УДК 81'42:82-3:004. ББК Ш300
ГРНТИ 16.31.21. Код ВАК 5.9.8

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ КАК ИНСТРУМЕНТ АНАЛИЗА ХУДОЖЕСТВЕННОГО ТЕКСТА

Колмогорова А. В.
Национальный исследовательский университет «Высшая школа экономики»
(Санкт-Петербург, Россия)
ORCID ID: https://orcid.org/0000-0002-6425-2050

Залевская Е. Д.
Национальный исследовательский университет «Высшая школа экономики»
(Санкт-Петербург, Россия)
ORCID ID: https://orcid.org/0009-0009-0929-722X

Аннотация. Статья посвящена проблеме эвристической продуктивности использования метода компьютерного тематического моделирования для филологического анализа художественного текста.

Анализируются результаты применения алгоритма Латентного размещения Дирехле (LDA) для поиска интертекстуальных связей мотивов в двух подкорпусах художественных текстов: 62 текстах разного жанра (рассказы, очерки, повести, критические статьи), принадлежащих перу С. Довлатова, с одной стороны, и 35 художественных произведениях, которые в одном из писем Т. Уржумовой писатель перечислил как произведения, которые оказали на него воздействие и которые должен прочитать каждый. Примененный алгоритм выявил 20 тем (топиков), по которым были распределены все тексты. Каждый полученный топик – это цепочка слов с весами значимости для реализации данной темы. В результате сопоставления текстов и тем были выявлены три соответствия «текст – тема». Одной общей теме принадлежат тексты в каждой из трех следующих групп: 1) роман Б. Пильняка «Голый год» и рассказ С. Довлатова «У реки»; 2) роман Г. Уэльса «Машина времени», повесть Э. Хемингуэя «Старик и море» и рассказ С. Довлатова «Эмигранты»; 3) рассказ А. Грина «Комендант порта» и очерк С. Довлатова «Мы говорим на разных языках».

Дальнейший филологический анализ позволил выявить пересечения мотивов в данных группах произведений.

Проведенное пилотное исследование показало, что методы компьютерного анализа текста, в том числе на основе машинного обучения, могут стать для филолога инструментом разведывательного поиска, направляя экспертную интуицию по пути, намеченному алгоритмом за счет обработки больших корпусных массивов.

Ключевые слова: художественный текст; метод компьютерного тематического моделирования; мотив; интертекстуальность; С. Довлатов

Благодарности: в данной научной работе использованы результаты проекта «Текст как Big Data: моделирование конвергентных процессов в языке и речи цифровыми методами», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2023 году.

Для цитирования: Колмогорова, А. В. Компьютерное моделирование как инструмент анализа художественного текста / А. В. Колмогорова, Е. Д. Залевская. – Текст : непосредственный // Филологический класс. – 2023. – Т. 28, № 2. – С. 22–33.

COMPUTER-ASSISTED MODELING AS AN INSTRUMENT FOR FICTION TEXT ANALYSIS

Anastasia V. Kolmogorova
National Research University Higher School of Economics (Saint Petersburg, Russia)
ORCID ID: https://orcid.org/0000-0002-6425-2050

Ekaterina D. Zalevskaya
National Research University Higher School of Economics (Saint Petersburg, Russia)
ORCIDE ID: https://orcid.org/0009-0009-0929-722X

Abstract. The article investigates the issue of heuristic productivity of using the method of computer-assisted topic modeling for philological analysis of fiction text.

The study analyzes the results of applying the algorithm of Latent Placement Dirichlet (LDA) for searching intertextual connections of motifs in two sub-corpora of fiction texts: 62 texts of different genres (stories, essays, novels, critical articles) belonging to S. Dovlatov, on the one hand, and 35 fiction works, which the writer listed in one of the letters to T. Urzhumova as the works that had deeply influenced him and should be read by everybody. The algorithm has revealed 20 themes (topics), into which all the texts were distributed. Each topic obtained was a chain of words with weights of significance for the realization of that topic. As a result of the comparison of the texts and the topics, three "text – topic" correspondences were discovered. The texts in each of the following three groups belong to one common topic: 1) B. Pilyniak's novel "The Bare Year" and Dovlatov's story "By the River"; 2) G. Wells's novel "The Time Machine", E. Hemingway's story "The Old Man and the Sea" and Dovlatov's story "Emigrants"; 3) A. Grin's story "The Commandant of the Port" and Dovlatov's essay "We Speak Different Languages".

Further philological analysis demonstrated the intersection of motifs in these groups of works of fiction.

The pilot study under consideration has shown that methods of computer-assisted text analysis, including those based on machine learning, can become a philologist's tool for experimental search, guiding the expert intuition along the path outlined by the algorithm via processing large corpus arrays.

Keywords: fiction text; computer-assisted topic modeling method; motif; intertextuality; S. Dovlatov

Acknowledgments: This research paper uses the results of the project "Text as Big Data: Modeling Convergent Processes in Language and Speech by Digital Methods", implemented as part of the HSE University Basic Research Program in 2023.

For citation: Kolmogorova, A. V., Zalevskaya, E. D. (2023). Computer-Assisted Modeling as an Instrument for Fiction Text Analysis. In Philological Class. Vol. 28. No. 2, pp. 22–33.

Введение

С развитием слабого и сильного искусственного интеллекта у филологов и лингвистов появились новые инструменты для анализа текстов, в том числе художественных.

Хотя использование количественных методов для получения качественных выводов о стилистической и семантической специфике художественного произведения имеет давнюю традицию в отечественной квантитативной лингвистике [Андреев 2016; Андреев 2019] и стилометрии [Мартыненко 2021], переход к использованию моделей машинного и глубинного обучения с целью получения интерпретируемых результатов для дальнейшего филологического анализа текста нетривиален для мировоззрения филолога и филологических рутин в целом.

Наряду с признанием необходимости применения цифровых методов в традиционно экспертных областях науки, прежде всего – гуманитарных, нам, исследователям, нелегко признать тот факт, что в ряде случаев возможно делегировать некоторые из своих профессиональных полномочий другому когнитивному агенту – искусственному интеллекту (ИИ). Такое делегирование может происходить по трем осям [Вахштайн 2021: 135–137]: ось определения ситуации, ось принятия решения и ось реализации решения. Так, представляется, что мы можем вполне довериться алгоритмам на оси определения ситуации, чтобы получить общее представление о кластерах слов («темах»), имеющих статистическую значимость для текста. Тем не менее у нас, экспертов, пока нет оснований передавать ИИ полномочия принятия решения и/или его реализации – например, на основе проведенного алгоритмом стилометрического анализа или вопреки ему исследователь

сам делает окончательный вывод об атрибуции текста [Skorinkin, Orekhov 2023] и /или сообщает, скажем, экспертному сообществу о том, что данный текст не стоит включать в перечень произведений автора X.

Цель данной публикации состоит в том, чтобы на основе отдельного кейса обосновать идею о том, что именно для определения «ситуации» – предоставления исследователю некоторых статистически валидных системных данных о лексико-семантической структуре некоторого текста – инструменты на основе машинного и глубинного обучения могут быть полезными филологу, став основой для дальнейшей экспертной интерпретации.

Основная польза такого применения ИИ, по-видимому, состоит в том, что аффордансы [Gibson 1986], предоставляемые текстом искусственному интеллекту, отличаются от тех, которые тот же самый текст считает «позволительными» для человека-интерпретатора. В дальнейшем тексте мы постараемся проиллюстрировать эту идею.

Проектируя дизайн исследования, мы выдвинули гипотезу о том, что книги, которые читал писатель на протяжении своей жизни, влияют на мотивную структуру его собственных произведений. Для того чтобы ее проверить, мы использовали два корпуса текстов: тексты 35 книг, которые в своем известном письме Тамаре Уржумовой [Список Сергея Довлатова] С. Довлатов перечисляет как книги, которые повлияли на него – с одной стороны, и тексты самого Довлатова – с другой. К этим корпусам мы применили метод тематического моделирования на основе одного из алгоритмов машинного обучения. Полученные же результаты постарались интерпретировать с точки зрения интертекстуальных пересечений мотивов у С. Довлатова и его любимых авторов. В итоге у нас получился своеобразный опыт описания писателя как читателя.

Тема в филологических и лингвистических штудиях. Понятие мотива

Предваряя мотивный анализ, опосредованный компьютерным тематическим моделированием, рассмотрим основные «реперные точки» относительно понятий «мотив» и «тема».

Данный обзор ни в коем случае не претендует на всеохватность, ибо наполнение данных терминов в литературоведении и лингвистике – сюжет чрезвычайно обширный и многоаспектный.

Цель же данной структурной части скромна – обозначить некоторые вехи в понимании данных текстовых категорий в соответствующих предметных сферах для того, чтобы сравнить их затем с результатами работы модели на основе методов машинного обучения и представить попытку перебросить некий эвристический мостик между первыми и вторыми.

Так, говоря о теме литературного произведения, Б. В. Томашевский отмечает: «В художественном выражении отдельные предложения, сочетаясь между собой по их значению, дают в результате некоторую конструкцию, объединенную общностью мысли или темы. Тема (о чем говорится) является единством значений отдельных элементов произведений» [Томашевский 1999: 116]. При этом тема должна обладать признаками субъективной привлекательности для читателя и вызывать у него эмоциональный отклик [Там же: 117]. В такой широкой трактовке тема приближается к понятию мотива. По словам И. В. Силантьева, «наряду с фабулой и сюжетом, тема – ближайшая к мотиву категория» [Силантьев 1999: 49].

В концепции мотива Б. М. Гаспарова основным его свойством является повторяемость: «...мотив, раз возникнув, повторяется множество раз, выступая при этом каждый раз в новом варианте, новых очертаниях и во все новых сочетаниях с другими мотивами. При этом в роли мотива может выступать любой феномен, любое пятно – событие, черта характера, элемент ландшафта, любой предмет, произнесенное слово, краска, звук и т. д.; определяет мотив, – это его репродукция» [Гаспаров 1993: 30–31].

Идея исследователя о диалектическом единстве смыслового инварианта мотива и его способности к вариативному воплощению в разных частях текста или разных текстах приводит литературоведов к идее систематики мотивов, способных образовывать гнезда благодаря структурным отношениям, существующим между ними [Жолковский,

Щеглов 1986: 120] (ср.: мотивема – алломотив). Как следствие того, что «мотив, взятый не в системе "фабула-сюжет", а в системе "текст-смысл", утрачивает свою специфическую связь с событием как основной единицей фабульного ряда повествования» [Силантьев 1999: 51–52], становясь категорией текста, он полноправно входит в систему координат теории интертекстуальности: «мотивы репрезентируют смыслы и связывают тексты в единое смысловое пространство» [Там же: 52].

Если мотив близок к теме, способен образовывать системные отношения, выходящие за пределы одного текста, инвариантно-вариативен, следовательно, на уровне формального выражения он также должен обнаруживать себя некоторым системным образом в виде номинационных цепочек разной длины. Изучением таких цепочек и их свойств занимается лингвистика. Единицами номинационных цепочек являются слова и словосочетания, связанные внутри одной цепочки отношениями референтного (единство референта) и сигнификативного тождества (эквивалентность или близость значения) [Матвеева 1990: 21–22].

Существует сложившаяся лингвистическая традиция экспертного выявления таких систем номинационных цепочек для моделирования тематического содержания текста. Можем ли мы автоматизировать эту задачу, чтобы затем приступить собственно к мотивному филологическому анализу? Существующие модели тематического моделирования текстовых данных претендуют на этот функционал.

Использование компьютерных методов тематического моделирования для анализа художественных текстов

Тематическое моделирование – метод, разработанный в компьютерных науках для кластеризации формально выраженных элементов текста, чаще всего – слов.

Алгоритм строит вероятностную тематическую модель коллекции текстовых документов, которая описывает каждый документ дискретным вероятностным распределением на множестве тем, а каждую тему – дискретным вероятностным распределением на множестве слов. Наряду со словами могут использоваться словосочетания, теги, категории и даже нетекстовые сущности [Булатов 2020: 14].

Наиболее распространенным способом осуществления тематического моделирования является применение алгоритма Latent Dirichlet Allocation (LDA) и его производных. Однако его мощным конкурентом сегодня становится модель на основе BERT – BERTopic [Grootendorst 2022].

В тематическом моделировании, темы не определяются явно, они рассматриваются как латентные наборы слов, которые наиболее часто встречаются в текстах корпуса. Важно, что тексты сравниваются не друг с другом, а с этими наборами слов / темами и приписываются к этим темам с различной вероятностью [Nikolenko et al. 2017: 89].

Подобные модели давно и успешно используются для анализа блогосферы [Ritter et al. 2010], социальных сетей [Quercia et al. 2012], научных статей [Jelisavčić et al. 2012], новостей [Koltsov et al. 2018], политического дискурса [Jacobs & Tschötschel 2019].

В последние несколько лет данный метод вторгся и в филологические пределы.

Так, исследовав при помощи алгоритма латентного размещения Дирихле (LDA) объемную коллекцию из 890 драматических произведений французских драматургов, опубликованных в период с 1610 по 1810 гг., К. Счёч установил, что в целом (топики) хорошо отражают жанровую специфику и позволяют с высокой точностью кластеризовать тексты пьес согласно их жанру: комедии, трагедии, трагикомедии [Schöch 2017].

На материале русского языка О. А. Митрофанова при помощи алгоритма LDA моделировала темы в романе «Мастер и Маргарита» М. А. Булгакова [Митрофанова 2019]. Анализируя результаты, исследователь отмечает, что выделенные алгоритмом темы в целом соответствуют основным сюжетным линиям и могут быть использованы для описания идиостиля автора.

Активно в этом же направлении работает на основе корпуса русского рассказа начала XX в. группа Т. Ю. Шерстиновой. Так, исследователи [Шерстинова и др. 2021] построили, используя тот же алгоритм LDA, 9 тематических моделей (по 3 выборки разного размера

для каждого из периодов) для рассказов: 1) начала XX века до 1913 г. включительно, 2) военно-революционного периода (1914–1922) и 3) раннесоветского периода (1923–1930) – и выявили частотные для каждого периода темы. Авторы констатируют, что темы отличаются по разным временным периодам, что позволяет считать их тематико-стилистическими маркерами анализируемых коллекций текстов наряду с более традиционными квантитативными мерами анализа текстов.

Таким образом, проведенное нами исследование, с одной стороны, вписывается в формирующуюся традицию филологических исследований, проводимых цифровыми методами, а с другой – обладает элементом новизны: мы используем тематическое моделирование для поиска интертекстуальных связей мотивов двух групп художественных текстов.

Методология исследования

Материалом для анализа послужили 31 произведение мировой литературы, которые С. Довлатов охарактеризовал как «книги, которые стоит прочесть» и «книги, которые мне нравятся». В этот список входят тексты русских, британских, французских и американских авторов (список приводится в Приложении). Среди авторов: Ф. Достоевский, А. Куприн, А. Грин, Е. Замятин, Г. Уэллс, Г. де Мопассан и т. д.

Второй подкорпус составили 62 текста С. Довлатова, написанные в период с 1974 по 1990 гг. и собранные методом сплошной выборки из пятитомника, вышедшего в издательстве «Азбука»: рассказы, повести, очерки, критические статьи.

Все тексты прошли предобработку. Они были представлены в табличном формате с тремя колонками: name – название текста, год издания и автор; author – метку other получали тексты «не-Довлатова», соответственно, метку Dovlatov – тексты С. Довлатова; text – собственно тексты, где каждая строка соответствовала одному тексту (рис. 1).

	name	author	text
0	Прелести культуры.Михаил Зощенко.1934	other	Я всегда симпатизировал центральным убеждениям...
1	Былое и думы.Александр Герцен.1868	other	Н П. Огареву В этой книге всего больше говори...
2	Зима тревоги нашей.Джон Стейнбек.1961	other	Читателям, которые ...
3	Машина времени.Герберт Уэллс.1895	other	1. ИЗОБРЕТАТЕЛЬ Путешественник по Времен...
4	Морские рассказы.Виктор Конецкий	other	Морские сны Тихая жизнь 15.09.69 Идем из Синга...

Рис. 1. Образец представления текстового материала

Тексты были приведены к нижнему регистру, токенизированы, из них удалены знаки препинания и стоп-слова, проведена лемматизация при помощи pymorphy2.

Для проведения тематического моделирования использовалась библиотека **gensim**. Это популярная открытая библиотека для тематического моделирования, в которой есть модель – LDA. Для тематического моделирования из лемматизированного текста был создан словарь, при помощи метода **filter_extremes** отфильтрованы слова, встречающиеся слишком часто или слишком редко в текстах. Затем был создан корпус в виде «мешка слов» (bag of words).

Для тематического моделирования, помимо создания корпуса и словаря, необходимо указать количество «обходов», которые будет делать алгоритм (чем больше, тем точнее и медленнее будет создаваться модель), и количество тем, которые мы хотим выделить. Мы установили гиперпараметры следующим образом: 20 топиков и 10 «обходов» (passes).

Каждый документ был представлен в виде «мешка слов» (Bag of words), а затем к нему был применен метод **get_document_topics.**

data

```
local/lib/python3.9/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `sh
 should_run_async(code)
```

name	author	text	text_processed	topic_20	probability_20
Прелести ультуры.Михаил Зощенко.1934	other	Я всегда симпатизировал центральным убеждениям...	[симпатизировать, центральный, убеждение, эпох...	12	0.362519
Былое и думы.Александр Герцен.1868	other	Н П. Огареву В этой книге всего больше говори...	[н, п., огарев, книга, говориться, личность, о...	5	0.708167
Зима тревоги нашей.Джон Стейнбек.1961	other	Читателям, которые ...	[читатель, стать, доискиваться, какой, реальны...	10	0.880703

Рис. 2. Фрагмент таблицы, полученной в результате работы алгоритма

В итоге была получена таблица, в которой к уже существующим трем колонкам были добавлены еще три: text_processed – собственно слова (термы), кластеризованные алгоритмом и составляющие тему (топик); topic_20 – номер темы из 20-ти выделенных, которая имеет для данного текста наибольший вес; probability_20 – вероятность, с которой данная тема является самой важной для данного текста (рис. 2). Теперь, используя данные из колонок 2 и 5 (тип корпуса: «не-Довлатов» или «Довлатов» и номер самой важной темы текста), мы провели своеобразное картирование, чтобы узнать, есть ли такие «важные» темы, которые встречаются и в текстах корпуса «Довлатов», и в текстах корпуса «не-Довлатов» (или, лучше сказать, «писатели Довлатова-читателя»).

Результаты и обсуждение

Применив библиотеку seaborn, мы получили график (рис. 3), где по горизонтали отложены авторы (Довлатов (Dovlatov) / не-Довлатов (other)), а по вертикали – темы (их номера). Стало понятно, что в двух подкорпусах совпадают только три темы: 0, 7 и 9.

В таблице 1 представлено по 10 термов, составляющих каждый из данных трех топиков, для каждого терма указан его вес.

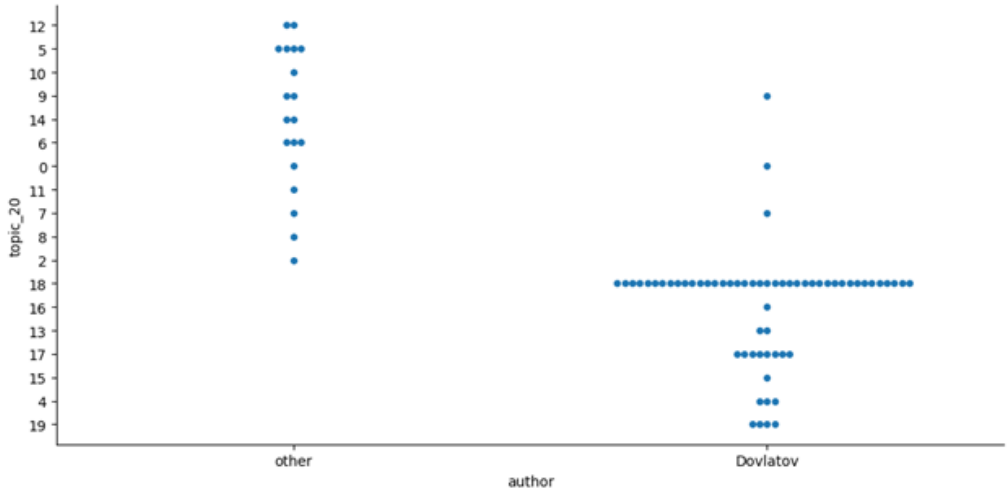

Рис. 3. Дистрибуция тем (топиков) по подкорпусам «Довлатов», «не-Довлатов»

Таблица 1. Темы, совпавшие в двух подкорпусах («Довлатов» и «не-Довлатов») и их термы с весами

тема	Вес, терм 0	Вес, терм 1	Вес, терм 2	Вес, терм 3	Вес, терм 4	Вес, терм 5	Вес, терм 6	Вес, терм 7	Вес, терм 8	Вес, терм 9
0	0.007*"иван"	0.007 "товарищ"	0.005"город"	0.004*"андрей"	0.004*"уйти"	0.004*"семён"	0.004*"отец"	0.004*"глеб"	0.004*" земля"	0.003*"степь"
7	0.013*"комендант"	0.006*"капитан"	0.005*" моряк"	0.004*" фон"	0.004*"тренер"	0.003*"вечер"	0.003*"надеяться"	0.003*"парень"	0.003*" буфет"	0.003*"рейс"
9	0.018*"старик"	0.017*"рыба"	0.007*" вода"	0.007*"лодка"	0.005*" море"	0.005*"солнце"	0.005*"мальчик"	0.004*"лес"	0.004*" акула"	0.004*"темнота"

В следующей таблице представлены значения вероятности, с которой тексты, имеющие ту или иную тему в качестве главной, принадлежат ей. Мы видим, что роман Бориса Пильняка «Голый год» с очень высокой вероятностью принадлежит теме 0, а вероятность того, что рассказ С. Довлатова «У реки» принадлежит ей же, – в два раза меньше. Обратная ситуация наблюдается для рассказа А. Грина «Комендант порта» и очерка «Мы с вами говорим на разных языках» С. Довлатова: для первого вероятность принадлежности теме 7 равна 44%, а для второго – 99%. Теме 9 практически с одинаковой, очень высокой, степенью вероятности принадлежат и роман Г. Уэллса, и повесть Э. Хемингуэя; рассказ Довлатова принадлежит этой теме с вероятностью вполовину меньшей.

Таблица 2. Вероятность принадлежности текста теме в двух подкорпусах текстов: текстов из довлатовского списка для чтения («не-Довлатов») и в текстах, написанных самим С. Довлатовым

Тема	Произведение не-Довлатова	Вероятность	Произведение Довлатова	Вероятность
0	Голый год. Борис Пильняк. 1922	0.99	У реки. 1983	0.45
7	Комендант порта. Александр Грин. 1933	0.44	Мы с вами говорим на разных языка. 1976	0.99
9	Машина времени Г. Уэллс	0.94	Эмигранты. 1985	0.48
	Старик и море. Эрнест Хемингуэй. 1952	0.99		

Проведем филологический анализ интертекстуальных пересечений тем.

Тема 0. Сюжет рассказа Довлатова «У реки» строится на том, что молодой человек по имени Федя, получив отказ от своей юной возлюбленной Зиночки, идет топиться на реку, но, войдя в воду, замечает воришку, уносящего прочь единственные брюки несчастного влюбленного. Федя пускается в погоню, вместе с милиционером хватает вора и становится поселковым героем, вернув тем самым благосклонность Зиночки. Таким образом, в тексте хорошо просматривается тема резкой и случайной смены социальной роли – из самоубийцы в героя. Трудно сравнивать роман Б. Пильняка с коротким текстом Довлатова,

но мотив упомянутого социального перевертыша красной нитью проходит и через роман: Донат Ратчин, наследник состоятельного купеческого рода, примыкает к красным и безжалостно крушит отчий дом (из богатого наследника превращается в агрессивного революционера); мещанка Олечка Кунц, делопроизводитель пролетарского исполкома, пылает страстью к комиссару Лайтису, но арестована как враждебный элемент (из возлюбленной комиссара превращается во враждебный революции элемент, в арестантку); Наталья Ордынина, княгиня, становится революционеркой и т. д.

Тема 7. Оба произведения: А. Грина «Комендант порта» и С. Довлатова «Мы говорим на разных языках» – посвящены теме, которую можно было бы обозначить как «человек, идущий навстречу другим людям»: у Грина это старик Тильс, ставший своеобразной ниточкой, связующей всех моряков, приходящих в порт судов; у Довлатова это Фрицас Маркузас, немецкий врач, участвовавший в молодые годы в революционном движении, прошедший войну и в мирное время помогающий спортсменам бороться с травмами. Оба героя живут в «мужском мире»: Тильс – среди моряков, Маркузас – среди рабочих, затем – среди солдат, позже – среди спортсменов, которых лечит.

Анализ показывает, что в обоих текстах присутствует мотив важности общения, некоего общего языка для становления мировоззрения человека или коллектива людей.

Герой Грина, старик Тильс, ходит от судна к судну, от одной компании моряков к другой, где-то над ним подшучивают, где-то привечают, где-то гонят, но он – своего рода социальный медиатор: рассказывает последние новости, вспоминает, расспрашивает, желает здоровья и удачи. Когда настает момент сообщить официантке печальную новость о смерти возлюбленного, все чувствуют, что выполнить ее сможет только Тильс. После его смерти все понимают, что старик был чем-то очень важным для городка, для порта – заменить его никто не сможет. Его ценность в том, что он создавал нечто наподобие «общего языка» между всеми. Эта потеря становится осязаемой, когда один из местных парней пытается взять на себя роль Тильса (пример 1):

(1) *– Нет, нет, – ответил с палубы, не обижаясь на дурака, Ластон. – Подделка налицо. Никогда твоя пасть не спросит как надо о том, «был ли хорош рейс».*

У Довлатова тот же мотив раскрывается «в обратном порядке»: несмотря на то, что для протагониста (Ф. Маркузас) и антагониста (Гейнц фон Книбуш) родным языком является немецкий, прожив разную жизнь, совершив в ней разные выборы, они в итоге «говорят на разных языках» (пример 2):

(2) *Гранд-отель в Мюнхене. На лестнице беседуют двое пожилых мужчин.*

– Так мы увидимся? – спрашивает Гейнц фон Книбуш. – Не забывай, мы старые приятели. Мы говорим на одном языке.

– О нет, ты ошибаешься, Гейнц. Мы говорим на разных языках, – отвечает Маркузас.

Тема 9. Для «Машины времени» и короткого рассказа Довлатова «Эмигранты» общим оказывается мотив перемещения в иное пространство: у Герберта Уэллса это перемещение изобретателя из XIX в. в город будущего; у Довлатова – мнимое перемещение двух случайных приятелей из советского Ленинграда в Голландию (которая, на самом деле, не более чем район Ленинграда – Новая Голландия). Интересно, что символом попадания в «иное» в обоих текстах становится солнце, сравните (примеры 3–5 – у Уэллса, 6–7 – у Довлатова):

(3) *Пока я мчался таким образом, ночи сменялись днями, подобно взмахам крыльев. Скоро смутные очертания моей лаборатории исчезли, и я увидел солнце, каждую минуту делавшее скачок по небу от востока до запада, и каждую минуту наступал новый день.*

(4) Скоро я заметил, что полоса, в которую превратилось солнце, колеблется то к северу, то к югу – от летнего солнцестояния к зимнему, – показывая, что я пролетал более года в минуту, и каждую минуту снег покрывал землю и сменялся яркой весенней зеленью.

(5) Наконец я отвел от него глаза и увидел, что завеса града прорвалась, небо прояснилось и скоро должно появиться солнце.

(6) Солнце вставало неохотно. Оно задевало фабричные трубы. Бросалось под колеса машин на холодный асфальт. Блуждало в зарослях телевизионных антенн. В грязном маленьком сквере *проснулись одновременно Чикваидзе и Шаповалов.*

(7) Дома обступили маленький сквер. Бледное солнце вставало у них за плечами. Остатки ночной темноты прятались среди мусорных баков.

Терм *солнце* является частью данной темы с весом 0.005 (таблица 1). У Уэллса солнце появляется, чтобы маркировать смену времен во время полета в будущее (пример 3–4) и затем – обозначить момент попадания в «иное» (пример 5). У Довлатова появление солнца также символизирует «озарение» протагонистов о том, что они случайно оказались за границей. Интересен и параллелизм в описании восприятия героями «иного» у Уэллса и Довлатова, сравните у Довлатова (пример 8):

(8) Друзья шли по набережной. Свернули на людную улицу. Поблескивали витрины. Таяло мороженое. Улыбались женщины и светофоры.

– Посмотри, благодать-то какая! – неожиданно воскликнул Шаповалов.

– Живут неплохо, – поддакнул Чикваидзе.

– А как одеты!

– Ведь это – Запад!

– Кругом асфальт! Полно машин! А солнце?!

у Уэллса (пример 9):

(9) Подбежавший человек показался мне удивительно прекрасным, грациозным, но чрезвычайно хрупким существом… я был весь увешан гирляндами цветов и окружен волнующейся толпой людей, облаченных в светлые, нежных расцветок одежды, сверкавших белизной обнаженных рук и смеявшихся и мелодично ворковавших.

Мир «иного» воспринимается как олицетворение счастья, где все улыбаются (улыбались женщины и светофоры…; смеявшихся…), хорошо одеты (а как одеты!; облаченных в светлые, нежных расцветок одежды).

Своеобразным антиподом прекрасного «иного» становится образ темноты, олицетворяющей «свой мир» (терм темнота имеет вес 0.004, таблица 1). Так, когда герои Довлатова проснулись после ночной драки (их реальный и обыденный мир) и вознамерились узнать, где они находятся, нарратор констатирует:

(10) Остатки ночной темноты прятались среди мусорных баков.

С наступлением темноты изобретатель Уэллса может, наконец, погрузиться в себя и осмыслить то, что он увидел в «новом прекрасном мире»:

(11) Пока я сидел в сгущавшейся темноте, мне казалось, что этим простым объяснением я разрешил загадку мира и постиг тайну прелестного маленького народа.

Что касается интертекстуальной связи между проанализированными выше текстами и произведением Э. Хемингуэя «Старик и море», то она проявилась, по-видимому, в теме смены дня и ночи как символа движения во времени и пространстве: как путешественник во времени Уэллса замечает свое движение по тому, как сменяют друг друга солнце и темнота, как Шаповалов и Чикваидзе, прожив банальную в своей обыденности ночь (пьяная драка, знакомство, похмельный сон на куче щебня), оказываются с восходом солнца в «новой жизни» в Новой Голландии, так и старик и рыба движутся через время и пространство, не имея других вех, кроме смены темноты солнцем и снова прихода темноты. На рисунке 4 показано распределение частот морфем *солнц** и *темн** по мере развертывания текста «Старик и море» – они идут попарно на протяжении всего текста.

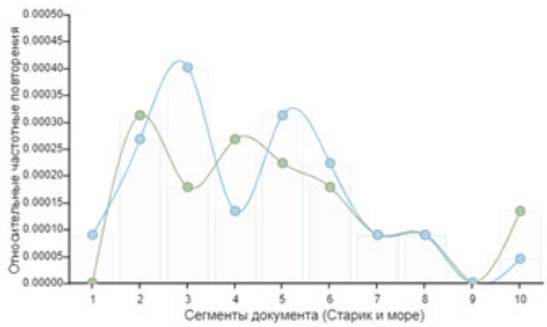

Рис. 4. Распределение частот употребления морфем солнц- и темн- по тексту повести Э. Хемингуэя «Старик и море»

Таким образом, проведенный анализ дает основания предполагать, что в коротких текстах С. Довлатова проявляются некоторые мотивы, почерпнутые им из прочитанных в период своего становления произведений художественной литературы. Эти мотивы иногда могут находить причудливое воплощение, узнать их нелегко, тем более – обнаружить интертекстуальные связи между текстами таких разных писателей, как С. Довлатов и А. Грин, Г. Уэллс, Б. Пильняк или Э. Хемингуэй. На наш взгляд, в этом и состоит эвристическая ценность применения моделей ИИ для анализа больших текстовых данных – результаты могут стать направляющим вектором для дальнейших, уже филологических, изысканий.

Заключение

В ходе исследования мы ставили своей целью найти в творчестве Сергея Довлатова его черты как читателя, ответив на вопрос, что из прочитанного им и в какой форме могло проявиться в его текстах. Оттолкнувшись от предположения о том, что мотив может проявляться как повторяющаяся тема и потому его можно будет узнать, проанализировав статистически важные лексические единицы, мы провели тематическое моделирование текстов, составляющих достояние писателя как читателя, с одной стороны, и написанных им, с другой. Полученный результат частично подтвердил нашу гипотезу: из полученных двадцати тем три были с разной вероятностью определены как ключевые для текстов как Довлатова, так и его любимых писателей. Дальнейший мотивный анализ показал, что в целом выявленные темы и их термы интерпретируемы и позволяют по-новому взглянуть на интертекстуальные переплетения в столь несхожих текстах.

Вторая цель, которую мы преследовали данным пилотным экспериментом, – желание обосновать потенциальную продуктивность использования ИИ в рамках филологического анализа художественных текстов. Представляется, что проанализированный в качестве примера кейс в некоторой мере это желание реализовал.

ЛИТЕРАТУРА

Андреев, В. С. «Светлый» Лонгфелло: концепт Свет в меняющемся стиле / В. С. Андреев // Известия Смоленского государственного университета. – 2019. – № 3 (47). – С. 201–210.

Андреев, С. Н. Распределение триграмм в тексте (динамический аспект изучения стихотворного текста) / С. Н. Андреев // Квантитативная лингвистика. – 2016. – Т. 4. – С. 20–30.

Булатов, В. Г. Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet : дис. … канд. техн. наук / Булатов В. Г. – М., 2020. – 147 с.

Вахштайн, В. Техника / В. Вахштайн. – СПб. : Издательство Европейского университета в Санкт-Петербурге, 2021. – 156 с.

Гаспаров, Б. М. Литературные лейтмотивы. Очерки по русской литературе XX в. / Б. М. Гаспаров. – М. : Наука. Издательская фирма «Восточная литература», 1993. – 304 с.

Жолковский, А. К. Мир автора и структура текста / А. К. Жолковский, Ю. К. Щеглов. – Tenafly, 1986. – 348 с.

Мартыненко, Г. Я. Методы математической лингвистики в стилистических исследованиях / Г. Я. Мартыненко. – Санкт-Петербург : Нестор-История, 2019. – 295 с.

Матвеева, Г. В. Функциональные стили в аспекте текстовых категорий: Синхронно-сопоставительный очерк / Г. В. Матвеева. – Свердловск : Изд-во Уральского университета, 1990. – 172 с.

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

Список Сергея Довлатова. – URL: http://sergeidovlatov.com/books/urzhumova.html. – Текст : электронный.

Томашевский, Б. В. Теория литературы. Поэтика : учеб. пособие / Б. В. Томашевский ; вступ. статья Н. Д. Тамарченко ; комм. С. Н. Бройтмана при участии Н. Д. Тамарченко. – М. : Аспект Пресс, 1999. – 334 с.

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

Gibson, J. The Ecological approach to visual perception / J. Gibson. – Tailor and Francis, 1986. – 359 p.

Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure / M. Grootendorst // ArXiv. – 2022. – arXiv:2203.05794.

Jacobs, T. Topic models meet discourse analysis: a quantitative tool for a qualitative approach / T. Jacobs, R. Tschötschel // International Journal of Social Research Methodology. – 2019. – No. 22:5. – P. 469–485. – DOI: 10.1080/13645579.2019.1576317.

Jelisavčić, V. Topic models and advanced algorithms for profiling of knowledge in scientific papers / V. Jelisavčić, B. Furlan, J. Protić, C. Milutinović // Proceedings of the 35th International Convention. – 2012. – P. 1030–1035.

Koltsov, S. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research / S. Koltsov, S. Pashakhin, S. Dokuka // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in 7 Bioinformatics). 10th International Conference on Social Informatics, SocInfo 2018. – St. Petersburg ; Cham : Springer, 2018. – P. 308–321.

Nikolenko, S. I. Topic modelling for qualitative studies / S. Nikolenko, S. Koltcov, O. Koltsova // Journal of Information Science. – 2017. – No. 43 (1). – P. 88–102. – DOI: https://doi.org/10.1177/0165551515617393.

Quercia, D. Tweet LDA: supervised topic classification and link prediction in twitter / D. Quercia, H. Askham, J. Crowcroft // Proceedings of the ACM Web science conference, 2012. – New York : ACM, 2012. – P. 247–250.

Ritter, A. Unsupervised Modeling of Twitter Conversations / A. Ritter, C. Cherry, B. Dolan // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – 2010. – P. 172–180.

Schöch, Ch. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama / Ch. Schöch. – Text : electronic // Digital Humanities Quarterly. – 2017. – Vol. 11, No. 2. – URL: http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html (mode of access: 29.04.2023).

Skorinkin, D. Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta / D. Skorinkin, B. Orekhov // Digital Scholarship in the Humanities. – 2023. – DOI: https://doi.org/10.1093/llc/fqad012.

REFERENCES

Andreev, S. N. (2016). Raspredelenie trigramm v tekste (dinamicheskii aspekt izucheniya stikhotvornogo teksta) [Distribution of Trigrams in the Text (the Dynamic Aspect of the Study of the Poetic Text)]. In *Kvantitativnaya lingvistika*. Vol. 4, pp. 20–30.

Andreev, V. S. (2019). «Svetlyi» Longfello: kontsept Svet v menyayushchemsya stile [«Lighted» Longfellow: Concept Light in Changing Style]. *In Izvestiya Smolenskogo gosudarstvennogo universiteta*. No. 3 (47), pp. 201–210.

Bulatov, V. G. (2020). M*etody otsenivaniya kachestva i mnogokriterial'noi optimizatsii tematicheskikh modelei v biblioteke TopicNet [Methods for Quality Assessment and Multi-criteria Optimization of Topic Models in TopicNet Library]*. Dis ... kand. tekhn. nauk. Moscow. 147 p.

Gasparov, B. M. (1993). *Literaturnye leitmotivy. Ocherki po russkoi literature XX v.* [Literary Leitmotifs. Essays on Russian Literature of the 20th Century]. Moscow, Nauka. Izdatel'skaya firma «Vostochnaya literatura». 304 p.

Gibson, J. (1986). The Ecological Approach to Visual Perception. Tailor and Francis. 359 p.

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. In *ArXiv*. arXiv:2203.05794.

Jacobs, T., Tschötschel, R. (2019). Topic Models Meet Discourse Analysis: A Quantitative Tool for a Qualitative Approach. In *International Journal of Social Research Methodology*. No. 22:5, pp. 469–485. DOI: 10.1080/13645579.2019.1576317.

Jelisavčić, V., Furlan, B., Protić, J., Milutinović, C. (2012). Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers. In *Proceedings of the 35th International Convention*, pp.1030–1035.

Koltsov, S., Pashakhin, S., Dokuka, S. (2018). A Full-Cycle Methodology for News Topic Modeling and User Feedback Research. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in 7 Bioinformatics). 10th International Conference on Social Informatics, SocInfo 2018*. Saint Petersburg, Cham, Springer, pp. 308–321.

Martynenko, G. Ya. (2019). *Metody matematicheskoi lingvistiki v stilisticheskikh issledovaniyakh* [Methods of Mathematical Linguistics in Stylistic Research]. Saint Petersburg, Nestor-Istoriya. 295 p.

Matveeva, G. V. (1990). *Funktsional'nye stili v aspekte tekstovykh kategorii: Sinkhronno-sopostavitel'nyi ocherk [Functional Styles in the Aspect of Textual Categories: A Synchronous and Comparative Essay]*. Sverdlovsk, Izdatel'stvo Ural'skogo universiteta. 172 p.

Mitrofanova, O. A. (2019). Issledovanie strukturnoi organizatsii khudozhestvennogo proizvedeniya s pomoshch'yu tematicheskogo modelirovaniya: opyt raboty s tekstom romana «Master i Margarita» M. A. Bulgakova [The Study of the Structural Organization of a Work of Fiction through Thematic Modeling: Experience with the Text of the Novel "The Master and Margarita" by M. A. Bulgakov]. In *Korpusnaya lingvistika-2019*. Saint Petersburg, pp. 387–394.

Nikolenko, S. I., Koltcov, S., Koltsova, O. (2017). Topic Modelling for Qualitative Studies. In *Journal of Information Science*. No. 43 (1), pp. 88–102. DOI: https://doi.org/10.1177/0165551515617393.

Quercia, D., Askham, H., Crowcroft, J. (2012). Tweet LDA: Supervised Topic Classification and Link Prediction in Twitter. In *Proceedings of the ACM Web science conference, 2012*. New York, ACM, pp. 247–250.

Ritter, A., Cherry, C., Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180.

Schöch, Ch. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In *Digital Humanities Quarterly*. Vol. 11. No. 2. URL: http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html (mode of access: 29.04.2023).

Sherstinova, T. Yu., Moskvina, A. D., Kirina, M. A. et al. (2022). Tematicheskoe modelirovanie russkogo rasskaza 1900–1930: naibolee chastotnye temy i ikh dinamika [Thematic Modeling of the Russian Story 1900–1930: The Most Frequent Themes and Their Dynamics]. In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam mezhdunarodnoi konferentsii «Dialog 2022»*. Issue 21, pp. 512–526.

Skorinkin, D., Orekhov, B. (2023). Hacking Stylometry with Multiple Voices: Imaginary Writers Can Override Authorial Signal in Delta. In *Digital Scholarship in the Humanities*. DOI: https://doi.org/10.1093/llc/fqad012.

Spisok Sergeya Dovlatova [Sergei Dovlatov's List]. URL: http://sergeidovlatov.com/books/urzhumova.html.

Tomashevsky, B. V. (1999). *Teoriya literatury. Poetika* [Theory of literature. Poetics]. Moscow, Aspekt Press. 334 p.

Vakhshtayn, V. (2021). *Tekhnika* [Technics]. Saint Petersburg, Izdatel'stvo Evropeiskogo universiteta v Sankt-Peterburge. 156 p.

Zholkovsky, A. K., Shcheglov, Yu. K. (1986). *Mir avtora i struktura teksta* [Author's World and Structure of the Text]. Tenafly. 348 p.

ПРИЛОЖЕНИЕ

Тексты подкорпуса «не-Довлатов»:

1. Белые слоны. Хемингуэй. 1927.
2. Бесы. Федор Достоевский. 1872.
3. Бильярд в половине десятого. Генрих Белль. 1959.
4. Былое и думы. Александр Герцен. 1868.
5. В сторону Свана. Марсель Пруст. 1913.
6. Голубой Отель. Крейн. 1958.
7. Голый год. Борис Пильняк. 1922.
8. Гранатовый браслет. Александр Куприн. 1911.
9. Зависть. Олеша. 1927.
10. Зима тревоги нашей. Джон Стейнбек. 1961.
11. Казаки. Лев Толстой. 1863.
12. Комендант порта. Александр Грин. 1933.
13. Машина времени. Герберт Уэллс. 1895.
14. Милый друг. Ги де Мопассан. 1885.
15. Морские рассказы. Виктор Конецкий.
16. Мы. Евгений Замятин. 1924.
17. Наследник. Лев Славин. 1930.
18. Нетерпение сердца. Стефан Цвейг. 1939.
19. Обыкновенная женщина. Аркадий Аверченко. 1917.
20. Поединок. Куприн. 1905.
21. Прелести культуры. Михаил Зощенко. 1934.
22. Рыжик. Ренар. 1894.
23. Севастопольские рассказы. Лев Толстой. 1855.
24. Смерть героя. Олдингтон. 1929.
25. Старик и море. Эрнест Хемингуэй. 1952.
26. Темные аллеи. Бунин. 1938.
27. Тихий Дон. Шолохов. 1925–1940.
28. Фиеста. Хемингуэй. 1926.
29. Чудаки (рассказы). Том 1. Алексей Толстой. 1908–1911.
30. Чужая жена и муж под кроватью. Достоевский. 1848.
31. Шагреневая кожа. Оноре де Бальзак. 1831.

Данные об авторах

Колмогорова Анастасия Владимировна – доктор филологических наук, профессор департамента филологии Санкт-Петербургской школы гуманитарных наук и искусств, заместитель заведующего лабораторией языковой конвергенции, Национальный исследовательский университет «Высшая школа экономики» (Санкт-Петербург, Россия).

Адрес: 190068, Россия, Санкт-Петербург, наб. канала Грибоедова, 119–121.

E-mail: akolmogorova@hse.ru.

Залевская Екатерина Дмитриевна – стажер-исследователь лаборатории языковой конвергенции Санкт-Петербургской школы гуманитарных наук и искусств, Национальный исследовательский университет «Высшая школа экономики» (Санкт-Петербург, Россия).

Адрес: 190068, Россия, Санкт-Петербург, наб. канала Грибоедова, 119–121.

E-mail: edzalevskaya@edu.hse.ru.

Дата поступления: 02.05.2023; дата публикации: 30.06.2023

Authors' information

Kolmogorova Anastasia Vladimirovna – Doctor of Philology, Professor of Department of Philology, Saint Petersburg School of Humanities and Arts, Deputy Head of the Laboratory of Language Convergence, National Research University Higher School of Economics (Saint Petersburg, Russia).

Zalevskaya Ekaterina Dmitrievna – Research Intern of the Laboratory of Language Convergence, Saint Petersburg School of Humanities and Arts, National Research University Higher School of Economics (Saint Petersburg, Russia).

Date of receipt: 02.05.2023; date of publication: 30.06.2023