reuters [Coronavirus on the agenda of the news agencies RIA Novosti and Reuters]. *Voprosy teorii i praktiki zhurnalistiki*. 2 (10). pp. 338–350.

14. Karasik, V.I. (2020) Epidemiya v zerkale mediynogo diskursa: fakty, otsenki, pozitsii [The epidemic in the mirror of media discourse: facts, assessments, positions]. *Politicheskaya lingvistika*. 2 (80). pp. 25–34. doi: 10.26170/pl20-02-02

15. Erofeeva, I.V., Tolstokulakova, Yu.V. & Murav'ev, A.V. (2021) Pandemiya koronavirusa v kontseptual'noy sfere mediadiskursa Rossii i Kitaya: strategiya vyzhivaniya [Coronavirus pandemic in the conceptual sphere of media discourse in Russia and China: survival strategy]. *Voprosy teorii i praktiki zhurnalistiki*. 1 (10). pp. 78–93. doi: 10.17150/2308-6203.2021.10(1)

16. Charu, C. (2018) *Aggarwal Machine Learning for Text*. Mohegan Lake, NY: Springer. doi: 10.1007/978-3-030-96623-2

17. Deveaud, R., SanJuan, E. & Bellot, P. (2014) Accurate and effective Latent Concept Modeling for ad hoc information retrieval. *Document Numerique*. 1 (17). pp. 61–84.

18. Alokaili, A., Aletras, N., & Stevenson, M. (2020) *Automatic Generation of Topic Labels*. /doi: 10.1145/3397271.3401185

19. Amparo, C.B. & Xu, R. (2014) Automatic Labelling of Topic Models Learned from Twitter by Summarisation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore. June 2014. Association for Computational Linguistics. pp. 618–624. doi: 10.3115/v1/P14-2101

20. Kou, W., Fang, L. & Baldwin, T. (2015) Automatic labelling of topic models using word vectors and letter trigram vectors. *AIRS 2015*. Proceedings of the 11th Asian Information Retrieval Societies Conference. Brisbane. 2–4 December 2015. Springer. pp. 229–240.

21. Allahyari, M. & Kochut, K. (n.d.) *Automatic Topic Labeling using Ontology-based Topic Models*. [Online] Available from: http://linkeddata.org/ (Accessed: 20.09.2023).

22. Mitrofanova, O.A. (2014) Modelirovanie tematiki spetsial'nykh tekstov na osnove algoritma LDA [Modeling the subject matter of special texts based on the LDA algorithm]. *Proceedings of the 42nd International Conference. Saint Petersburg. 11–15 March 2014*. Saint Petersburg: Saint Petersburg State University. pp. 220–233. (In Russian).

23. Sherstinova, T.Yu. et al. (2022) [Thematic modeling of the Russian story 1900–1930: the most frequent topics and their dynamics]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computer Linguistics and Intelligent Technologies]. Proceedings of the International Conference Dialog 2022. Moscow. 15–18 June 2022. [Online] Available from: https://www.dialog-21.ru/media/5790/sherstinovatyplusetal042.pdf/ (In Russian),

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

Информация об авторах:

Резанова З.И. – д-р филол. наук, заведующая кафедрой общей, компьютерной и когнитивной лингвистики, заместитель заведующего Лабораторией лингвистической антропологии Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: rezanovazi@mail.ru

Степаненко А.А. – заведующий лабораторией «Когнитивные исследования языка» Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: stepanenkone@mail.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

Z.I. Rezanova, Dr. Sci. (Philology), head of the Department of General, Computational and Cognitive Linguistics; deputy head of the Laboratory for Cognitive Studies of Language, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: rezanovazi@mail.ru

A.A. Stepanenko, head of the Laboratory for Cognitive Studies of Language, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: stepanenkone@mail.ru

The authors declare no conflicts of interests.

Статья поступила в редакцию 14.06.2023;
одобрена после рецензирования 02.10.2023; принята к публикации 26.12.2023.

The article was submitted 14.06.2023;
approved after reviewing 02.10.2023; accepted for publication 26.12.2023.