

ВЫЯВЛЕНИЕ ЗНАЧИМЫХ ТЕМ ПРИ ПОМОЩИ АЛГОРИТМА LDA НА МАТЕРИАЛЕ НЕМЕЦКОГО МЕДИАДИСКУРСА*

Для цитирования: *Корышев М. В., Хохлова М. В., Куликова Л. А., Мазин К. В. Выявление значимых тем при помощи алгоритма LDA на материале немецкого медиадискурса // Немецкая филология в Санкт-Петербургском государственном университете. 2024. Вып. 14. С. 418–433.
<https://doi.org/10.21638/spbu33.2024.122>*

Методы тематического моделирования позволяют получить представление о тематическом наполнении текстов и выявить скрытые семантические структуры. Каждый текст может быть представлен несколькими темами, таким образом также можно определить схожесть текстов и более широко — общие тенденции, характерные для текстов, нацеленных на конкретную аудиторию. Целью статьи является формирование круга тем, которые интересуют читателей, на материале современного немецкоязычного дискурса, посвященного разным сферам жизни. В работе продемонстрированы результаты построения нескольких моделей тематического моделирования при помощи алгоритма латентного размещения Дирихле (LDA) применительно к текстам немецкоязычных источников: журнала «Zeitschrift für Ideengeschichte», который посвящен вопросам истории и развития политических, религиозных, философских и литературных идей и мыслей, и студенческого периодического издания «Moritz. Magazin» Грайфсвальдского университета. Выделенные ключевые слова получили экспертную оценку. Результаты применения алгоритма LDA показывают, что с течением времени в журнале «Zeitschrift für Ideengeschichte» можно заметить переход от более узких тем к более широким. Проведенный анализ продемонстрировал достаточно низкую схожесть между текстами разных годов данного журнала, однако в рамках выборок одного года были выявлены похожие тексты согласно мере TF-IDF. Несмотря на первоначально заявленную разноплановость рассмотренных журналов, именно политическая составляющая является для них общей. Однако в издании «Moritz. Magazin» данная тема прослеживается более отчетливо, в то время как в статьях журнала «Zeitschrift für Ideengeschichte» политический настрой подается опосредованно через отсылку к определенным мыслителям и темам. В ходе исследования был предварительно определен перечень тем, интересующих представителей двух значимых кругов современной Германии, относящихся к образованному университетскому сообществу, и тем самым был описан некоторый идейный ландшафт страны.

Ключевые слова: тематические маркеры, ключевые слова, латентное размещение Дирихле (LDA), медиадискурс, немецкий язык.

* Исследование выполнено за счет гранта Российского научного фонда № 24-28-00937, <https://rscf.ru/project/24-28-00937/>.

IDENTIFICATION OF SIGNIFICANT THEMES USING THE LDA ALGORITHM ON THE MATERIAL OF GERMAN MEDIA DISCOURSE*

For citation: Koryshev M. V., Khokhlova M. V., Kulikova L. A., Mazin K. V. Identification of significant themes using the LDA algorithm on the material of German media discourse. *German Philology in St. Petersburg State University*, 2024, iss. 14, pp. 418–433. <https://doi.org/10.21638/spbu33.2024.122> (In Russian)

Topic modeling methods allow obtaining insights into the thematic content of texts and identifying latent semantic structures. Each text can be represented by multiple topics, thereby enabling the identification of text similarity and, more broadly, common trends specific to texts targeting a particular audience. The goal of this paper is to form a set of themes that interest readers based on German discourse covering various aspects of life. The results of constructing several thematic models using the Latent Dirichlet Allocation (LDA) algorithm are demonstrated in relation to the texts of the German journals “Zeitschrift für Ideengeschichte” which covers the history and development of political, religious, philosophical, and literary ideas and thoughts, and the student periodical “Moritz.Magazin” of the Greifswald University. The extracted keywords received expert evaluation. The results show that over time, the “Zeitschrift für Ideengeschichte” transits from narrower to broader topics. The analysis demonstrated relatively low similarity between texts from different years of this journal. However, texts published in the same year share common features and this similarity was identified according to the TF-IDF measure. Despite the initially declared diversity of the examined journals, the political component is common to them, with the “Moritz.Magazin” displaying it more prominently, while in the “Zeitschrift für Ideengeschichte” articles, the political orientation is conveyed indirectly through references to prominent thinkers and themes. During the research, a preliminary set of topics interesting to representatives of two significant circles in contemporary Germany, belonging to the educated university community, was defined, thus describing a certain intellectual landscape of the country.

Keywords: topics, keywords, Latent Dirichlet Allocation (LDA), media discourse, German language.

1. Введение

Тематическое моделирование относится к современным и перспективным направлениям компьютерной лингвистики, его алгоритмы позволяют определить значимые для того или иного источника смысловые элементы. Каждый текст может быть описан

* The study was supported by the Russian Science Foundation, project no. 24-28-00937, <https://rscf.ru/en/project/24-28-00937/>.

несколькими темами (топиками¹), которые могут свидетельствовать о схожести текстов. В свою очередь, топики отражены в виде набора ключевых слов (термов, или терминов), которые позволяют получить представление о содержании текстов и выявлять скрытые семантические структуры. Для решения задач, связанных с тематическим моделированием, в работах был предложен целый ряд алгоритмов: латентно-семантический анализ (LSA) [Deerwester et al. 1990; Landauer et al. 1998], вероятностный латентно-семантический анализ (pLSA) [Hofmann 1999], латентное размещение Дирихле (LDA) [Blei et al. 2013], неотрицательное матричное разложение (или неотрицательная матричная факторизация) [Lee, Seung 1999], структурное тематическое моделирование (STM) [Roberts et al. 2013] и другие.

В исследованиях наибольшее внимание уделяется новостным текстам и сообщениям в социальных сетях. Как и многие другие задачи прикладной лингвистики, изучение роли тематических маркеров и ключевых слов при помощи автоматических методов первоначально проводилось на англоязычном материале. Применительно к анализу политического дискурса выделение тематических маркеров позволяет продемонстрировать, какие темы пользуются популярностью, дать представление об основных идеях, которые заключены в тексте, а также — шире — выявить закономерности, лежащие в основе гетерогенного, на первый взгляд, материала. Отдельное внимание уделяется ключевым политическим проблемам и их отражению в публицистике, в том числе немецкоязычной. Например, отношение к закону о возобновляемых источниках энергии (Erneuerbare-Energien-Gesetz) было исследовано авторами на материале немецких газет при помощи алгоритма STM в работе [Dehler-Holland et al. 2021]. Представления о южных странах (Португалия, Италия, Греция, Испания) анализировались при помощи этого же метода в немецкоязычной прессе с 1946 по 2009 г. на материале газеты «Die Welt» [Küsters, Garrido 2020].

В статье представлены результаты автоматического выделения ключевых слов, которые могут быть рассмотрены в качестве маркеров темы, а также построения тематических моделей применительно к текстам немецкоязычных журналов. Были выделе-

¹ Вслед за [Кирина 2022] мы предлагаем использовать термин «топик» для обозначения результатов выполнения процедур тематического моделирования и термин «тема» для отсылки к содержательному наполнению документов (статей).

ны наиболее частотные темы, которым уделяется внимание в рассматриваемой публицистике, а также проанализирована схожесть статей между собой.

2. Материал и методика

2.1. Отбор текстов

В качестве материала были выбраны два немецкоязычных источника: журналы «Zeitschrift für Ideengeschichte» и «Moritz. Magazin». Их разноплановость представляется сильной стороной для целей настоящей работы: отдельного внимания заслуживает выявление тех тем, которые в них пересекаются, поскольку эти пересечения важны для определения неявных векторов развития умонастроений современного германского общества в целом. Журнал «Zeitschrift für Ideengeschichte», выпускаемый с 2007 г., представляет собой пример немецкоязычного научно-просветительского журнала о философии и истории. Выходящий четыре раза в год «Zeitschrift für Ideengeschichte» публикует эссе по различным предметным областям, задается вопросами об изменчивости идей, будь то философских, религиозных, политических или литературных, и их роли в современном обществе и мышлении людей. Журнал рассматривает как наиболее знаменитых немецких мыслителей, таких как, например, Карл Маркс и Мартин Хайдеггер, так и выдающихся зарубежных деятелей, повлиявших на мир, таких как Дьёрдь Лукач и Роберт Оппенгеймер. У каждого выпуска есть тематическая направленность, в которой публикуются несколько эссе. В разделе «Архив» размещены различные документы, посвященные какой-то конкретной идее, личности, которые затем подробно разбираются одним из авторов журнала. Сильными сторонами журнала являются междисциплинарная направленность и высокое качество подготовки текста, что делает его интересным как для непрофессионалов, так и для гуманитариев, специализирующихся на тематике тех или иных выпусков.

«Moritz.Magazin» — студенческое периодическое издание, которое с 1998 г. освещает события из жизни Грайфсвальдского университета. В первую очередь журнал ориентирован на студентов, однако его статьями регулярно интересуются и другие жители города. Поскольку редакция журнала состоит из студентов, которые

учатся на разных направлениях, выпуски охватывают максимально широкий спектр тем. Ключевой из них с самого начала существования журнала является развитие университетской политики: самые актуальные и спорные вопросы о реорганизации структуры университета, студенческих стипендиях, бытовых условиях обучения, а также о вопросах управления университетом. Не менее интересны для читателей и «неуниверситетские» разделы журнала: политика, новости города и в особенности молодежная культура. «Moritz.Magazin» дает возможность студентам взглянуть на мир за пределами лекционного зала и библиотеки, узнать о новинках кино, книг и местных театральных постановок. В год выпускается в среднем шесть выпусков, однако материалы выходят не только в печатном формате: у издания имеются также собственные подкаст и видеоблог, в рамках которых редакция берет интервью у студентов на разнообразные темы.

Нами были отобраны статьи с 2013 по 2018 г. в случае журнала «Zeitschrift für Ideengeschichte» и с 2013 по 2023 г. для журнала «Moritz.Magazin», что позволяет прицельно рассмотреть те центральные для редакции материалы, которые, с учетом сказанного выше, как раз и должны служить созиданию общеевропейского гуманитарного пространства через интеграцию в него собственно германских умонастроений. Объем собранных текстов составил около 1 млн словоупотреблений для каждого журнала, что и обуславливает разный временной охват привлекавшихся для анализа изданий.

2.2. Методы

Цель нашего исследования заключается в выделении при помощи автоматических процедур возможных тем, которые получают экспертную оценку, на материале двух разноплановых источников.

Предварительная обработка текстов осуществлялась при помощи следующей процедуры: была проведена лемматизация с использованием теггера HanTa [Wartena 2019] (который демонстрирует наиболее высокие результаты при обработке немецкоязычных текстов), а также удалялись стоп-слова и служебные части речи с привлечением библиотеки NLTK. Дополнительно были отфильтрованы редкая лексика и высокочастотные слова, которые могут

«зашумлять» темы: не рассматривались слова, которые встречаются менее чем в пяти документах и более чем в 75 % документов. Нами было принято решение сохранить в текстах числовые комплексы, так как в случае текстов на исторические и политические темы они могут быть значимыми: указание года может служить отсылкой к некоторой теме, которая затрагивается в выпуске. При помощи алгоритма LDA, реализованного в библиотеке gensim [Řehůřek, Sojka 2010], были построены модели, которые позволили выделить наиболее частотные темы для рассматриваемых источников.

Для отбора оптимального количества тем была использована мера когерентности C_v (coherence measure), которая демонстрирует наиболее удачные результаты при решении данной задачи [Röder et al. 2015]. Значения меры варьируются от 0 до 1: чем выше значение, тем большая когерентность наблюдается между словами и, соответственно, тем более удачной является модель или выделяемые темы оказываются более интерпретируемыми.

3. Результаты

3.1. Когерентность для подкорпусов

В качестве диапазона количества возможных топиков был выбран интервал от 4 до 20. Графики когерентности позволяют определить оптимальное количество выделенных тем для рассмотренных коллекций статей (рис. 1).

Отметим существующий риск того, что с увеличением количества единиц возрастает их «дробность», что может затруднять интерпретируемость тематических составляющих текстов. Наибольшее значение меры когерентности наблюдалось для шести топиков для статей журнала «Zeitschrift für Ideengeschichte» и для восьми топиков для статей журнала «Moritz.Magazin», которые и были далее нами рассмотрены более подробно.

3.2. Выделенные темы

Результатом выполнения процедуры тематического моделирования являются списки: 1) топиков с соответствующими им ключевыми словами с весами; 2) документов с соответствующими им топиками (они сопровождаются указанием вероятности того, что

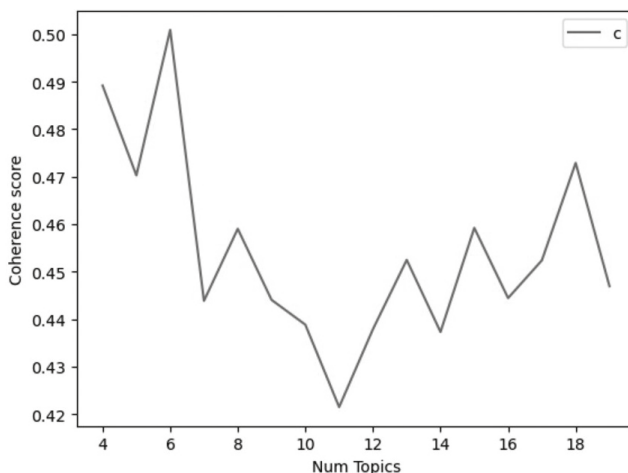


Рис. 1. График меры когерентности для статей из журнала «Zeitschrift für Ideengeschichte» [Скриншот из Google Colab]. Ось X — количество тем, ось Y — значения меры когерентности

документ принадлежит теме). Например, топик под номером пять для текстов журнала «Zeitschrift für Ideengeschichte» сформирован при помощи следующих имен собственных:

(‘Marx’, 0.012080658),
 (‘Burke’, 0.004636104),
 (‘Koselleck’, 0.0039058619),
 (‘Blumenberg’, 0.0033647367).

На следующем этапе работы автоматически выделенные ключевые слова получили экспертную оценку (были использованы списки, состоящие из 50 слов для каждого тематического кластера). Для корпуса текстов «Ideengeschichte» были предложены следующие темы с выделенными на основе ключевых слов тематическими маркерами:

1. WECHSELSCHICKSALE DER EUROPÄISCHEN ZIVILISATION

Наиболее значимые тематические маркеры: *Lukác, Plessner, Jasper, Melancholie, Warburg, Untergang, Krankheit.*

2. KULTURGESCHICHTLICHE WISSENSKONTEXTE EUROPAS ZWISCHEN LOKAL UND GLOBAL

Наиболее значимые тематические маркеры: *Winckelmann, theologisch, Marquard, Auerbach, Panofsky, räumlich, Holz, Zweig, Garten.*

3. WELTKRIEGE UND EUROPÄISCHES KULTURGEDÄCHTNIS

Наиболее значимые тематические маркеры: *Tillich, 1914, Soldat, Auerbach, Koselleck, Burckhardt, Front, Adorno, Arendt, Hannah, Schlacht, 1918, George.*

4. ALPTRÄUME UND ABGRÜNDE EINER (UN)GEREGELTEN WELT ALS KULTURGESCHICHTLICHES ERBE

Наиболее значимые тематические маркеры: *Unsel, Adorno, Jahn, Kracauer, Augustinus, Verwaltung, Schmerz, Konzentrationslager.*

5. LINKES UND RADIKALES IN DER BILDUNGSBÜRGERLICHEN HERMENEUTIK

Наиболее значимые тематические маркеры: *Marx, Burke, Koselleck, Engels, Kapital, sowjetisch, DDR, Dialektik, Mommsen, kommunistisch, Leiche, Leibniz, islamisch, Horkheimer.*

6. GESCHICHTSPHILOSOPHIE UND POSTMODERNER INDIVIDUALISMUS IN GROSSEN UND KLEINEN SCHRITTEN

Наиболее значимые тематические маркеры: *Burckhardt, Foucault, Eliot, Koselleck, Adorno, Kantorowicz, Szene, Sartre, Kulturgeschichte, Hermeneutikum, heroisch, mythisch.*

Для каждой из тем в качестве значимых было отмечено разное количество лексем, что можно объяснить разным весом, который они вносят в соответствующую тему. В таблице приведены результа-

**Таблица. Распределение документов по темам
для корпуса «Zeitschrift für Ideengeschichte»**

№	Тема	Количество документов
1.	WECHSELSCHICKSALE DER EUROPÄISCHEN ZIVILISATION	5
2.	KULTURGESCHICHTLICHE WISSENSKONTEXTE EUROPAS ZWISCHEN LOKAL UND GLOBAL	3
3.	WELTKRIEGE UND EUROPÄISCHES KULTURGEDÄCHTNIS	2
4.	ALPTRÄUME UND ABGRÜNDE EINER (UN)GEREGELTEN WELT ALS KULTURGESCHICHTLICHES ERBE	4
5.	LINKES UND RADIKALES IN DER BILDUNGSBÜRGERLICHEN HERMENEUTIK	5
6.	GESCHICHTSPHILOSOPHIE UND POSTMODERNER INDIVIDUALISMUS IN GROSSEN UND KLEINEN SCHRITTEN	5

ты распределения выпусков журнала «Zeitschrift für Ideengeschichte» по принадлежности к темам с вероятностью более 0,5.

Можно отметить равномерное распределение документов по первой, пятой и шестой темам, в то время как только два выпуска журнала относятся к третьей теме. При этом, как указывалось выше, некоторые документы могут относиться к нескольким темам одновременно, поскольку алгоритмы тематического моделирования выполняют так называемую нечеткую кластеризацию. Примером может служить текст выпуска «Kilcher, Košenina, Siegel, Reitinger, Herz, Raulff» (весна 2013), содержание которого соответствует первой, пятой и шестой темам с вероятностями 0,22, 0,56 и 0,10 соответственно. В текстах из статей журнала «Zeitschrift für Ideengeschichte» было выделено большое количество имен собственных, что потребовало дальнейшего экспертного уточнения. Например, в ряде тем упоминался немецкий мыслитель и философ Т. Адорно — представитель умеренно-левого политического крыла, который придерживался конформистской идеологии, приближающейся к центристам. Однако в целом основные темы, которым посвящены выпуски, успешно выделяются при помощи алгоритма.

Для корпуса текстов «Moritz.Magazin» были выбраны следующие темы, обозначенные связанным с ними лексическим наполнением:

1. SOZIALES

Наиболее значимые тематические маркеры: *Frau, Lebensmittel, Euro, Mitglied, Projekt, erhalten, unterstützen, Verein, Hilfe, Gewalt, Geflüchtete.*

2. POLITIK AUSSERHALB DES UNIVERSITÄTSALLTAGS

Наиболее значимые тематические маркеры: *politisch, Ungehorsam, zivil, Studierende, Aktion, Ort, handeln, Gesellschaft, Veranstaltung, Protest, Innenstadt, unterschiedlich, verschieden.*

3. RUND UM DAS STUDIUM

Наиболее значимые тематические маркеры: *Studium, lernen, Student, Seminar, Konferenz, Studierende, Projekt, Gremium, Semester, Institut, Erfahrung.*

4. ALLTAG AUSSERHALB DES UNTERRICHTS

Наиболее значимые тематические маркеры: *Ehe, Dompassage, Einkaufszentrum, Bürgerinitiative, Heirat, Fürsorge, Lebensweg, Liegenschaft, Greit, heiraten, Innenstadt, Glaubwürdigkeit, Eigenheim, Internetaufttritt, Einkaufsmöglichkeit, Umbau, Konkurrenz.*

5. BILDUNG UND FINANZEN

Наиболее значимые тематические маркеры: *Euro, Geld, Professor, AStA, nutzen, Angebot, Antrag, Studierendenparlament, Bildung, Stupist.*

6. FREIZEIT AKTIV

Наиболее значимые тематические маркеры: *Kunst, Festival, Ort, Künstlerin, Kultur, City, Verein, Freiheit, Ausstellung.*

7. POLITIK AN DER UNIVERSITÄT

Наиболее значимые тематические маркеры: *Studierende, studentisch, politisch, Stupa, Studierendenschaft, Partei, Wahl, Prozent, AStA, wählen, Person, Gremium, Studierendenausschuss, Senat, Staat, Amt, Regierung, Veranstaltung, Ziel.*

8. FREIZEIT INTELLEKTUELL

Наиболее значимые тематические маркеры: *Geschichte, Buch, Film, Bild, Theater, Musik, Stück, erzählen, Idee, Kunst, lesen.*

В случае данного издания список выявленных ключевых слов содержит большое количество глаголов и только несколько имен собственных. По сравнению с результатами, полученными для текстов из журнала «*Zeitschrift für Ideengeschichte*», обозначение тем вызвало меньше затруднений ввиду большей семантической однородности термов, которые были предложены алгоритмом.

3.3. Сходство между текстами

Нами было вычислено сходство по косинусной мере на TF-IDF векторах при помощи функции *TfidfModel* и матрицы сходства *MatrixSimilarity* (на рис. 2 приведен пример для статей журнала «*Zeitschrift für Ideengeschichte*», по осям цифрами от 0 до 23 обозначены выпуски журнала). Мера принимает значения от 0 (в случае отсутствия какого-либо подобия) до 1 (в случае идентичных по своему содержанию документов). Элементы матрицы, таким образом, представляют собой значения данной меры для попарно сравниваемых между собой выпусков журнала.

Несмотря на однородный характер материала, статьи журнала «*Zeitschrift für Ideengeschichte*» в основном гетерогенны по своей структуре и при попарном сравнении демонстрируют низкое сходство. При анализе текстов по годам наибольшее сходство (значение меры равно 0,47) было продемонстрировано выпусками «*Welthistoriker — Dilettant — Burckhardt*» (весна 2018) и «*Ich*» (осень 2018). В обоих случаях статьи объединяет тема индивидуализма. В выпуске «*Ich*» речь идет

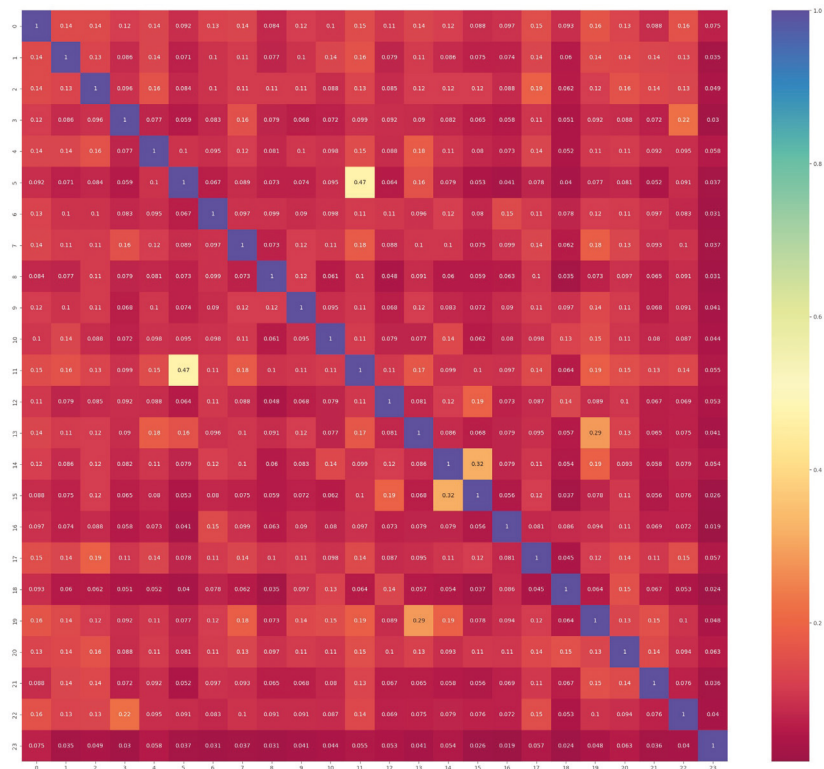


Рис. 2. Тепловая карта для статей из журнала «Zeitschrift für Ideengeschichte»
[Скриншот из Google Colab]

об индивидуальном «я», в то время как в весеннем номере говорится о поиске собственного пути в какой-то сфере деятельности, который сам человек считает для себя важным. Таким образом, в статьях затрагиваются одни и те же вопросы, которые, однако, освещаются по-разному: я — личность, или индивидуальность vs свобода следовать своему выбору (в том числе та автономия, которую дает эта свобода).

4. Заключение

В рамках исследования была сделана попытка выявить наиболее частотные темы для текстов статей журналов «Zeitschrift für Ideengeschichte» и «Moritz.Magazin», опубликованных в 2010-е гг., при

этом анализировался примерно одинаковый объем текстового материала. Исползованный алгоритм LDA позволил выделить значимые тематические маркеры, которые в целом представляются ожидаемыми в свете данной выше общей характеристики этих изданий. Вместе с тем анализ показал, что темы этих изданий не выводятся напрямую из тематических маркеров, взятых по отдельности, — тематика выводима из анализа их в совокупности. В текстах обоих источников выводится идеальный образ интеллектуала-индивидуалиста, который свободно ориентируется в контексте, заданном европейским культурным каноном, при этом последовательно проводится мысль о включенности культурной проблематики в глобальный политический контекст. Представляется, что этот образ динамичен: если в нацеленном на публику, включающую зрелых интеллектуалов, «*Zeitschrift für Ideengeschichte*» мы находим законченный образ либерального индивидуала, то применительно к читательской аудитории журнала «*Moritz.Magazin*» такой индивидуалист только складывается как продукт целенаправленного политического воспитания. Ср. наблюдаемый по ключевым словам переход от идеала совместного действия и партиципации (например, *AStA*, *Gremium*, *Bürgerinitiative*, *Heirat*) к идеологии диверсификации и разнообразия (например, *unterschiedlich*, *verschieden*) и обсуждению дискурса недоброжелательности и зависти (например, *Greit*, *Konkurrenz*), что косвенно подводит аудиторию к принятию идеи атомизации общества как данности. Таким образом, несмотря на первоначально заявленную разноплановость журналов, именно косвенно предъявляемая аудитории политическая составляющая является для них общей. При этом в издании «*Moritz.Magazin*» линия политического воспитания прослеживается более отчетливо, в то время как в статьях журнала «*Zeitschrift für Ideengeschichte*» политический настрой подается опосредованно через отсылку к определенным мыслителям и темам (что также свидетельствует о важности упоминаемых имен для общеевропейского контекста).

В целом алгоритм LDA продемонстрировал успешные результаты, однако требуется дополнительное исследование полученных данных для такого сложного публицистического материала. Требуется дальнейшее увеличение объема данных, привлечение иных источников, а также оценка тем при помощи ключевых слов, выявленных другими алгоритмами. В дальнейшем планируется также проследить изменение тем, которые затрагиваются в журнальных публикациях, с течением времени.

Источники

Moritz.Magazin. URL: <https://webmoritz.de/moritz-magazin/> (дата обращения: 18.01.2024).

Zeitschrift für Ideengeschichte. URL: <https://www.wiko-berlin.de> (дата обращения: 18.01.2024).

Литература

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.

Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2013. Vol. 3 (4–5). P. 993–1022.

Deerwester S., Dumais S. T., Furnas G. W., Landauer Th. K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. Vol. 41 (6). P. 391–407.

Dehler-Holland J., Schumacher K., Fichtner W. Topic Modeling Uncovers Shifts in Media Framing of the German Renewable Energy Act // Patterns. 2021. Vol. 2. P. 100–169.

Hofmann Th. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99). New York: Association for Computing Machinery, 1999. P. 50–57.

Küstners A., Garrido E. Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper Die Zeit (1946–2009) // Journal of Contemporary European Studies. 2020. Vol. 28: (4). P. 477–493. <https://doi.org/10.1080/14782804.2020.1784112>

Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25. P. 259–284.

Lee D., Seung H. Learning the Parts of Objects by Non-Negative Matrix Factorization // Nature. 1999. Vol. 401. P. 788–791.

Řehůřek R., Sojka P. Software Framework for Topic Modelling with Large Corpora // Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta: University of Malta, 2010. P. 46–50.

Roberts M., Stewart B., Tingley D., Airolidi E. The Structural Topic Model and Applied Social Science // Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation. 2013. URL: <https://projects.iq.harvard.edu/files/wcfia/files/stmnips2013.pdf> (дата обращения: 18.01.2024).

Röder M., Both A., Hinneburg A. Exploring the Space of Topic Coherence Measures // Proceedings of the Eight International Conference on Web Search and Data Mining, Shanghai, February 2–6. Shanghai: ACM, 2015. P. 399–408.

Wartena Ch. A probabilistic morphology model for German lemmatization // Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers / German Society for Computational Linguistics & Language Technology. Erlangen: Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. P. 40–49. <https://doi.org/10.25968/opus-1527>

References

Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2013, vol. 3 (4–5), pp. 993–1022.

Deerwester S., Dumais S. T., Furnas G. W., Landauer Th. K., Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, vol. 41 (6), pp. 391–407.

Dehler-Holland J., Schumacher K., Fichtner W. Topic Modeling Uncovers Shifts in Media Framing of the German Renewable Energy Act. *Patterns*, 2021, vol. 2, pp. 100–169.

Hofmann Th. Probabilistic Latent Semantic Indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*. New York: Association for Computing Machinery, 1999, pp. 50–57.

Kirina M. A. A Comparison of Topic Models Based on LDA, STM and NMF for Qualitative Studies of Russian Short Prose. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2022, vol. 20, no. 2, pp. 93–109. (In Russian)

Küstners A., Garrido E. Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper Die Zeit (1946–2009). *Journal of Contemporary European Studies*, 2020, vol. 28 (4), pp. 477–493. <https://doi.org/10.1080/14782804.2020.1784112>

Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis. *Discourse Processes*, 1998, vol. 25, pp. 259–284.

Lee D., Seung H. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 1999, vol. 401, pp. 788–791.

Řehůřek R., Sojka P. Software Framework for Topic Modelling with Large Corpora. *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, University of Malta, 2010, pp. 46–50.

Roberts M., Stewart B., Tingley D., Airolidi E. The Structural Topic Model and Applied Social Science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, 2013. Available at: <https://projects.iq.harvard.edu/files/wcfia/files/stmnips2013.pdf> (accessed: 18.01.2024).

Röder M., Both A., Hinneburg A. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eight International Conference on Web Search and Data Mining*, Shanghai, February 2–6. Shanghai, ACM, 2015, pp. 399–408.

Wartena Ch. A probabilistic morphology model for German lemmatization. *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, German Society for Computational Linguistics & Language Technology. Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019, pp. 40–49. <https://doi.org/10.25968/opus-1527>

Корышев Михаил Витальевич

доцент кафедры сопоставительного изучения языков и культур СПбГУ,
кандидат филологических наук

Адрес: Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Mikhail V. Koryshev

Associate Professor of Comparative Study of Languages and Cultures Department,
St. Petersburg State University, Candidate of Philological Sciences

Address: 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

ORCID: <https://orcid.org/0000-0001-8946-4431>

E-mail: m.koryshev@spbu.ru

Хохлова Мария Владимировна

доцент кафедры математической лингвистики СПбГУ,
кандидат филологических наук

Адрес: Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Maria V. Khokhlova

Associate Professor of Mathematical Linguistics Department, St. Petersburg State
University, Candidate of Philological Sciences

Address: 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

ORCID: <https://orcid.org/0000-0001-9085-0284>

E-mail: m.khokhlova@spbu.ru

Куликова Любовь Алексеевна

лаборант-исследователь СПбГУ

Адрес: Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Liubov A. Kulikova

Research Assistant, St. Petersburg State University

Address: 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

ORCID: <https://orcid.org/0009-0004-6438-1161>

E-mail: st110434@student.spbu.ru

Мазин Константин Владимирович

лаборант-исследователь СПбГУ

Адрес: Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Konstantin V. Mazin

Research Assistant, St. Petersburg State University

Address: 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

ORCID: <https://orcid.org/0009-0001-7325-1973>

E-mail: st110126@student.spbu.ru

Статья поступила в редакцию 15 февраля 2024 г.

Принята к публикации 26 апреля 2024 г.