

Министерство науки и высшего образования Российской Федерации
Сибирский федеральный университет

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ

Материалы Международной научно-практической конференции

Красноярск, 25–28 сентября 2023 г.

Электронное научное издание

Красноярск
СФУ
2023

УДК 168.522:004.9(08)
ББК 71я4+32.97я4
И741

Ответственный редактор – Лаптева Марина Анатольевна

И741 Информационные технологии в гуманитарных исследованиях : материалы Междунар. науч.-практ. конф. Красноярск, 25–28 сентября 2023 г. / отв. ред. М. А. Лаптева. – Электрон. дан. (5,6 Мб). – Красноярск : Сиб. федер. ун-т, 2023. – Электрон. опт. диск (CD-ROM). – Систем. требования : PC не ниже класса Pentium I ; 128 Mb Ram ; Windows 98/XP/7 ; Adobe Reader v 8.0 и выше. – Загл. с экрана.
ISBN 978-5-7638-4900-4

В настоящем издании представлены материалы докладов Международной научно-практической конференции «Информационные технологии в гуманитарных исследованиях», которая состоялась в сентябре 2023 г. в Красноярске. Статьи посвящены актуальным вопросам развития цифровых гуманитарных наук в России и мире и размещены в шести разделах, соответствующих основным тематическим направлениям конференции: «Цифровое наследие»; «Цифровая история»; «Digital Humanities в образовании»; «Цифровые проекты»; «Компьютерный анализ текста»; «Пленарные доклады».

Предназначены специалистам, работающим в области цифровых гуманитарных исследований, а также всем, кому интересно современное развитие гуманитарного знания.

Ответственность за аутентичность и точность цитат, имён и иных сведений, а также за соблюдение законов об интеллектуальной собственности несут авторы публикуемых материалов.

УДК 168.522:004.9(08)
ББК 71я4+32.97я4

ISBN 978-5-7638-4900-4

© Сибирский федеральный университет, 2023

Электронное научное издание

Корректор Л. В. Боос
Компьютерная вёрстка Е. А. Сафиной

Подписано в свет 18.12.2023. Заказ № 20342
Тиражируется на машиночитаемых носителях
Библиотечно-издательский комплекс
Сибирского федерального университета
660041, г. Красноярск, пр. Свободный, 82а
Тел.: (391)206-26-16; <http://rio.sfu-kras.ru>
E-mail: publishing_house@sfu-kras.ru

СОДЕРЖАНИЕ

РАЗДЕЛ 1. ЦИФРОВОЕ НАСЛЕДИЕ.....	7
<i>Гайдук В. Л.</i> Можно ли сделать датасет из архивной базы: опыт Медиатеки ГМИИ им. А. С. Пушкина.....	7
<i>Кожин К. Д., Кижнер И. А.</i> Оценка вариативности описания пропусков в значениях полей метаданных учётных записей объектов коллекции Госкаталога.....	12
<i>Мазин Е. И., Полоян А. В.</i> Digital Humanities в сохранении культуры и традиций: говорящий город.....	20
<i>Гук Д. Ю., Пермяков В. А.</i> Электронная энциклопедия Эрмитажа. Итоги первой пятилетки.....	24
<i>Пиков Н. О., Дворецкая А. П., Меховский В. А., Дяченко А. С.</i> Виртуальная репрезентация сакрального пространства Енисейской Сибири.....	27
<i>Шемелина Д. С., Радченко Л. К., Декунцов Н. С.</i> Архитектурно-градостроительное наследие сибирских оборонительных линий XVIII века: поиск перспектив ревалоризации с помощью геоинформационной модели.....	32
РАЗДЕЛ 2. ЦИФРОВАЯ ИСТОРИЯ.....	40
<i>Бабайцев М. Н.</i> Обработка исторических источников инструментарием OCR, табличных процессоров и ГИС на примере списка населённых мест Тверской губернии.....	40
<i>Владимиров В. Н., Брюханова Е. А., Неженцева Н. В.</i> Информационная система по исторической статистике Алтая: особенности формирования комплекса региональных данных.....	46
<i>Дамдынчап В. М.</i> Пространственная репрезентация и анализ социодемографического пространства территории Тувы советского периода на основе ГИС-технологий.....	51
<i>Данилов Е. В.</i> Опыт применения статистических методов для анализа процессов обращаемости за медицинской помощью в России в конце XIX века.....	60
<i>Саломатина С. А.</i> Расстояния между региональными рынками Российской империи во второй половине XIX века: сетевой и статистический анализ.....	64

<i>Солощенко Н. В.</i> Образ «нового рабочего» пищевой промышленности СССР в годы первой и второй пятилеток в отражении публицистических материалов (опыт сетевого анализа и контент-анализа).....	70
<i>Стальмакова А. В.</i> Проблемы советской истории на страницах калининградской партийной прессы в период перестройки.....	74
<i>Стороженко А. А.</i> Старообрядческие миграции в 30–50-е годы XX века в Сибирь: опыт применения цифровых технологий.....	79
<i>Воронина П. С.</i> База данных «Карбеженцы: люди и судьбы (1920-е годы)»: концепция и структура.....	82
<i>Акашева А. А.</i> Наборы данных, локальные проекты, онлайн-ресурсы об изданиях Первой всеобщей переписи населения 1897 года Российской империи.....	87
РАЗДЕЛ 3. DIGITAL HUMANITIES В ОБРАЗОВАНИИ.....	89
<i>Водолажская Л. Н.</i> Трансдисциплинарный курс «Цифровые и естественно-научные методы в археологии».....	89
<i>Жила В. В., Михайлова М. А., Ермолаев Е. А., Кочнева Я. В.</i> Цифровое гуманитарное обучение как основа разработки учебных программ.....	93
<i>Ишанходжаева З. Р.</i> Цифровые технологии в высшем историческом образовании в Узбекистане.....	98
РАЗДЕЛ 4. ЦИФРОВЫЕ ПРОЕКТЫ.....	102
<i>Брюханова Е. А., Рыгалова М. В.</i> Справочные издания как источники по городскому населению Сибири конца XIX – начала XX века.....	102
<i>Matkin N. A., Klyuev N. A.</i> A city in a mirror of social media: automated analysis approach to urban representations...	107
<i>Мадрахимова Д. С.</i> Применение искусственного интеллекта для экспертной оценки успешности карьеры российских пореволюционных эмигрантов с учётом их дореволюционного положения в академическом сообществе России.....	114
<i>Косовская К. К., Тихонов Г. А., Долгозвягов Д. Д.</i> Цифровые инструменты и ресурсы, применяемые в гуманитарных науках.....	119

<i>Дяченко А. С., Кижнер И. А.</i> Подход к мультимодальному анализу периодического издания «Курьер ЮНЕСКО»: сегментация и извлечение изображений.....	124
<i>Гурьянов И. С.</i> Тенденции развития исторических ГИС.....	134
<i>Акашева А. А.</i> Информационные ресурсы Ассоциации «История и компьютер».....	137
<i>Брылевская Е. В.</i> Некоторый алгоритм для построения семантических сетей.....	139
РАЗДЕЛ 5. КОМПЬЮТЕРНЫЙ АНАЛИЗ ТЕКСТА.....	142
<i>Бутусова М. А.</i> Сетевой анализ речи героев пьес А. П. Чехова «Три сестры» и «Вишнёвый сад»....	142
<i>Ведерникова В. Д., Северина Е. М.</i> Цифровой анализ идиостиля автора и переводчика (на материале текстов произведений Дж. К. Роулинг).....	147
<i>Максименко П. И.</i> Русскоязычная электронная база фанфикшн-текстов: принципы создания и анализ метаданных.....	151
<i>Панкова М. С.</i> Корпус донских говоров как opportunistic corpus.....	160
<i>Белковский В. А., Максаков И. П.</i> Перспективы использования компьютеризированного контент-анализа как метода гуманитарного исследования в современной практике.....	164
<i>Артемяева А. И.</i> Технологии оценки доступности текстовой информации в практике создания безбарьерной коммуникационной среды.....	169
<i>Сальников Е. А., Бонч-Осмоловская А. А.</i> Применение стилометрии для определения сгенерированных текстов.....	176
<i>Долуденко Е. А., Горохова Л. А., Горохова А. В.</i> Опыт корпусного анализа дневниковых записей студентов-переводчиков.....	183
<i>Emmanuel Ngue Um</i> The erasure of Africa in the era of ChatGPT.....	189
<i>Полоян А. В., Приходько К. Д.</i> Составление промпта в нейросети Midjourney для генерации иллюстраций к творчеству А. П. Чехова.....	194

<i>Шерстинова Т. Ю., Кирина М. А., Хлусова Я. К.</i> Корпус русского рассказа как база для проведения социолингвистических исследований русской литературы.....	200
<i>Шилова М. В.</i> Визуализация исторических текстов: выявление географических связей цифровыми средствами.....	212
<i>Синицын В. Ю., Назаровская В. С.</i> Тестирование эффективности больших языковых моделей при выполнении заданий единого государственного экзамена по русскому языку.....	221
<i>Северина Е. М.</i> Проект Chekhov Digital: цифровой индекс имён и названий.....	228
<i>Баранов В. А.</i> Дистрибуция и семантика славянского слова в историческом корпусе.....	232
<i>Михайлова А. Н.</i> Атрибуция мемуаров Г. К. Жукова и К. К. Рокоссовского с помощью метода Delta...	238
РАЗДЕЛ 6. ПЛЕНАРНЫЕ ДОКЛАДЫ.....	246
<i>Бородкин Л. И.</i> Нечёткая кластеризация с элементами обучения в историко-типологических исследованиях.....	246
<i>Норманская Ю. В.</i> Возможности «ЛингвоДока» для выявления степени близости языков и диалектов и описания графико-фонетических особенностей.....	251
<i>Orekhov B. V.</i> Hard cases in cross-corpora drama annotation.....	258

РАЗДЕЛ 1. ЦИФРОВОЕ НАСЛЕДИЕ

УДК 930.25

В. Л. Гайдук,

начальник Медиатеки

Государственный музей изобразительных искусств им. А. С. Пушкина, Москва, Россия

e-mail: bastetm9u@gmail.com

ORCID: 0000-0002-1131-5732

МОЖНО ЛИ СДЕЛАТЬ ДАТАСЕТ ИЗ АРХИВНОЙ БАЗЫ: ОПЫТ МЕДИАТЕКИ ГМИИ ИМ. А. С. ПУШКИНА

Аннотация. Оцифровка архивных и музейных коллекций создала иллюзию возможного объединения материалов, раздробленных в силу исторических обстоятельств. Большинство российских консолидированных архивных коллекций представляет собой банк цифровых копий архивных документов, при этом даже внутри одной базы разнятся качество и формат изображений, отсутствует стандартизация содержания полей описания. В докладе будут проанализированы основные проблемы, связанные с превращением архивов из скопления разнородных данных в действенный инструмент работы с ними.

Ключевые слова: *датасет, цифровизация, КАМИС, стандартизация данных, медиатека*

Оцифровка архивных и музейных коллекций создала иллюзию возможного объединения материалов, раздробленных в силу исторических обстоятельств. Большинство примеров современных российских консолидированных архивов представляет собой банк цифровых копий документов [1; 4; 12], при этом даже внутри одной базы разнятся качество и формат изображений, отсутствует стандартизация содержания полей описания. Возможной причиной такого поверхностного понимания оцифровки стало представление о базе данных как о конечном продукте цифровизации.

В начале 2000-х гг. в ходе дискуссии о роли библиотек и архивов в историческом процессе на страницах «Нового литературного обозрения» [9] рассматривался вопрос о демократизации архивов посредством использования цифровых технологий. Главной проблемой, которую, по мнению исследователей, могли решить новые технологии, была организация открытого дистанционного доступа к архивным материалам. Сценарий публикации материалов виделся таким образом: «поэтапное размещение в свободном доступе в Сети справочников сначала на уровне фондов, затем на уровне описей и только потом на уровне дел» [3].

По прошествии 20 лет практически каждый федеральный, государственный и муниципальный архив в рамках проекта цифровизации разместил на сайте свой научно-справочный аппарат, а именно: описи, путеводители и каталоги. Следующим закономерным этапом стала консолидация всех архивных материалов в рамках единого общегосударственного каталога – Центрального фондового каталога архивного фонда РФ (ЦФК).

Главной задачей ЦФК является «предоставление пользователям информации об архивных *фондах* (выделено – В. Г.), хранящихся во всех федеральных архивах, государственных и муниципальных архивах субъектов РФ» [6]. Исходя из информации, опубликованной на сайте, мельчайшей единицей описания ЦФК является архивный фонд, хотя система осуществляет поиск по единицам хранения, полное архивное описание единиц хранения в системе не представлено, равно как и цифровые образы документов.

Неоднократно в исследованиях, инициированных ВНИИДАДом и Росархивом, звучало, что одной из главных задач при использовании информационных технологий в архивной сфере является «Необходимость выработки единых, унифицированных для всей архивной сферы подходов, регулирующих перенос информации из учётной документации архивов в автоматизированные архивные комплексы» [2; 14]. Для этих целей был разработан единый порядок внесения информации в Центральный фондовый каталог, в которых подробно описаны различные способы унификации информации, представленной в учётных документах архива [5]. Порядок включает схемы описания различных типов документов: письменных, аудиовизуальных и электронных, но, к сожалению, в нём отсутствует принцип жёсткой стандартизации данных, что затрудняет использование данных из ЦФК в обобщённом виде для дальнейших исследований.

Пример ЦФК поставил под вопрос возможность разработки архивной базы данных, главная цель которой не сводилась бы к простой презентации, визуализации и популяризации материалов. Безусловно, приведение к единому знаменателю большого массива разнородных и разнотипных документов является трудоемкой и практически невыполнимой задачей, поэтому возникает стремление попробовать создать на существенно меньшем и более однородном объеме материалов базу данных, которая бы стала не финальной, а начальной точкой исследований. Именно такая задача стоит перед командой Медиатеки ГМИИ им. А. С. Пушкина.

Медиатека ГМИИ представляет собой комплексное хранение документов по истории русского искусства второй половины XX – начала XXI в., которое включает книжный фонд, фонд документальных материалов и фонд материалов на цифровых носителях. Синтетический характер фондов Медиатеки обусловлен спецификой передаваемых документов: современные художники работают в разных техниках, с различными медиумами, хранение которых не вписывается в привычные рамки архива и/или библиотеки.

В настоящее время архив Медиатеки представлен несколькими документальными собраниями: среди которых коллекция документов художников второй половины XX – начала XXI в., коллекция афиш и плакатов второй половины XX – начала XXI в., коллекция газетных вырезок по истории искусства конца XX – начала XXI в., коллекция приглашений и буклетов российских и зарубежных институций современного искусства. Данные о материалах Медиатеки равно как и сами документы, а также их цифровые копии, вносятся в базу музейного учета КАМИС.

Изначально главной целью Медиатеки была фиксация художественного процесса, что включало сбор, систематизацию и хранение, в первую очередь, информации об акторах и событиях художественной жизни. Большая часть документов Медиатеки не является уникальной по формальным критериям, определённым Росархивом¹, но содержат ценную, хотя и однотипную информацию о выставках, показах, встречах, художниках, кураторах и пр.

Афиши, приглашительные билеты, пресс-релизы на основе признаков, выделенных Б. Г. Литваком, подпадают под определение массовых источников². Смысловым ядром такого рода источников становится информация о событии, которую можно разложить на несколько составных частей: название события, тип события, место проведения события, дата проведения события, участники события, институция-организатор. Эти пять составных элементов представляют формуляр, типичный для афиш, пресс-релизов и приглашений. Именно формуляр карточки «событие» было решено положить в основу разрабатываемого сотрудниками Медиатеки датасета³ по истории русского искусства 1990–2000-х гг.

В настоящее время существуют датасеты по русской литературе, литературоведению и фольклору [см., например, 8; 11]. Портал открытых данных Министерства культуры предоставляет датасеты по различной тематике [10], в т. ч., например, по событиям в сфере культуры, но, во-первых, датасет включает только мероприятия, которые были опубликованы на портале *PRO.культура.рф*, во-вторых, самые ранние события относятся к 2014 г. В рамках проекта мы предлагаем сделать открытый набор данных, который охватит в т. ч. и более ранний период истории русского искусства – 1990-е гг.

При внесении карточек архивных документов в базу музейного учёта КАМИС, если карточка аффилирована с событием (выставкой, дискуссией, концертом и пр.), в ней создаётся или выбирается из списка соответствующее событие. Справочник по событиям можно выгрузить отдельно. Справочник содержит следующие характеристики: название, тип события,

¹ Согласно п. 6 ст. 3 ФЗ «Об архивном деле в РФ»: «уникальный документ – особо ценный документ, не имеющий себе подобных по содержащейся в нём информации и (или) его внешним признакам, невозможный при утрате с точки зрения его значения и (или) автографичности» [13].

² Борис Григорьевич Литвак выделял следующие признаки массовых источников: однородность обстоятельств происхождения; однородность, аналогичность или повторяемость содержания; однотипность формы, тяготеющая к стандартизации. [7, с. 7].

³ Под датасетом в данном случае понимается обработанный и структурированный набор данных.

место проведения, дата проведения, персоналии (участники события), организация. В процессе работы с базой изначально не было разработано стандарта заполнения справочника по событиям, это привело, во-первых, к дублированию записей, во-вторых, к разному формату данных, которые представлены в справочнике и, в-третьих, к разной степени полноты этих данных.

В процессе нормализации и стандартизации данных по событиям художественной жизни 1990–2000-х гг. стало очевидно, что дальнейшее внесение информации в справочники возможно только при условии разработки единого стандарта заполнения полей. Порядок, который используется для заполнения ЦФК оказался нерелевантным в данном случае, т. к. не предполагал последующего использования данных, в нашем же случае мы планируем получить датасет, пригодный для дальнейшего научного использования.

Из шести полей формуляра события, лишь одно поле является уникальным – «название», остальные поля являются справочными, т. к. предполагают выбор из предлагаемых вариантов, если необходимого варианта нет, то этот вариант добавляется в справочник. Наиболее уязвимым атрибутом события оказалось поле «дата», т. к. согласно архивным правилам при указании даты могут быть использованы буквы, цифры и другие знаки, например, квадратные скобки для указания примерной датировки (допустимы значения: «май 1935 г.», «15 июня 1984 г.», «1870-е», «[1876]» и т. д.). К сожалению, в датасетах возможно только указание точной даты в числовом формате, поэтому для уточнения датировки используются внешние ресурсы, в которых зафиксирована точная дата. При невозможности указания точной даты поле остаётся незаполненным.

Таким образом, на примере небольшого объёма данных из справочника «Событие» становится очевидным, что любой работе по заполнению базы данных должна предшествовать разработка единых стандартов заполнения полей, главной целью которой должно быть представление о дальнейших возможностях использования данных. Только в этом случае база не будет мыслиться как «кладбище» данных, а станет удобным инструментом для работы с ними.

Список источников

1. Автограф. XX в.: электронный архив русской литературы. URL: literature-archive.ru/ru.
2. Изучение типологии и эволюции исторических форм архивной учётной документации и возможности их представления в электронной среде: аналит. обзор. М., 2014.
3. Боброва Е. В. Архивы via Интернет / Е. В. Боброва // Новое литературное обозрение. 2005. № 74. URL: magazines.russ.ru/nlo/2005/74/bob31.html.
4. Документы советской эпохи. URL: sovdoc.rusarchives.ru.
5. Единый порядок заполнения полей единой автоматизированной информационной системы, состоящей из программы комплексов «Архивный фонд», «Фондовый каталог», «Центральный фондовый каталог». URL: archives.gov.ru/sites/default/files/poryadok-zapoln-poley-af-fc-cfc.pdf.

6. Информация о Центральном фондовом каталоге. URL: cfc.rusarchives.ru/CFC-search/Search/About.
7. Литвак Б. Г. Очерки источниковедения массовой документации XIX – начала XX в. / Б. Г. Литвак. М., 1979.
8. Набор данных «Назирова». URL: github.com/nevmenandr/nazirov-texts-dataset.
9. Новое литературное обозрение. 2005. № 74 // Журнальный зал. URL: magazines.russ.ru/nlo/2005/74.
10. Открытые данные Министерства культуры РФ. URL: opendata.mkrf.ru/opendata.
11. Репозиторий открытых данных по русской литературе и фольклору. URL: dataverse.pushdom.ru; kaggle.com/datasets/d0rj3228/russian-literature.
12. Сеть архивов российского искусства (RAAN). URL: russianartarchive.net/ru.
13. Об архивном деле в РФ: Федеральный закон № 125-ФЗ от 22.10.2004. URL: pravo.gov.ru/proxy/ips/?docbody=&link_id=17&nd=102089077.
14. Юмашева Ю. Ю. Центральный фондовый каталог: идеология системы и анализ результатов первого года функционирования / Ю. Ю. Юмашева // Вестник архивиста. 2013. № 2. URL: vestarchive.ru/2013-2/2607-centralnyi-fondovyi-katalog-ideologiya-sistemy-i-analiz-rezultatov-pervogo-goda-funkcionirovaniia.html.

К. Д. Кожин¹, И. А. Кижнер²,

¹ лаборант-исследователь лаборатории искусственного интеллекта СФУ;

² кандидат культурологии, доцент, старший научный сотрудник лаборатории
Digital Humanities

^{1,2} Сибирский федеральный университет, Красноярск, Россия

e-mail: ¹ kkozhin@sfu-kras.ru, ² ikizhner@sfu-kras.ru

ОЦЕНКА ВАРИАТИВНОСТИ ОПИСАНИЯ ПРОПУСКОВ В ЗНАЧЕНИЯХ ПОЛЕЙ МЕТАДАННЫХ УЧЁТНЫХ ЗАПИСЕЙ ОБЪЕКТОВ КОЛЛЕКЦИИ ГОСКАТАЛОГА

Аннотация. Цифровые коллекции культурных и исторических объектов являются основным источником данных для цифровых гуманитарных исследований. Анализ цифровых коллекций становится возможен благодаря анализу метаданных, характеризующих цифровые объекты. В этой работе мы исследуем пропуски в значениях полей метаданных учётных записей объектов коллекции Государственного каталога Музейного фонда РФ и утверждаем, что понятие пропуска значения в поле метаданных является вариативным значением, которое, как и другие варианты значений полей метаданных, обусловлено социальными и культурными причинами. Мы показываем, что распознавание вариативности заполнения поля дает возможность лучше понять степень качества метаданных и эксплицировать пропуски в данных и/или недостаточную репрезентативность данных.

Ключевые слова: пропуски в данных, стандартизация, цифровые коллекции культурного наследия, музейные коллекции, Государственный каталог Музейного фонда РФ

Введение

Цифровые коллекции культурных и исторических объектов являются основным источником данных для цифровых гуманитарных исследований. Анализ цифровых коллекций становится возможен благодаря анализу метаданных, характеризующих цифровые объекты. Проблемы, связанные с качеством метаданных и совместимостью метаданных как в рамках одной коллекции, так и в случае агрегатора объектов цифрового наследия, осложняют анализ коллекций [1–4]. Одной из серьёзных проблем, связанных с качеством метаданных, является проблема пропусков в метаданных [4–7]. Для некоторых полей метаданных, таких как место публикации или жанр произведения, пропуски могут составлять более 70 % всех значений для этого поля [8; 9].

В случае когда значение поля метаданных неизвестно, например, неизвестен год создания произведения, сотрудник музея, библиотеки или архива часто использует спектр вариантов, выбирая тот вариант, который принят в его окружении. В этой работе мы исследуем пропуски в значениях полей метаданных учётных записей объектов коллекции Государст-

венного каталога Музейного фонда РФ⁴ и утверждаем, что понятие «пропуска значения в поле метаданных» является вариативным значением, которое, как и другие варианты значений полей метаданных, обусловлено социальными и культурными причинами. Мы показываем, что распознавание вариативности заполнения поля дает возможность лучше понять степень качества метаданных и эксплицировать пропуски в данных и/или недостаточную репрезентативность данных.

1. Источник данных

В качестве источника данных был использован набор метаданных учётных записей объектов Государственного каталога Музейного фонда РФ. Госкаталог является крупнейшим агрегатором культурного наследия России. Количество учётных записей объектов на июнь 2023 г. составляет около 37 млн, это больше половины предметов основного Музейного фонда, включающего в себя около 70 млн культурных предметов⁵. Каждая учётная запись объекта отнесена к одной из 15 категорий (фотографии, документы, скульптура, нумизматика, редкие книги, предметы быта и этнографии и т. д.). Каждая учётная запись объекта содержит значения более 20 полей метаданных, таких как название, автор, место создания, описание, период создания, размер, музей, техника изготовления и т. д.

Таким образом, набор учётных записей объектов Государственного каталога Музейного фонда РФ подходит для нашего исследования, т. к. он является большим массивом разнообразных культурных данных. Также исследователям предоставляется возможность, хотя и с ограничениями, получить доступ к данным с помощью интерфейса прикладного программирования (*API*), что является более удобным и быстрым способом сбора информации.

2. Методы исследования

2.1. Метод получения данных

Метаданные учётных записей объектов Госкаталога были получены с помощью программы *SGAT 2.0* [10], которая позволяет быстро собрать данные в *csv*-файл, используя *API* сайта Открытых данных Минкультуры РФ⁶. Так, был получен файл, включающий метаданные 34 187 161 учётной записи объекта. Нам удалось собрать более 90 % данных коллекции Госкаталога. Данные Госкаталога постоянно пополняются, также существуют ограничения *API* Открытых данных Минкультуры России (например, за один запрос можно получить метаданные не более чем 100 тыс. учётных записей). Однако массив данных, составляющий более 90 % от общего числа учётных записей Госкаталога, является большим датасетом, который хорошо продемонстрирует проблему пропуска данных. Окончательный *csv*-файл включает метаданные для 34 187 161 учётной записи объекта коллекции Госкаталога.

⁴ URL: goskatalog.ru/portal.

⁵ URL: opendata.mkrf.ru/opendata/7705851331-stat_museum_svod.

⁶ URL: opendata.mkrf.ru.

2.2. Метод определения пропусков в значениях полей метаданных

Обычно под пропуском в данных подразумевают пустое значение, отсутствие информации. Однако некоторые специалисты для описания пропусков используют прочерк, который в цифровом виде имеет четыре варианта написания (тире, короткое тире, дефис и минус).

Так, при подсчёте пропусков в значениях полей метаданных будут учитываться не только пустые значения, но также и варианты записи прочерка, такие как тире, короткое тире, дефис и минус.

2.3. Метод поиска вариантов описаний пропусков в значениях полей метаданных

Учётные записи объектов Госкаталога включают примеры, где значением поля метаданных является информация об отсутствии значения поля метаданных (например, «нет автора», «неизвестно»). Поскольку подобные значения не имеют стандартов, мы осуществляли поиск, используя слово «неизвестно», а также всевозможные формы этого слова, синонимы и варианты с опечатками.

Таким образом, из-за большой вариативности и отсутствия стандартов значений полей метаданных на момент написания этой статьи поиск вариантов описаний пропусков в значениях полей метаданных возможен только полуавтоматическим способом.

2.4. Метод определения пропусков в значениях поля «описание»

В поле «описание» обычно содержится расширенная информация об учётной записи объекта. Но также среди учётных записей объектов встречаются такие, где значение поля «описание» и «название» совпадают. Поэтому для определения пропусков в значениях поля «описание» необходимо использовать значения поля «название». Поле «название» является обязательным для заполнения⁷, т. к. с помощью него учётная запись индексируется в поисковой системе сайта. Зная это, мы можем считать, что в учётной записи объекта, с полностью одинаковыми значениями полей «название» и «описание», отсутствует значение поля «описание».

3. Результаты исследования

3.1. Пропуски в значениях полей метаданных

С помощью собранных данных мы определили процент пропусков в значениях четырёх основных полей метаданных учётных записей объектов Госкаталога. В результате была сформирована гистограмма с накоплением, показывающая процент отсутствующих данных по каждому из четырёх полей метаданных (рис. 1).

⁷ URL: consultant.ru/document/cons_doc_LAW_310837/a13afd91b69fc676f206c828817225b03bd832b2.

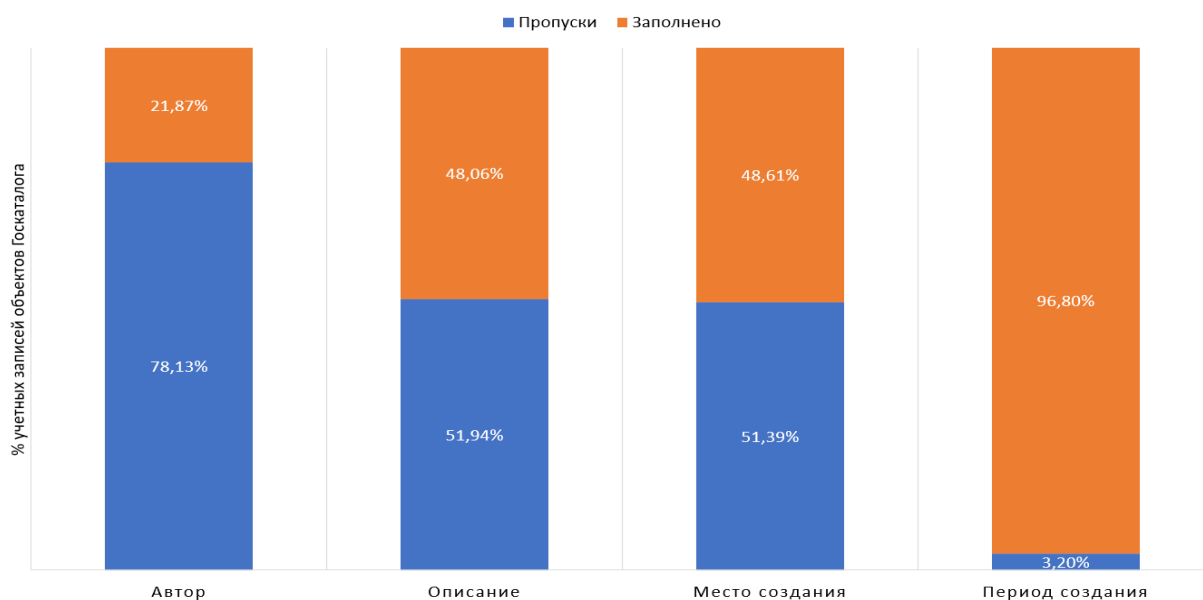


Рис. 1. Проценты учётных записей объектов Госкаталога с пропусками в значениях (тире, короткое тире, дефис, минус и пустое значение) полей метаданных (автор, описание, место создания, период создания). Синий цвет показывает процент учётных записей с такими значениями в полях метаданных, как тире, короткое тире, дефис, минус и пустое значение. Оранжевый цвет показывает процент с остальными значениями в полях метаданных

На диаграмме видно, что описание, а также информация об авторе и месте создания не указана для большей части учётных записей объектов Госкаталога.

3.2. Пропуски в значениях полей метаданных по годам

Важно показать, как ситуация с пропусками в значениях четырёх основных полей метаданных учётных записей объектов Госкаталога менялась на протяжении всего времени существования коллекции. Это позволит отследить динамику изменения пропусков и оценить, усугубилась ли проблема пропусков со временем. Используя информацию о дате регистрации каждого объекта, мы построили линейный график, показывающий пропуски в значениях полей метаданных учётных записей с 2012 по 2022 гг. (рис. 2). График не учитывает данные за 2011 и 2023 гг., т. к. коллекция начала заполняться не с начала года, а текущий год ещё не закончился.

График показывает, что по мере пополнения коллекции Госкаталога количество учётных записей с пропусками в значениях полей метаданных, таких как «автор», «описание» и «место создания», росло. Ситуация же с полем «период создания» с годами улучшилась.

Также важно показать, как менялось процентное соотношение пропусков в значениях полей метаданных по годам. Такой подход помогает лучше оценить масштаб проблемы, учитывая увеличение числа записей с каждым годом. Процентное соотношение позволяет точно отследить динамику изменений и сделать более объективные выводы. Из графика (рис. 3) следует, что пропуски в значении поля «автор» на протяжении всего времени остаются примерно на одном и том же уровне. Пропуски в зна-

чениях поля «период создания» после 2016 г. стали резко уменьшаться, что, вероятно, связано с новыми правилами заполнения метаданных, в которых данное поле стало обязательным для заполнения⁸.

Таким образом, мы получили графики, показывающие высокий процент пропусков в значениях основных полей метаданных учётных записей объектов Госкаталога. Но учитывают ли эти результаты все пропуски?

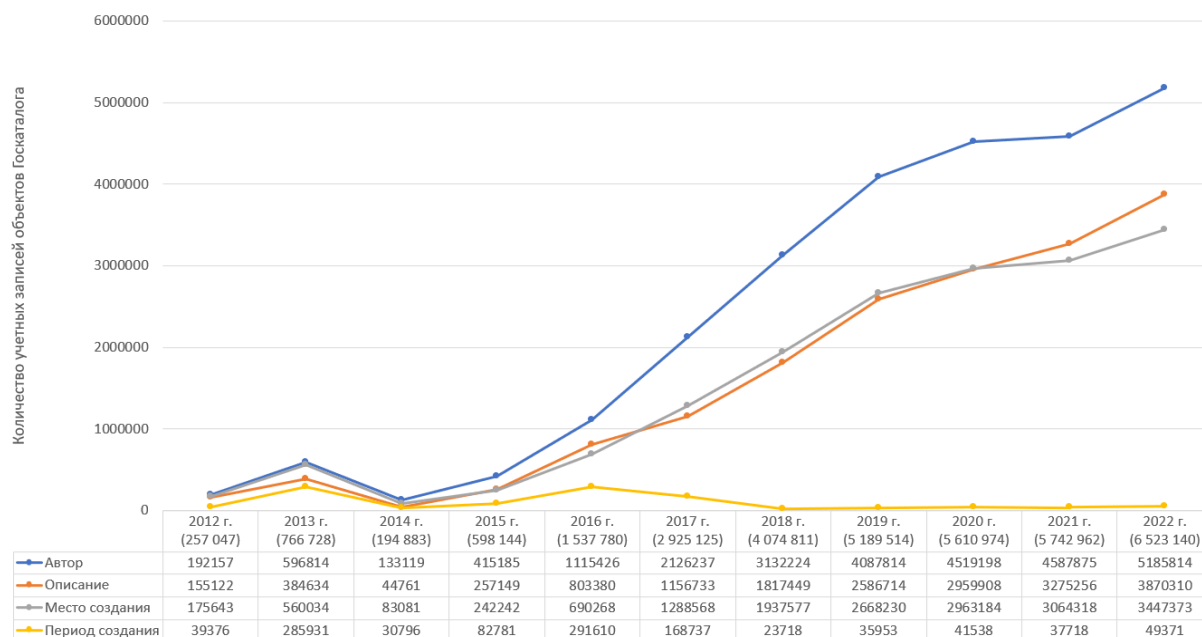


Рис. 2. Количество учётных записей объектов Госкаталога с пропусками в значениях (тире, короткое тире, дефис, минус и пустое значение) полей метаданных (автор, описание, место создания, период создания) по годам (2012–2022 гг.)

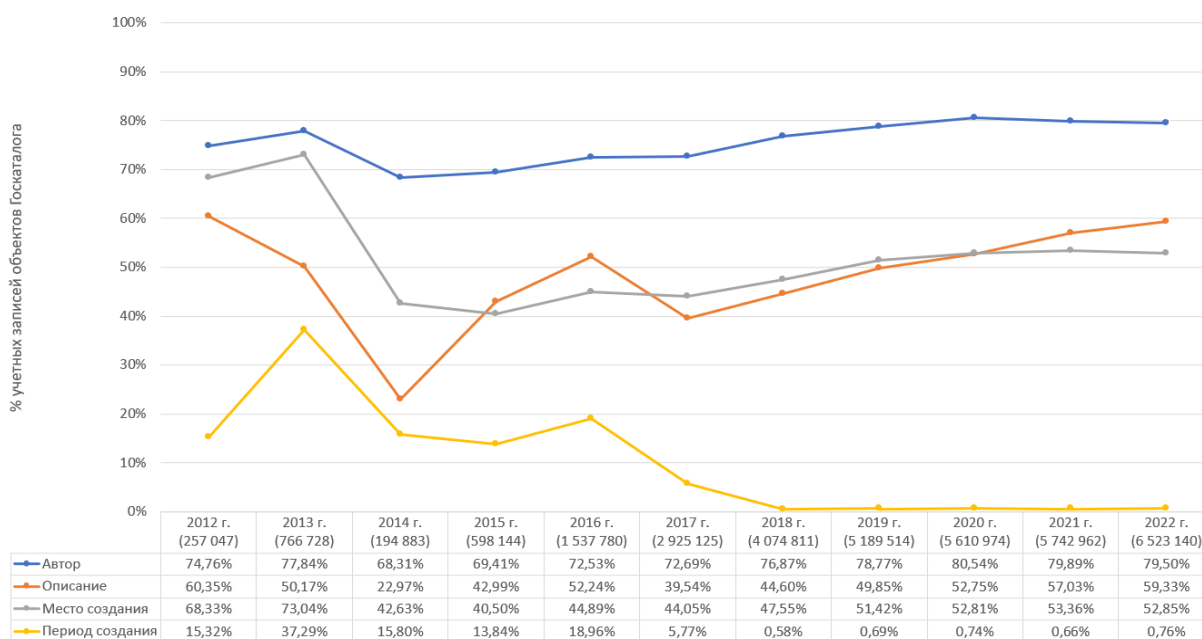


Рис. 3. Процентное соотношение учётных записей объектов Госкаталога с пропусками в значениях (тире, короткое тире, дефис, минус и пустое значение) полей метаданных (автор, описание, место создания, период создания) по годам (2012–2022 гг.)

⁸ URL: consultant.ru/document/cons_doc_LAW_310837/a13afd91b69fc676f206c828817225b03bd832b2.

3.3. Вариативность описания пропусков в значениях полей метаданных

Для получения полноценного представления о пропусках в значениях полей метаданных необходимо учесть все варианты описаний отсутствия данных и определить, в какой мере они влияют на общий результат.

Среди значений поля «автор» нам удалось найти 497 вариантов описания отсутствия данных. Помимо простых вариантов (например, «неизвестно», «нет данных»), иногда указывают род деятельности автора или предприятия (например, «Неизвестный гравёр», «Неизвестная самоварная фабрика»), страну происхождения или национальность автора (например, «Неизвестный мастер. Китай», «Неизвестный итальянский гравёр»), период деятельности автора (например, «Неизвестный художник 1-й половины XX в.»), пол автора (например, «Неизвестная мастерица»). Также встречаются опечатки и ошибки в написании слов (например, «неизвесен», «Автор не известен»). Таким образом, было найдено ещё 2 % учётных записей объектов Госкаталога, имеющих пропуск в значении поля метаданных.

Среди значений поля «место создания» было обнаружено 119 вариантов описания отсутствия данных. Среди вариантов можно заметить как простую информацию об отсутствии данных (например, «Место создания неизвестно»), так и сокращения (например, «б/м.», «Без вых. дан.»). Также встречаются опечатки и ошибки в написании слов (например, «неизвесто»). Таким образом, было найдено ещё 0,65 % учётных записей объектов Госкаталога, имеющих пропуск в значении поля метаданных.

Среди значений поля «период создания» мы нашли 163 варианта описания отсутствия данных. Среди вариантов можно заметить как простую информацию об отсутствии данных (например, «Дата не установлена», «Не известе»), так и сокращения (например, «б.г.», «Без г.с»). Также встречаются опечатки и ошибки в написании слов (например, «не узвестно»). Таким образом, было найдено ещё 1,6 % учётных записей объектов Госкаталога, имеющих пропуск в значении поля метаданных.

Также пользуясь методом определения пропусков в значениях поля «описание» (см. п. 2.4) было найдено ещё 1,1 % учётных записей объектов Госкаталога, имеющих пропуск в значении поля метаданных.

Учитывая полученные данные о количестве вариантов описаний пропусков в значениях полей метаданных, мы объединили все результаты обнаружения пропусков в значениях полей метаданных учётных записей объектов Госкаталога в одну таблицу (табл. 1).

Табл. 1 показывает, что вариативность описаний пропусков может значительно повлиять на результат поиска пропусков в значениях метаданных, ведь в нашем наборе данных 1 % соответствует 340 тыс. учётных записей объектов.

Пропуски в значениях полей метаданных
учётных записей объектов Госкаталога

Название поля метаданных	% пропусков в значениях полей метаданных (тире, короткое тире, дефис, минус и пустое значение)	% пропусков в значениях полей метаданных (описания отсутствия данных)	Количество найденных вариантов описаний пропусков в значениях полей метаданных	Общий % пропусков в значениях полей метаданных
Автор	78,13 %	2 %	497	80,13 %
Описание	51,94 %	1,1 %	–	53,04 %
Место создания	51,39 %	0,65 %	119	52,04 %
Период создания	3,2 %	1,6 %	163	4,8 %

Заключение

Результаты исследования свидетельствуют о сложностях, связанных с совместимостью значений полей метаданных не только в случае, где имеется несколько различных наименований для одного населённого пункта или автора (например, «Москва», «город Москва»), но и в случае где есть несколько различных описаний отсутствия данных, таких как «неизвестно» или «нет данных». Без возможности подсчета всех вариантов описаний пропусков в значениях полей метаданных нельзя точно утверждать, что данные являются полными и надёжными.

Отсутствие стандартизации затрудняет поиск вариантов описаний пропусков в значениях полей метаданных, поэтому, чтобы не допустить ошибок при подсчёте, необходимо вручную отбирать значения. Мы опубликовали найденные варианты описаний пропусков на *GitHub*⁹ и планируем в дальнейшем пополнять созданный словарь.

В следующих исследованиях мы планируем оценить проблему вариативности значений пропусков в метаданных объектов крупных онлайн-коллекций культурного наследия, таких как Европеана или Музей Метрополитен в Нью-Йорке.

Список источников

1. Глазунов Е. В. Унификация данных музейного Госкаталога РФ / Е. В. Глазунов и др. // Сибирский антропологический журнал. 2020. Т. 4. № 3. С. 154–168.
2. Schreiber G. Semantic Annotation and Search of Cultural-heritage Collections: the MultimediaN E-Culture Demonstrator / G. Schreiber et al. // Journal of Web Semantics 6.4. 2008. Pp. 243–249.
3. Hyvönen E. Semantic Portals for Cultural Heritage. Handbook on Ontologies / E. Hyvönen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. Pp. 757–778.
4. Lorenzini. On Assessing Metadata Completeness in Digital Cultural Heritage Repositories / Lorenzini et al. // Digital Scholarship in the Humanities. 2021. Vol. 36. Pp. ii182–ii188.

⁹ URL: github.com/konstantinkozhin/missing_values.

5. Ochoa X. Automatic Evaluation of Metadata Quality in Digital Repositories / X. Ochoa, E. Duval // International Journal on Digital Libraries. 2009. No. 10 (2–3). Pp. 67–91.
6. Margaritopoulos T. A Fine-grained Metric System for the Completeness of Metadata / T. Margaritopoulos, M. Margaritopoulos, I. Mavridis et al. // Research Conference on Metadata and Semantic Research. Berlin, Heidelberg: Springer, 2009. Pp. 83–94.
7. Margaritopoulos M. Quantifying and Measuring Metadata Completeness / M. Margaritopoulos, T. Margaritopoulos, I. Mavridis et al. // Journal of the American Society for Information Science and Technology. 2012. No. 63 (4). Pp. 724–737.
8. Havens L. The Library Catalogue as Dataset: Exploring Data Science Approaches to Analyse Collections at Scale / L. Havens et al. 2022.
9. Shang W. Exploring Metadata Quality Issues in Non-English Corpora: Preliminary Assessments of HathiTrust Records of Late Imperial Chinese Books / W. Shang, J. Jett, J. S. Downie.
10. Кожин К. Д. Программа для сбора и анализа открытых данных культурного наследия (SGAT 2.0) / К. Д. Кожин; св-во о гос. рег-ции в Реестре программ для ЭВМ № 2023666033 от 04.08.2023.

Е. И. Мазин¹, А. В. Полоян²,

² кандидат филологических наук

^{1,2} Южный федеральный университет, Ростов-на-Дону, Россия

e-mail: ¹ emazin@sfedu.ru, ² avpoloyan@sfedu.ru

ORCID: ¹ 0009-0000-1572-2813, ² 0000-0002-5620-1624

DIGITAL HUMANITIES В СОХРАНЕНИИ КУЛЬТУРЫ И ТРАДИЦИЙ: ГОВОРЯЩИЙ ГОРОД

Аннотация. Ежегодно в России делается упор на развитие какой-либо ключевой темы для привлечения к ней общественного внимания. Например, 2021 г. получил название Года науки и технологий, 2022 – был объявлен Годом народного искусства и культурного наследия, а 2023 – Годом педагога и наставника. Данная статья объединяет все три темы, поскольку посвящена технологии создания аудиогuida, имеющей неоценимый при грамотном использовании в педагогической деятельности потенциал для сохранения культурного наследия. В статье предложены и обоснованы различные способы внедрения ресурса *izi.TRAVEL* в образовательный процесс, в т. ч. с таким перспективным направлением *Digital Humanities*, как краудсорсинг.

Ключевые слова: *сохранение культуры, краудсорсинг, информационные технологии, аудиогид, обучение иностранному языку*

В России повсеместно создаются центры, целью которых является сохранение наследия в виде визуализации и реконструкции. В различных направлениях *Digital Humanities*, практикующих применение информационных технологий в рамках гуманитарных дисциплин, активно принимают участие университеты и университетские объединения: Сибирский федеральный университет (Красноярск) в сотрудничестве с Государственным Русским музеем с оцифровкой музейных коллекций [2, с. 47], НИУ ВШЭ (Москва) с проектом «Семантическая публикация 90-томного собрания сочинений Л. Н. Толстого» [2, с. 46], а также с проектом *Chekhov Digital* (URL: chekhov-digital.sfedu.ru) в рамках работы «Зеркальной лаборатории» совместно с Южным научным центром Российской академии наук и Южным федеральным университетом (Ростов-на-Дону).

Проблема сохранения и приумножения культуры и традиций остро стоит как в России, так и в других странах. Одним из ярких примеров применения *Digital Humanities* в создании фольклорного архива стал краудсорсинговый проект в Латвии, где для расшифровки оцифрованных текстов были приглашены волонтеры [3]. Возрастающее количество подобных проектов свидетельствует о необходимости оцифровки культурного наследия не только по причине его сохранности, но и в целях популяризации культуры среди молодого населения.

Времена меняются, вкусы и предпочтения молодёжи поддаются влиянию новых технологий. Книги всё реже открываются молодыми

людьми, всё чаще в их руках смартфон. И в музеи они приходят значительно реже, в основном с организованными образовательной организацией экскурсиями. Подросткам легче исследовать мир через мобильные устройства, поэтому современному педагогу и наставнику не стоит препятствовать такому формату познания и потребления информации, а постараться использовать этот формат в своих педагогических целях.

Поскольку подростки практически повсюду берут с собой телефон, использование аудиогuida для озвучивания культурных объектов города на пути следования подростка представляется нам наиболее целесообразным. Достаточно полнофункциональным сервисом в этом плане является платформа *izi.TRAVEL*, которая позволяет создавать аудиотуры по городу, виртуальные музеи, а также составлять городские квесты, доступные как на экране ПК, так и на мобильных устройствах [1]. Принцип работы аудиогuida достаточно прост: в приложении или в браузере по ссылке для выбранного маршрута срабатывает аудиодорожка озвученного поблизости объекта. Элемент игры, включённый в квесты, поддерживает интерес пешехода. Следует отметить, что желание пройтись по маршруту и послушать интересные записи о встречаемых на пути объектах могут служить ещё одним стимулом молодежи выйти на свежий воздух, т. е. способствует оздоровлению молодого поколения.

Примечательно то, что на указанной платформе в большом количестве (более 40) присутствуют аудиогиды по г. Красноярску (рис. 1). При этом использованы все возможности платформы: помимо туров по городу присутствуют аудиогиды по музеям (Красноярский художественный музей им. В. И. Сурикова, Красноярский краевой краеведческий музей, Музей художника Бориса Рязова и др.), а также Красноярск-квест «По следам Беринга».

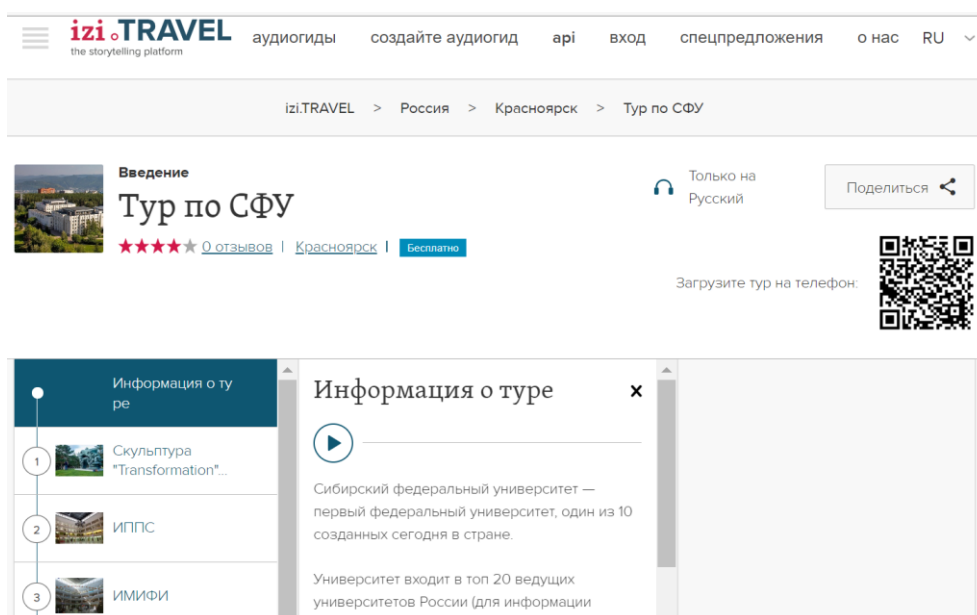


Рис. 1. Аудиогид на *izi.TRAVEL* «Тур по СФУ»

Потенциал такого ресурса в образовательном процессе огромен не только для сохранения и воспроизведения информации о культурных объектах, но и для обучения иностранному языку. Преподаватель в качестве заданий по развитию устной и письменной речи может подобрать студентам аудиогиды зарубежных музеев на иностранном языке. Студенты, переводя услышанное на родной язык с опорой на текст или без таковой, знакомятся с культурой страны изучаемого языка.

Платформа позволяет добавлять аудиофайлы, текст и фотографии, что даёт широкое поле деятельности для проектной работы. Преподаватель может использовать в работе с данным ресурсом поисковый метод, дав студентам задание найти информацию о заданных объектах самостоятельно, составить тексты как на родном языке, так и перевести их на иностранный язык, озвучить тексты. Платформа *izi.TRAVEL* имеет интуитивно понятный интерфейс с интегрированной картой *Google*, т. е. для создания аудиогидов пользователю предлагается карта мира с уже имеющимися на ней объектами (с возможностью отмечать новые).

Ресурс позволяет создавать мультязычные слои для одного и того же объекта, что расширяет возможности использования аудиогидов в смешанной языковой группе и служит определённым стимулом для привлечения иностранных туристов.

В августе 2022 г. группой студентов Южного федерального университета в рамках проекта «Говорящий город» был создан аудиогид для иногородних студентов с целью помочь абитуриентам познакомиться с городом, в котором им вскоре предстоит учиться (рис. 2). В дальнейшем аудиогид планируется дополнить другими иностранными языками в рамках выполнения проектной работы на занятиях по иностранному языку. Таким образом, аудиогид в дальнейшем будет полезен как иногородним, так и иностранным студентам.



Рис. 2. Пример аудиогидов по Ростову-на-Дону для иногородних студентов

Аудиогид для иногородних студентов ЮФУ был представлен на внутриуниверситетской студенческой конференции осенью 2022 г. Заинтересованные студенты предложили и другие идеи для создания аудио-

гида – например, аудиогид по студенческому городку для вновь прибывших жильцов. Также прозвучала идея о создании путеводителя по рекомендуемым магазинам города с советами от студентов старших курсов, как первокурснику экономить и тратить средства грамотно. Но если выходить за рамки одного института, идею создания аудиогидов можно развить до межфакультетского взаимодействия. Например, аудиогид по культурным объектам города может сплотить команды института иностранных языков с командой студентов исторического факультета для совместной разработки проекта. Особенно такой союз будет полезен для создания маршрута по местам боевой славы в рамках патриотического воспитания молодёжи.

Также взаимодействие студентов на описываемом ресурсе возможно расширить не только внутри страны, но и за её пределами. Совместная работа со студентами иностранных университетов, где изучается русский язык, позволит создавать интересные проекты, например, по озвучиванию описания наших достопримечательностей на китайский язык, а интересных объектов в Китае, соответственно, на русский.

Аудиогиды имеют большой потенциал применения в сфере инклюзии (для определённых групп нозологии – например, слабовидящих). В Красноярске создан тифлоаудиогид по Музею-усадьбе В. И. Сурикова «Суриков. Точка доступа», а также Инклюзивный тур по природным достопримечательностям.

Вывод

Потенциал платформы *izi.TRAVEL* для создания аудиогидов трудно переоценить. С его помощью становится возможным выполнение задач различных направлений *Digital Humanities*, а с применением краудсорсинга работа над проектом становится более интересной и менее затратной по времени.

Список источников

1. Как создать аудиогид на платформе *izi.TRAVEL* // Роза ветров. Север. Информационные технологии для начинающих. Просто о сложном. URL: ikt-masterilki.ru/izi-travel.
2. Лаптева М. А. Digital Humanities в России: перспективы развития / М. А. Лаптева, Е. А. Гордеева // Прикладная информатика. 2018. Т. 13. № 1 (73). С. 44–51. EDN YSCDYB.
3. Перкова Н. Краудсорсинг в Digital Humanities: опыт Латвийского фольклорного архива / Н. Перкова // СИСТЕМНЫЙ БЛОКЪ. 2020. URL: sysblok.ru/linguistics/kraudsorsing-v-digital-humanities-opyt-latvijskogo-folklornogo-arhiva.

Д. Ю. Гук¹, В. А. Пермяков²,

¹ старший научный сотрудник, кандидат филологических наук;

² специалист лаборатории Digital Humanities

¹ Государственный Эрмитаж, Санкт-Петербург, Россия

² Сибирский федеральный университет, Красноярск, Россия

e-mail: ¹ hookk@hermitage.ru, ² vapermyakov@sfu-kras.ru

ЭЛЕКТРОННАЯ ЭНЦИКЛОПЕДИЯ ЭРМИТАЖА. ИТОГИ ПЕРВОЙ ПЯТИЛЕТКИ

Аннотация. Статья посвящена разработке и поддержанию мультимедийного двуязычного веб-ресурса об истории археологических исследований и формирования археологических коллекций Государственного Эрмитажа. Авторы рассматривают проблему показа археологических коллекций музея в виртуальном пространстве с ориентацией на разную глубину интереса, а также на разную степень готовности информационного и визуального обеспечения. Эксплуатация веб-сайта как информационной системы на протяжении пяти лет подтвердила правильность выбранной структуры данных и привела к ряду важных выводов, связанных с вопросами администрирования и информационной безопасности.

Ключевые слова: археология, Эрмитаж, электронная энциклопедия, CIDOC-CRM, информационная безопасность

Публичный показ археологических коллекций в музее представляет собой не менее сложную задачу, чем сохранение информации об исследованиях в целом. Музейные предметы из собрания древностей требуют освещения исторического и археологического контекста и рассказа об исследователях, связанных с этим археологическим памятником. Если в прошлом веке можно было довольствоваться текстами и иллюстрациями на бумажной основе, то теперь появилось множество новых форм, доступных исключительно в цифровом виде через интернет. Данные, находящиеся в открытом доступе, нуждаются в систематизации и проверке на актуальность и достоверность.

Концепция *CIDOC-CRM (ISO 21127:2014)*, в основе модели которой лежит некое «событие», позволяет перейти от трёхмерного пространства на поверхности Земли, где это событие происходило или длилось какое-то время, в плоскость участников событий, к археологическим объектам, которые были изучены, а также музейным экспозициям и фондам. Сущность «Событие» включает в себя всевозможные категории: «открытие памятника», «находка», «раскопки», «исследования», «реставрация», «выставка», «конференция», «публикация» и т. д. Глубина описания во всех случаях разная: от простого указания места события на «Яндекс Карты» до ссылок на публикации материалов многолетних раскопок. Инструменты гипертекста обеспечивают доступ к внешним сетевым ресурсам (энцикло-

педиям, архивам, видеозаписям, панорамам и 3D-моделям). Каждая из сущностей «лицо», «локация», «памятник» связана с «Событием» отношением «многие-к-многим», что позволяет выполнять контекстные поисковые запросы, отталкиваясь от любых исходных данных. Таким образом, модель данных становится многомерной и сетевой.

Силами молодых специалистов из разных стран была разработана рабочая модель такого ресурса, посвящённого археологии в Государственном Эрмитаже (URL: archaeoglobus.com). Сотрудники музея принимали участие в подготовке данных, руководствуясь принципами *Web 2.0*, т. е. добровольно, проявляя инициативу по мере сил и возможностей. Ссылка на источник данных обязательна, как и примечание в том случае, если были выявлены ошибки или неточности. Также важно отметить тот факт, что история формирования археологических коллекций тесно связана с историей музея, России и с международным сотрудничеством, и это можно увидеть прямо на интерактивной карте на заглавной странице сайта [1–4].

Инструменты веб-аналитики продемонстрировали позитивную динамику использования разработки в период пандемии *COVID-19* [5], в частности рост числа виртуальных посетителей и среднего времени посещения, превышающего средний показатель для сайтов в области культуры. Однако недостаточное внимание к вопросам администрирования и безопасности повлекло те последствия, которым посвящено настоящее сообщение.

Восстановление работы сайта «Электронная Энциклопедия Эрмитажа» было сложной и трудоёмкой задачей. Сайт был подвергнут атаке, и злоумышленник получил доступ к административной панели через уязвимость в версии *CMS WordPress*, что привело к публикации неуместной информации.

Для начала мы провели полную очистку сайта от вредоносных файлов. Это включало сканирование всех файлов и папок на наличие подозрительного и вредоносного кода. После обнаружения и удаления такого кода мы установили антивирусный и защитный плагин, чтобы предотвратить повторное внедрение вредоносного кода.

Далее мы удалили всех неизвестных пользователей и администраторов, чтобы устранить возможность несанкционированного доступа к административной панели. После этого мы обновили пароли административных аккаунтов, установив сильные, сложные комбинации для усиления безопасности.

Важной частью процесса было переключение сайта на протокол *HTTPS*, который обеспечивает безопасное соединение между пользователем и сайтом. Однако возникла проблема с тем, что не все страницы корректно открывались по этому протоколу. Чтобы решить это, нам пришлось вручную настроить перенаправление с протокола *HTTP* на *HTTPS* для всех страниц сайта.

После проведения всех этих операций мы смогли успешно восстановить работу сайта «Электронная Энциклопедия Эрмитажа», обеспечив его стабильную и безопасную работу.

Перспектива развития проекта была заложена в самой идее много-томной энциклопедии. Попытка добавить следующий том, посвящённый библиотекам, была предпринята два года назад, но потерпела фиаско из-за недостатка профессиональных навыков у исполнителя. Накапливаемый разнородный контент ждёт своего часа.

Авторы признательны В. Н. Бумастрову, первому разработчику структуры сайта, за консультации по вопросам администрирования и сопровождения.

Список источников

1. Гук Д. Ю. Применение информационных технологий в изучении и популяризации археологических коллекций / Д. Ю. Гук, М. В. Румянцев, Н. О. Пиков и др. // V (XXI) Всероссийский археологический съезд: сб. науч. тр. / отв. ред.: А. П. Деревянко, А. А. Тишкин. Барнаул: АлтГУ, 2017. С. 297.

2. Гук Д. Ю. О концепции распределённого мультимедийного информационного ресурса «Электронная Энциклопедия Эрмитажа. Том 1. Археология» / Д. Ю. Гук // Электронные ресурсы библиотек, музеев, архивов: матер. XII Всеросс. НПК (СПб., 2018). СПб.: Политехника-сервис, 2018. С. 35–47.

3. Гук Д. Ю. Молодые молодым: международное сотрудничество в области создания веб-ресурса по истории археологического собрания Государственного Эрмитажа / Д. Ю. Гук // Северный археологический конгресс. Ханты-Мансийск, 2019.

4. Гук Д. Ю. Археология Эрмитажа в формате электронной энциклопедии / Д. Ю. Гук // Труды ВАС-6. 2020. Т. 3. С. 118–119.

5. Гук Д. Ю. Релевантность онлайн-ресурсов музея в период пандемии / Д. Ю. Гук, Е. А. Сулова // Современные тенденции в развитии музеев и музееведения: матер. IV Всеросс. НПК с междунар. уч. (Новосибирск, 2020) / отв. ред.: О. Н. Шелегина, Г. М. Запорожченко. Новосибирск: НГУ, 2020. С. 141–147.

Н. О. Пиков¹, А. П. Дворецкая², В. А. Меховский³, А. С. Дяченко⁴,
¹ старший преподаватель;
² кандидат исторических наук, доцент;
^{3,4} специалист лаборатории Digital Humanities
^{1,2,3,4} Сибирский федеральный университет, Красноярск, Россия
e-mail: ¹ nikita.pikov@mail.ru, ² advoreckaya@mail.ru,
³ mehovsky.zenit-champion@yandex.ru, ⁴ Alinadyachenko12@gmail.com
ORCID: ² 0000-0002-0140-3704

ВИРТУАЛЬНАЯ РЕПРЕЗЕНТАЦИЯ САКРАЛЬНОГО ПРОСТРАНСТВА ЕНИСЕЙСКОЙ СИБИРИ

Аннотация. В статье рассматриваются методы виртуальной репрезентации сакрального пространства Енисейской Сибири. Показаны причины обращения к виртуальной репрезентации и её использования в современных исторических исследованиях на основе визуальных источников. Даются примеры подобного рода репрезентаций, подготовленные в рамках МГУ и СФУ. Дается описание интернет-ресурса *SACRA*, подготовленного специалистами лаборатории *Digital Humanities*, который бы смог совместить все виртуальные продукты и технологии. Делается вывод, что данный ресурс позволяет создать виртуальное открытое пространство для нужд познавательного туризма и музеев, школьных уроков.

Ключевые слова: технологии 3D-моделирования и фотограмметрии, виртуальные туры, Енисейская Сибирь

В настоящее время для воссоздания исторических ландшафтов огромную роль играют возможности технологий 3D-моделирования и фотограмметрии в воссоздании пространственных, визуальных образов исторической городской среды, эволюционировавшей на протяжении веков. Возникающие при этом источниковедческие задачи имеют сходство с традиционными, присущими любой реконструкции исторических процессов и явлений. Однако виртуальная реконструкция имеет свою специфику, связанную с ведущей ролью визуальных источников. Обращение ко всему комплексу доступных источников даёт возможность восстановить расположение и размеры монастырских строений, их изменявшийся облик путём сопоставления полученной из описательных источников информации с данными графических и изобразительных источников.

Создание 3D-модели позволяет проверить ряд гипотез о внешнем облике здания, его эволюции ещё на этапе макетирования, а также оценить степень достоверности и точности графических исторических источников.

Преимуществом таких 3D-моделей является возможность интерактивного просмотра в онлайн-режиме и верификации созданной виртуальной реконструкции. Предложенная в проекте процедура верификации подразумевает возможность взаимодействия пользователя с представленными

на сайте источниками реконструкции (текстовой документацией, чертежами, планами, фотографиями, живописными произведениями и т. п.) и созданной на их основе виртуальной 3D-моделью с подробным описанием методики её построения применительно к каждому зданию, а также с виртуальным туром по религиозному месту. При этом возникает как новая источниковедческая задача – презентация всех источников, использованных для восстановления рассматриваемого фрагмента здания (с соответствующей критикой источников), так и технологическая: как реализовать этот подход, какими программными инструментами пользоваться.

Главная проблема – формирования и использования электронной среды виртуальных реконструкций и туров. Создание такой среды даёт возможность презентации результатов проекта в интернете, включая детализированные трёхмерные модели. Такую возможность, например, даёт реконструкция трёхвековой эволюции монастырского комплекса и окружающей его Страстной площади Москвы, осуществлённая с помощью технологий 3D-моделирования, показывает те новые возможности в развитии исторической урбанистики, которые открылись перед историками. Полученные результаты представлены в открытом доступе на сайте исторического факультета МГУ им. М. В. Ломоносова (URL: hist.msu.ru/Strastnoy). Полученный опыт может быть использован в дальнейших исследованиях по созданию виртуальных реконструкций утраченных объектов историко-культурного наследия [1].

Виртуальная реконструкция построек Спасского мужского монастыря в г. Енисейске, включая утраченную надвратную церковь Захария и Елизаветы, выполненная в Сибирском федеральном университете, позволила создать виртуальные экскурсии по объектам, имеющим историческое и культурное значение, наделять их интерактивностью, т. е. возможностью перемещаться в любом направлении и взаимодействовать с объектами в реальном времени. Она была сделана в технологии, которая представляет собой информационную модель, своеобразный накопитель информации, с помощью которого создаются и воссоздаются утраченные элементы.

Необходимость оцифровки памятника архитектуры обуславливается очень часто практическими задачами, а конкретно – целью его дальнейшего использования при производстве работ по содержанию, реконструкции и реставрации. Так, Зашиверская Спасская церковь, в своё время перевезённая из Якутии в Новосибирск, подлежала реставрации, был повторён путь сборки здания при помощи *VIM*-технологии после его перемещения на новое место, что позволило паспортизировать все архитектурные элементы данного сооружения [2; 3].

Существенное расширение аудитории пользователей ведёт к техническому прогрессу при разработке программ, позволяющих создавать и демонстрировать модели, имитировать высокую степень реалистичности виртуального пространства, представить объёмно-пространственную и планировочную структуру объекта, декоративное убранство фасадов, со-

масштабность окружающей среде. При тщательной проработке хорошо видны техники кладки, материалы, декор здания.

Многовековая история религий Енисейской Сибири с каждым годом привлекает всё большее внимание и вызывает значительный интерес как жителей региона, так и гостей. Вероучения Енисейской Сибири хранят в себе огромное количество накопленной мудрости, которую мы стремимся с пониманием и уважением передать через наш интернет-ресурс *SACRA*.

В качестве основного тона сайта был выбран тёмно-синий цвет, т. к. он символизирует глубину и мудрость, присущие религиям. Также, чтобы передать святое и духовное значение конфессий, был использован белый цвет. Он олицетворяет чистоту и указывает путь к сакральным местам. Религии Енисейской Сибири могут иметь различные обычаи и традиции, но нам хотелось подчеркнуть их общую связь. В качестве знака гармонии в дизайне использовался голубой цвет, который воплощает реку Енисей, неразрывно связывающую народы, живущие вдоль её берегов.

Логотип сайта – слово *SACRA*, что в переводе с латинского означает «священная». Река Енисей, протекающая через все регионы Енисейской Сибири, фрагментарно представлена в каждой букве слова *SACRA*. Таким образом, рождается смысл, заложенный в фирменный знак – священные земли, объединённые рекой Енисей. Логотип *SACRA* визуальнo отражает особенность Енисея, который начинается от слияния двух рек – Малого и Большого Енисея. Так, буква *S* содержит две линии воды, вместо одной, отсылая к началу Енисея и демонстрируя идею о том, что религия начинается с веры, – так же, как река с истока.

Интернет-ресурс представляет собой информационный портал о религиях Енисейской Сибири, основным методом представления религиозных мест являются виртуальные туры, с помощью которых человек имеет возможность посетить это место и насладиться его интерьером в интерактивном формате. Здесь представлены все основные вероисповедания Енисейской Сибири. Также будет предусмотрена возможность добавления публикаций о культовом месте, что делает сайт научно-направленным. Помимо этого, на веб-ресурсе присутствует интерактивная карта для более наглядного представления религиозных мест.

Для того чтобы максимально передать атмосферу религиозного места использованы следующие методы представления пространства в цифровом формате.

1. Виртуальная реконструкция – одно из наиболее молодых прикладных научных направлений, к которому прибегают для моделирования и демонстрации своих экспериментов представители самых разных наук, в т. ч. исторических. Известными достоинствами виртуальной реконструкции являются короткие сроки проведения (в сравнении с традиционными методами), высокая точность, возможность использования полученных материалов в интернете и др.

2. Виртуальный тур – вид представления трёхмерного пространства в цифровом виде, который позволяет человеку оказаться в желаемом месте

лишь при помощи интернета. В контексте проекта Сакральное пространство Енисейской Сибири возможно не только ознакомиться с картинками и виртуальными реконструкциями религиозных мест, но и посетить это место, изучив его внешние архитектурные особенности и внутреннее убранство храмов и молитвенных мест.

Религиозное наследие Енисейского Севера формировалось на протяжении нескольких веков. Среди данных объектов особую ценность представляют сакральные объекты старожилов, выполненные в камне. К ним, например, относятся культовые объекты г. Енисейска. Часть из них подверглась реставрации, часть оказалась разрушена. В 2022 г. специалистами-архитекторами СФУ под руководством кандидата архитектуры А. В. Слабухи выполнена реконструкция следующих утраченных культовых сооружений г. Енисейска: Преображенская, Рождественская церкви, Крестовоздвиженская, часовня Св. Николая, синагога. Отреставрированные же церкви вошли в виртуальный тур по г. Енисейску, подготовленный сотрудниками лаборатории *Digital Humanities*. Уникальным явлением, связанным с Енисейским Севером, стала и традиция строительства деревянных церквей в Приангарье. В 2023 г. сотрудниками лаборатории *Digital Humanities* (консультант-архитектор А. В. Слабуха) подготовлена реконструкция культовых построек в деревнях Мозговой, Селенгино, Яркино Кежемского района Приангарья.

Ещё несколько объектов, подготовленных к съёмке, находятся в руинированном виде. Это церкви в с. Анциферово и Каменка Енисейского района, а также требующая серьёзной реконструкции домовая церковь в с. Широково Казачинского района. По сохранившимся документам будет выполнена виртуальная реконструкция, также будет проведена панорамная съёмка по этим местам.

Заключение

В настоящее время такие методы сохранения культурного наследия привлекают не только историков и специалистов в области прикладной информатики, но и архитекторов, художников, дизайнеров, музеологов, реставраторов. Новые технологии позволяют не только провести реконструкцию утраченных объектов и их отдельные элементы, но и попытаться с помощью заданных координат вписать их в виртуальное городское пространство, разместить на карте, показать эволюцию используемых строительных технологий, приёмы реставрации.

Кроме основной задачи воссоздания объектов с помощью визуальных источников возникают и другие аспекты образовательного, культурного и туристического характера – это моделирование окружающего ландшафта на протяжении веков для историко-культурной и археологической реконструкции, виртуальная музеефикация, в т. ч. действующих объектов, и мониторинг текущего состояния объекта культурно-исторического наследия, создание виртуальных экскурсий и виртуального открытого пространства для нужд познавательного туризма и музеев, школьных уроков.

Исследование выполнено в рамках проекта РНФ «№ 22-28-01668 “Комплексное исследование сакрального пространства старожилов Енисейского Севера: механизм сохранения и развития (на основе междисциплинарного исследования памятников истории и архитектуры)”» (рук. А. П. Дворецкая). URL: rscf.ru/project/22-28-01668.

Список источников

1. Бородкин Л. И. Виртуальная реконструкция монастырских комплексов Москвы: проекты в контексте Digital Humanities / Л. И. Бородкин // Вестник ПГНИУ. Сер.: История. 2014. № 3 (26). С. 259, 268, 271–273.

2. Бородкин Л. И. Виртуальная реконструкция Страстного монастыря (XVII–XX вв.): первый этап проекта / Л. И. Бородкин, Д. И. Жеребятъев, Р. Б. Кончаков и др. // Информационный бюллетень Ассоциации «История и компьютер». 2014. № 42. С. 216–218.

3. Майничева А. Ю. Новый подход к сохранению памятников русского деревянного зодчества: применение технологии BIM / А. Ю. Майничева // Уральский исторический вестник. 2018. № 1 (58). С. 135–140.

Д. С. Шемелина¹, Л. К. Радченко², Н. С. Декунцов³,

¹ кандидат архитектуры, доцент;

² кандидат технических наук, доцент;

³ магистрант направления «Архитектура»

^{1,3} Новосибирский государственный университет архитектуры, дизайна и искусств
им. А. Д. Крячкова, Новосибирск, Россия;

^{1,3} Лаборатория архитектурно-градостроительных исследований военно-исторического
наследия Сибири, Новосибирск, Россия;

² Сибирский государственный университет геосистем и технологий,
Новосибирск, Россия

e-mail: ¹ dasha-shem@yandex.ru, ² l.k.radchenko@sgugit.ru, ³ dns0607@yandex.ru

ORCID: ¹ 0009-0000-6948-5067, ² 0009-0009-0448-8094, ³ 0009-0008-0718-9850

АРХИТЕКТУРНО-ГРАДОСТРОИТЕЛЬНОЕ НАСЛЕДИЕ СИБИРСКИХ ОБОРОНИТЕЛЬНЫХ ЛИНИЙ XVIII ВЕКА: ПОИСК ПЕРСПЕКТИВ РЕВАЛОРИЗАЦИИ С ПОМОЩЬЮ ГЕОИНФОРМАЦИОННОЙ МОДЕЛИ

Аннотация. В работе представлены первые итоги разработки геоинформационной модели морфологии архитектурно-градостроительного наследия сибирских оборонительных линий XVIII в. с целью определения перспектив ревалоризации соответствующих памятников. Обозначены особенности применения геоинформационного моделирования в историко-архитектурном исследовании. Представлены сведения о выполнении задач по разработке структуры базы данных для создания основы геоинформационной модели, картографической визуализации базы данных, разработке библиотеки условных знаков, выведению статистических данных, формированию производных тематических карт.

Ключевые слова: сибирские оборонительные линии, архитектурно-градостроительное наследие, геоинформационная модель

В настоящее время в Новосибирском государственном университете архитектуры, дизайна и искусств им. А. Д. Крячкова в Лаборатории архитектурно-градостроительных исследований военно-исторического наследия Сибири в сотрудничестве с кафедрой Картографии и геоинформатики Сибирского государственного университета геосистем и технологий (Новосибирск) начата разработка геоинформационной модели «Архитектурно-градостроительное наследие сибирских оборонительных линий XVIII в.».

Этот протяжённый фортификационный комплекс, состоявший из Иртышской, Горькой и Колывано-Кузнецкой линий, создавался в XVIII в. для защиты российских границ в Сибири. Входившие в их состав военно-инженерные объекты разной мощности – долговременные (крупные крепости) и полевые укрепления – простирались от Урала и шли далее на восток. Являясь уникальными объектами военного урбанизма, укрепления сибирских линий сочетали в себе отечественные подходы охраны

рубежей с применением основ европейской теории фортификации (см., например, [4]). К сожалению, на сегодняшний день ценность архитектурно-градостроительного наследия сибирских линий раскрыта очень слабо, а историческая память жителей сибирского региона о данных объектах крайне размыта. Государственная граница, проведённая между Россией и Казахстаном после развала СССР, разделила единый фортификационный комплекс между двумя странами. В связи с этим остатки укреплений сегодня находятся по обе стороны границы, что дополнительно усложняет вопрос их сохранения как архитектурно-градостроительного наследия [5].

В данной работе представлен опыт применения методов цифровой гуманитаристики, в частности геоинформационного моделирования, к области истории архитектуры. В рамках нашего исследования геоинформационное моделирование направлено на определение перспектив ревалоризации¹⁰ архитектурно-градостроительного наследия сибирских линий и создания основы для последующего мониторинга происходящих с ним изменений. Геоинформационная модель рассматривается нами как инструмент, с помощью которого представляется возможным достичь обозначенной в исследовании цели по выявлению перспектив ревалоризации изучаемых объектов. В связи с этим на данном этапе геоинформационная модель разрабатывается нами в узкоспециальных научно-исследовательских целях, что не исключает возможности её дальнейшей адаптации и использования в просветительско-образовательной деятельности.

В отличие от многих геоинформационных проектов, посвящённых историко-культурному наследию, разрабатываемая нами модель имеет специализированный характер, оставаясь в поле теории и истории архитектуры, а также в области военного урбанизма – модель объединяет как данные по истории создания укреплений, так и данные, полученные из анализа их фортификационных и планировочных характеристик, сведений об их современном состоянии, а также из анализа мер охраны, принимаемых для сохранения укреплений.

Обращение в ходе нашей работы к методам геоинформационного моделирования продиктовано системным характером фортификационного комплекса сибирских линий, большим количеством составлявших его укреплений, разнородностью связанных с ними данных и характеристик, на основе которых в конечном счёте будут определены адекватные методы ревалоризации архитектурно-градостроительного наследия сибирских линий. Ещё одной причиной применения геоинформационных технологий к исследуемым объектам стала необходимость многократной картографической визуализации данных для представления различных характеристик наследия сибирских линий. Заметим, что до сих пор в сферах деятельности, связанных с архитектурой и градостроительством, широко использу-

¹⁰ В данном случае под термином «ревалоризация» понимается совокупность научно обоснованных мероприятий, направленных на сохранение объекта культурного наследия и позволяющих при минимальном вмешательстве в конструктивную систему наделить объект новой современной функцией и вернуть ему утраченную ценность [1; 3].

ется менее технологичное и неинтерактивное традиционное картографирование в графических редакторах. Кроме того, обращение к геоинформационным технологиям вызвано также и необходимостью создания современной сводной карты сибирских линий, которая до сих пор отсутствует в исследовательском поле обозначенной тематики.

Территориальные границы, в которых разрабатывается модель, охватывают современную территорию Западной Сибири, Алтая и прилегающих районов северного и северо-восточного Казахстана, где были возведены Иртышская, Горькая и Колывано-Кузнецкая оборонительные линии. Хронологические границы определены периодом с 1763 г. (когда был издан указ Екатерины II о необходимости модернизации сибирских оборонительных линий XVIII в., согласно которому началась перестройка укреплений [4]) по настоящий момент времени, в который можно оценить современное состояние укреплений.

В рамках создания геоинформационной модели были поставлены следующие задачи:

- разработка структуры базы данных и её наполнение для создания основы геоинформационной модели;
- картографическая визуализация базы данных с помощью программного обеспечения *MapInfo Professional*;
- разработка библиотеки условных знаков для выполнения картографической визуализации;
- выведение статистических данных в виде инфографики;
- формирование производных тематических карт;
- определение перспектив ревалоризации архитектурно-градостроительного наследия сибирских оборонительных линий.

Ниже представим сведения о выполнении перечисленных выше задач.

Разработка структуры базы данных и её наполнение для создания основы геоинформационной модели

Необходимость применения системного подхода к фортификационным комплексам как к системе памятников, рекомендованного Руководством по укреплениям и военному наследию ИКОМОС¹¹ [7], множественность элементов, входивших в состав сибирских линий, их разнообразие и задача учёта большого количества их характеристик обусловили отказ авторов данной работы от текстового, нарративного представления данных в пользу их организации в виде базы данных с последующим созданием на её основе геоинформационной модели.

Источниками и материалами, используемыми для заполнения базы данных и создания геоинформационной модели, стали:

- копии графических, картографических и текстовых материалов, хранящихся в федеральных и региональных архивных собраниях РГИА,

¹¹ ИКОМОС – Международный Совет по сохранению памятников и достопримечательных мест (*ICOMOS*).

РГАВМФ, РГВИА, ИАОО, ЦХАФ АК¹², из коллекций канд. арх. Д. С. Шемелиной и канд. арх. А. А. Гудкова, в т. ч. фиксационные планы и проекты перепланировки линейных укреплений, а также карты трассировки сибирских оборонительных линий XVIII в.;

– копии гравированных планов укреплений из европейских трактатов по теории фортификации (XVII в.) и их переводов на русский язык, выполненные в XVIII в., из европейских и российских собраний редких книг;

– данные спутниковых карт («Яндекс Карты», *Google Maps*), сведения из официальных документов и списков охраны объектов культурного наследия, а также данные натурных визуальных обследований, фотофиксаций и результаты наблюдений, полученные краеведами.

Известно, что для данных, которые используются в гуманитарных исследованиях, свойственны неполнота, фрагментарность, неопределённость, разнообразие и противоречивость. Не стали исключением данные, на которые опирается наша работа. Это выражено, например, в том, что для некоторых крепостей по линиям нами пока не выявлены в архивных собраниях планы, составленные в рамках перепланировки укреплений по указу 1763 г. Кроме того, на текущий момент для некоторых укреплений пока не определена теоретическая основа планировки и фортификации, т. е. не установлены конкретные теоретические разработки из европейских трактатов по фортификации XVII–XVIII вв., к которым обращались в процессе проектирования укреплений сибирские военные инженеры.

В соответствии с общей исследовательской целью по поиску перспектив ревалоризации архитектурно-градостроительного наследия сибирских оборонительных линий XVIII в. при разработке структуры базы данных был определён набор сущностей предметной области (объектов картографирования), а также параметры, по которым они описывались.

На *этапе концептуального проектирования* базы данных в качестве объектов картографирования были обозначены материальные объекты архитектурно-градостроительного наследия сибирских линий, представленные 19 крупными крепостями и рассматриваемые одновременно в двух аспектах – как исторические объекты XVIII в. и как объекты современности (объекты культурного наследия). В качестве ещё одной сущности были определены сами сибирские оборонительные линии. При этом полевые укрепления, т. е. более мелкие укрепления, также входившие в состав линий, на текущий момент при разработке геоинформационной модели и проектировании соответствующей базы данных в качестве сущностей не определялись.

Для описания крупных крепостей, рассматриваемых как исторические объекты, были определены перечисленные ниже параметры. Для на-

¹² РГИА – Российский государственный исторический архив; РГАВМФ – Российский государственный архив военно-морского флота; РГВИА – Российский государственный военно-исторический архив, ИАОО – Исторический архив Омской области; ЦХАФ АК – Центральное хранилище архивного фонда Алтайского края.

полнения базы данных в данном случае использовались текстовый, числовой и графический типы данных. При этом данные выбирались на основе проектов перепланировки укреплений 1765 г., составленных после издания указа 1763 г. о необходимости модернизации линий. Приведём параметры, выбранные для описания крупных крепостей, рассматриваемых как исторические объекты:

- название укрепления;
- год основания;
- год создания и начала реализации проекта перепланировки;
- система обороны;
- теоретическая основа фортификации и планировки;
- типизация фортификации и планировки.

Для описания крупных крепостей, рассматриваемых как объекты современности (объекты культурного наследия – ОКН), были определены перечисленные ниже параметры. Для наполнения базы данных в данном случае использовались текстовый, графический и пространственный типы данных. При этом данные о сохранности фортификации основных территорий укреплений и планировки форштадтов (территорий крепостей, отведённых под гражданскую застройку) определялись также на основе проектов перепланировки укреплений 1765 г. в сравнении с данными спутниковых снимков. Приведём параметры (геометрические и семантические составляющие базы данных), выбранные для описания крупных крепостей, рассматриваемых в качестве объектов современности:

- координаты;
- название населённого пункта, основанного на территории укрепления;
- государственная принадлежность территории укрепления (Россия/Казахстан);
- тип поселения, основанного на территории укрепления (городское/сельское);
- сохранность фортификации основной территории укрепления (частичная / не сохранилась);
- сохранность планировки форштадта (частичная / не сохранилась);
- наличие статуса ОКН (присвоен/отсутствует);
- функционирование статуса ОКН (функционирует/нарушается);
- проведение работ по ревалоризации (проведены / не проводились);
- транспортная доступность (недоступно / труднодоступно / средняя доступность / доступно).

Для описания сибирских оборонительных линий как для объекта картографирования были определены такие параметры, как название линии и её трассировка на местности (для наполнения базы данных в данном случае использовались текстовый и пространственный типы данных).

Отметим, что отбор перечисленных параметров для описания крепостей осуществлялся исходя из значения этих параметров для поиска перспектив ревалоризации архитектурно-градостроительного наследия сибир-

ских линий. В общей сложности для анализа объектов, представленных в базе данных, было определено около двух десятков таких параметров.

Крепости, рассматриваемые как исторические объекты XVIII в., и крепости, рассматриваемые как объекты современности (объекты культурного наследия), были определены как сущности с типом связей «один к одному». При этом оборонительные линии, объединяющие отдельные элементы в виде крепостей, связываются с ними как «один ко многим».

На *этапе логического проектирования базы данных* для набора выявленных сущностей и связей выбирался тип модели представления данных. Нужно отметить, что само устройство сибирских оборонительных линий как системы содержит в себе признаки иерархии и линейности. Так, отдельные укрепления входили в состав одной из линий – Иртышской, Горькой или Колывано-Кузнецкой, они в свою очередь объединялись в систему сибирских оборонительных линий, которые на следующем уровне входили в общероссийскую линейную систему обороны границ. При этом укрепления на линии располагались в определённой последовательности, обусловленной географическими координатами и их принадлежностью к той или иной линии.

Для проектирования базы данных была выбрана реляционная модель в виде двумерного массива с уникальным именем каждого однородного столбца. При этом ввиду описанных выше особенностей, связанных с иерархией и линейностью сибирских оборонительных линий, порядок следования строк и столбцов в обсуждаемой реляционной модели не является произвольным, как в классическом её варианте, а соответствует последовательности укреплений на линии.

Следующим этапом работы будет *этап физического проектирования*, в ходе которого выбранная модель будет реализована в конкретной системе управления базой данных (СУБД). На данный момент для структурирования и организации информации используется стандартный табличный редактор *Microsoft Excel*.

Картографическая визуализация базы данных

Для создания геоинформационной модели на основе разработанной базы данных было выбрано программное обеспечение *MapInfo Professional*. В качестве источника географической информации использовался открытый ресурс *Open Street Map*, с которого была произведена загрузка общегеографической основы для геоинформационной модели с целью нанесения тематической нагрузки.

В информационную модель из базы данных были перенесены характеристики крепостей, соответствующие перечисленным выше параметрам. Одновременно на карту были нанесены точечные объекты, представляющие собой крупные крепости сибирских оборонительных линий. Кроме того, на карте с привязкой к территории и современным административным границам была обозначена общая трассировка самих линий, условно объединяющая входившие в неё объекты. В дальнейшем получившаяся карта будет дополнена информацией о размещении полевых укреплений – более

мелких фортификационных объектов, также входивших в состав оборонительных линий.

Разработка библиотеки условных знаков для выполнения картографической визуализации

В ходе разработки библиотеки условных знаков, применяемых в геоинформационной модели, был проведён анализ традиционных карт и близких по тематике геоинформационных моделей – например, *Fortes Cultura* (2018) [6], *La France de Vauban* (2007) [8], «Археологические памятники г. Омска и окрестностей» (2023) [2]. При изучении аналогов было установлено, что изображение использованных при их создании условных знаков является слишком абстрактным, что заставляет чаще обращаться к легенде. Кроме того, в некоторых разработках используются условные знаки, не соответствующие изображаемому фортификационному объектам – например, бастионные крепости отображаются знаком, представляющим объект средневековой фортификации со стенами и башнями, что не только искажает восприятие, но и не соответствует реальному устройству укреплений. Наиболее проработанным с точки зрения качества условных обозначений проектом-аналогом на наш взгляд является туристская карта *La France de Vauban* [8], посвящённая фортификационному наследию французского военного инженера XVII в. Себастьяна Ле Престра де Вобана. Используемые в ней условные знаки сочетают в себе простоту визуального восприятия, эстетичность, высокую информативность и соответствие фортификационным особенностям укреплений. С опорой на данную систему условных обозначений в настоящее время нами разрабатывается собственная библиотека условных знаков.

Выведение статистических данных в виде инфографики

Из геоинформационной модели будут извлечены основные статистические данные, которые позволят на основе повторяемости характеристик и их сочетаний разделить изучаемые объекты по группам. Ввиду важности для области архитектуры визуального представления информации статистические данные будут выведены в формате наглядной инфографики. На их основе будут обозначены приоритетные направления работы с наследием и разработан комплекс предложений по ревалоризации архитектурно-градостроительного наследия сибирских оборонительных линий XVIII в.

Формирование производных тематических карт

На основе созданной геоинформационной модели, с использованием извлечённых из неё статистических данных и разработанной библиотеки условных обозначений будут сформированы тематические карты. Картографическая визуализация позволит в наглядной форме представить основные группы изучаемых объектов с географической привязкой к местности. Кроме того, планируемым итогом реализации данной стадии должна стать полноценная комплексная карта сибирских линий, которая, как уже упоминалось выше, до этого не создавалась. Существующие на сегодняшний день карты, касающиеся линейной системы обороны границ Сибири

XVIII в., были созданы несколько десятилетий назад. Они демонстрируют только отдельные линии, а не всю систему в комплексе, ограничиваются лишь указанием географического размещения укреплений по линиям и не учитывают современный контекст размещения объектов относительно актуальных административных границ, современное состояние и другие характеристики укреплений на данных картах не визуализируются. Комплексная карта сибирских оборонительных линий, которая будет создана на данной стадии, будет направлена на ликвидацию перечисленных недостатков.

Полагаем, что как разрабатываемая нами геоинформационная модель, так и тематические карты, будут способствовать устранению низкой информированности об архитектурно-градостроительном наследии сибирских оборонительных линий XVIII в., которая, к сожалению, наблюдается как в обществе в целом, так и среди специалистов в области архитектуры в частности.

Список источников

1. Благовидова Н. Г. Кластерный подход к валоризации малых исторических городов на примере Сестрорецка / Н. Г. Благовидова, Н. В. Юдина. М.: Буки-Веди, 2020. 168 с.

2. Археологические памятники г. Омска и окрестностей: геоинформационная система. 2023. URL: klis.omsu.ru.

3. Светличная О. С. Современное понимание термина «ревалоризация» в теории реставрации памятников архитектуры и его трактовка / О. С. Светличная // Вестник ДонНАСА. 2013. № 1 (99). С. 194–197.

4. Шемелина Д. С. Реконструкция линейных крепостей 1765 г.: переход к европейской фортификации на сибирских рубежах / Д. С. Шемелина // Фундаментальные, поисковые и прикладные исследования РААСН по научному обеспечению развития архитектуры, градостроительства и строительной отрасли РФ в 2018 г.: сб. тр. М.: АСВ, 2019. Т. 1. С. 185–197.

5. Шемелина Д. С. Проблема сохранения архитектурно-градостроительного наследия сибирских оборонительных линий XVIII в. / Д. С. Шемелина, Н. С. Декунцов // Патриот: актуальные вопросы военной истории России и её силовых структур. Военная история России и воспитание гражданственности: матер. IV Всеросс. НПК (Новосибирск, 2023). Новосибирск: НИПКПРО, 2023. С. 111–115.

6. Forte Cultura. European Cultural Route. Fortified Monuments. 2018. URL: fortecultura.eu/media/flip/FoCu-Flyer2018-EN.

7. ICOMOS Guidelines on Fortifications and Military Heritage. GA 2021 6-1. Ver. 04/30/2020 // ICOFORT. URL: icofort.org/fortificationsguidelines.

8. La France de Vauban [Doc. Cartographique] // Réalisé et éd. par l'Institut Géographique National. No. 923. 1:1 000 000. Paris: IGN, 2007.

РАЗДЕЛ 2. ЦИФРОВАЯ ИСТОРИЯ

УДК 004.9

М. Н. Бабайцев,

аспирант

Тверской государственной университет, Тверь, Россия

ОБРАБОТКА ИСТОРИЧЕСКИХ ИСТОЧНИКОВ ИНСТРУМЕНТАРИЕМ OCR, ТАБЛИЧНЫХ ПРОЦЕССОРОВ И ГИС НА ПРИМЕРЕ СПИСКА НАСЕЛЁННЫХ МЕСТ ТВЕРСКОЙ ГУБЕРНИИ

Обработка исторических источников инструментарием *OCR* (оптического распознавания текста), табличных процессоров и ГИС даёт возможность получить обработанные данные, численно отражающие экономическую, социальную, политическую сферы жизнедеятельности конкретного общества в прошлом. Преимуществами такого инструментария является: во-первых, экономия временных издержек по сравнению с ручными способами обработки текста; во-вторых, увеличением объективности полученного знания. Конечные данные могут опровергнуть или подтвердить, частично или полностью, выдвинутые гипотезы, связанные с воздействием тех или иных факторов на жизнь человека в историческом контексте.

Цель обработки источника заключается в формулировании общей характеристики уездов Тверской губернии середины XIX в. на основе статистических данных, полученных посредством компьютерного анализа исторического источника [1].

Работа велась поэтапно:

- 1) перевод сканированных изображений источника в машиночитаемый вид с помощью технологий оптического распознавания текста;
- 2) чистка и унификация полученной информации посредством табличного процессора;
- 3) разработка алгоритмов обработки данных и их интерпретация в формулах;
- 4) создание сводной таблицы;
- 5) разработка и реализация кода на языке программирования для работы с *API* онлайн-карт для получения координат населённого пункта;
- 6) интерпретация данных в геоинформационной системе.

Перевод в машиночитаемый вид производился с помощью специализированного программного обеспечения *FineReader* версии 11 компании

АВВУУ [2]. Существуют готовые базы данных по источнику, однако информация, содержащаяся в них, оказалась неполной (отсутствуют записи с № 479 по № 752 и запись № 9469); некоторые названия населённых пунктов набраны с ошибками, в целом форматирование и структура данных не позволяют обрабатывать их формулами [3]. Исходя из этого было принято решение заново перевести источник в машиночитаемый вид (см. прил. 1).

Данная работа – продолжение исследования, начатого в 2018 г. Формулы и обработанные данные после подсчёта количества, суммы, среднего арифметического, минимального и максимального значений применительно к дворам, расстоянию от уездного города, населения и населённых пунктов см. в предыдущей работе автора [4].

Был сделан акцент на расстояние от становой квартиры (полицейского участка), плотность населения населённых пунктов, также обновлены данные по дворам, населению с учётом уездных городов. В отдельную группу выделено количественное соотношение населённого пункта и источника воды. Расширены критерии типов населённых пунктов. Добавлены подсчёты моды, размаха, стандартного отклонения, медианы для значений по каждому критерию для каждого уезда.

Дальнейшая обработка очищенных и унифицированных данных показала следующее.

Замечено, что населённый пункт может находиться при нескольких источниках воды одновременно. При этом предполагается, что авторы источника указывали самые значимые источники воды для населённого пункта. Исключением может быть лишь ситуация, при которой основным источником воды является, например, несколько ручьёв, – в таком случае источник указывался во множественном числе. Кроме того, имеется ряд записей, где не указан источник воды. Был произведён подсчёт статистики по следующим критериям: «Близ реки (рек)», «При реке(-ах)», «При речке(-ах)», «н/д», «На правом берегу реки Волги», «При озере(-ах)», «При колодеце(-ах)», «При ручье(-ях)», «Близ водохранилища», «При пруде(-ах)», «При болоте», «При овраге», «При притоке», «При протоке», «При плёсе (-ах)», «При резервуаре», «При роднике(-ах)».

Для подсчёта использовалась формула (на примере водохранилищ): =СЧЁТЕСЛИ(Тверской!\$E\$2:\$E\$466;"*близ водохранилища*").

Отдельно выделяются записи: «При резервуаре» – 3 записи в Новоторжском уезде, «При протоке» – 1 запись и «При притоке» – 8 записей в Корчевском уезде (скорее всего «протока» и «приток» – одно и то же, только с орфографической ошибкой), что может говорить о том, что у переписчиков не была до конца согласована система обозначений характеристик населённых пунктов.

Анализ записей по источнику воды показал, что самым популярным источником воды является колодец (5 060 населённых пунктов), далее – река (3 512), речка (1 155).

Остаётся пока непонятным до конца, почему было важно считать расстояние от полицейского участка (становая квартира) до населённого пункта. Есть предположение, что это как-то связано с политикой и разного рода недовольством крестьян. Однако в среднем от населённого пункта до становой квартиры расстояние было равно 25,6 км, при этом минимальное – 1,07, максимальное – 709,42 км. В формуле важно было указать коэффициент перевода из старорусской версты в километры, т. к. в источнике используется верста, а на практике привычны километры. Формула перевода (на примере Старицкого уезда): =Старицкий!G773*1,0668.

При уточнении данных по типу населённого пункта (табл. 1) выяснилось, что переписчики могли классифицировать населённый пункт неоднозначно – например, «Погост и сельцо» или «Сельцо и деревня (сц. и д.)», поэтому при подсчёте статистики это было учтено. В табл. 1 в столбце «Количество» указана сначала сумма всех записей по уезду, далее – среднее арифметическое на уезд.

Таблица 1

Количественные показатели по типу населённого пункта

Тип населённого пункта	Количество
Деревня (д.), в т. ч. комбинированные	8 649 / 721
Сельцо (сц.), в т. ч. комбинированные	1 411 / 118
Село (с.), в т. ч. комбинированные	543 / 46
Нет данных (н/д)	239 / 20
Слобода (сл.)	25 / 3
Погост (погост), в т. ч. комбинированные	1 / 1
Монастырь (мон.), в т. ч. комбинированные	13 / 2
Хутор (хутор), в т. ч. комбинированные	19 / 2
Станция (станция)	4 / 1
Ферма (ферма)	2 / 1
Мыза (мыза)	8 / 1
Мельница (мельница)	3 / 1
Выселок (выс.)	12 / 1
Приселок (прис.)	4 / 1
Погост (пог.), в т. ч. комбинированные	209 / 18
Скотный двор (скотн. дв.)	75 / 7
Фабрика (фабрика)	1 / 1
Усадьба (усад.)	177 / 15
Постоялый двор (пост. дв.)	23 / 2
Харчевня (харчевня)	1 / 1
Посад (посад)	0 / 0
Хутор (хут.), в т. ч. комбинированные	1 / 1
Погост и сельцо (пог. и сц.)	0 / 0
Погост и село (пог. и с.)	1 / 1
? (вц.)	1 / 1
Завод (зав.)	2 / 1
Плотина-бейшлот (бейшлот)	1 / 1
Сельцо и деревня (сц. и д.)	2 / 1
? Монастырь (пос. мон.)	2 / 1
? Хутор (дв. хут.)	1 / 1
Почтовая станция (почт. ст.)	2 / 1

Ещё не были обработаны данные по принадлежности населённого пункта к государству, церкви или частному лицу. Это задача для дальнейшего исследования.

Таким образом, сельский уточнённый «портрет» Тверской губернии на середину XIX в. выглядит следующим образом: в среднем населённые пункты удалены от уездного города на 37 вёрст и от становой квартиры на 25,6 км. При этом самый близкорасположенный к городу населённый пункт находился в Тверском уезде (0,25 вёрст) а самый удалённый – в Новоторжском (232 версты). Близкорасположенные к становой квартире населённые пункты находились во всех уездах губернии в Тверском уезде (минимальное расстояние – 1,07 км / 1 верста), а самый удалённый – в Корчевском (709 км). В среднем уезд включал 14 303,5 двора (село), при этом максимальное количество дворов – в Бежецком уезде (27 684), минимальное – в Тверском (8 229). Всего дворов в губернии – 171 642. В сельской местности самый густонаселённый уезд – Бежецкий (184 294 чел.), самый малонаселённый – Тверской (66 399 чел.). В среднем в сельской местности в уезде проживало 108 061,4 чел. при сельском населении губернии в 1 404 798,4 чел. В среднем на двор приходилось 7,7 чел. Всего населённых пунктов – 11 329 без учёта городов. В среднем дворов на сельский населённый пункт – 15,8. В среднем на сельский населённый пункт приходилось по 121 чел. Большая часть сельских населённых пунктов находилась при искусственных источниках воды – колодцах, остальная – при реках.

Подробную статистику можно посмотреть на странице сайта [5].

Далее была написана простая программа на языке *Python*, которая взаимодействует с API «Яндекс Карт» и подтягивает координаты населённых пунктов (прил. 2). Стоит отметить, что данные были предварительно подготовлены для обработки алгоритмом. Так, запись «Щербово» с привязкой к уезду была переформирована в запись «Россия, Тверская область, Сандовский муниципальный округ, деревня Щербово». Однако этот способ не совсем точный: во-первых, могут встречаться одноимённые населённые пункты в разных уездах; во-вторых, старые названия и названия с современных карт могут не совпадать, что критично для автоматической обработки; в-третьих, Корчевской уезд был упразднён, а часть населённых пунктов расселена и затоплена при строительстве Иваньковского водохранилища. Таким образом, координаты населённых пунктов могут не соответствовать действительности или не быть найдены.

Для визуализации полученных данных была использована веб-ГИС *Google Maps* (сервис «Мои карты»). После размещения населённых пунктов на карте видны указанные выше недочёты. В частности, некоторые населённые пункты оказались далеко за пределами Тверской губернии и Тверской области, что является невозможным (прил. 3).

Таким образом, современный цифровой инструментарий позволяет обрабатывать статистическими методами и графически визуализировать большие объёмы числовых и текстовых данных, позволяя тем самым приблизиться к «истине» и взглянуть на исследуемый предмет и объект с нескольких ракурсов.

Список источников

1. Списки населённых мест Российской империи, составленные и издаваемые Центральным статистическим комитетом МВД: [по сведениям 1859 г.] / под ред. И. Вильсона. СПб.: Изд-во ЦСК МВД, 1861–1885. Вып. 43: Тверская губерния. 1862. XXXVIII, [2]. 454 с.

2. ABBYY FineReader. URL: pdf.abbyy.com.

3. Дубоделов П. С. Электронный вариант СНМ Тверской губернии / П. С. Дубоделов. URL: boxpris.ru/svg/?p=1634.

4. Бабайцев М. Н. Использование инструментария MS Excel для анализа исторических источников на примере списка населённых мест Тверской губернии 1859 г. / М. Н. Бабайцев, И. П. Комаров // Наука. Молодость. Талант: сб. ст. Вып. 8. Тверь: ТвГУ, 2018. С. 5–7.

5. Список населённых мест Тверской губернии. URL: dighistory.ru/loftp.

6. Исторические ГИС. URL: dighistory.ru/historical_gis.

Приложение 1

Оптическое распознавание исторического источника

№		НАВАНИЕ НАСЕЛЕННЫХ МЕСТ.	ПОЛОЖЕНИЕ.	Отъ уезд. станц.	Отъ станц. кварт.	Число дворов.	Число жителей.	Церкви и молельные здания; учебныя и благотворительныя заведенія; почтовая станція; ярмарка; базары; пристани; фабрики и заводы и т. п.
№	НАВАНИЕ НАСЕЛЕННЫХ МЕСТ.	ПОЛОЖЕНИЕ.	Отъ уезд. станц.	Отъ станц. кварт.	Число дворов.	Число жителей.		Церкви и молельные здания; учебныя и благотворительныя заведенія; почтовая станція; ярмарка; базары; пристани; фабрики и заводы и т. п.
						м. п.	ж. п.	
3995	Гарусово, с. вл.	при прудѣ и колодцахъ.	29	25	32	128	130	Часовня 1.
3996	Третьяково, д. вольн.	при безымянномъ ручьѣ.	29	19	25	50	59	
3997	Широкое, д. вольн. и вл.	при колодцахъ.	30	20	50	99	141	
3998	Заборько, пот.	при колодцахъ.	30	25	10	32	34	Церковь православная 1.
3999	Нашино, д. вольн.	при р. Погожевѣ.	30	20	22	44	50	
4000	Косино, д. вольн.	при колодцахъ.	30	20	22	47	63	
4001	Козьмино, д. вольн.	при колодцахъ.	32	28	21	47	47	
4002	Кузьмино, с. вл.	при прудѣ и колодцахъ.	32	28	20	67	83	Часовня 1.
4003	Никишино, д. вольн.	при р. Погожевѣ.	33	24	24	35	66	Часовня 1.
4004	Шушино, д. вольн.	при р. Чертковѣ.	34	24	30	104	133	Часовня 1.
4005	Матвеево, д. вл.	при колодцахъ.	35	25	23	44	38	
4006	Павлово, д. вл.	при р. Погожевѣ.	35	25	21	80	98	
4007	Жуково, д. вольн.	при р. Окозѣ.	35	30	20	52	54	
4008	Лужинское, с. вл.	при р. Колурѣ.	35	30	23	114	133	Церковь православная 2.
4009	Никитино, д. вл.	при колодцахъ.	38	33	15	67	67	
4010	Броннино, д. вольн.	при прудѣ и колодцахъ.	36	26	16	41	47	
4011	Дуваново, д. вл.	при колодцахъ.	36	34	7	32	30	
4012	Косино, д. вольн.	при р. Шитовѣ.	36	32	60	173	203	
4013	Васильевское, д. вольн.	при р. Сидѣ.	32	26	26	62	63	
4014	Масаринское, д. вл.	при колодцахъ.	36	31	42	140	164	
4015	Горки, д. вольн.	при колодцахъ.	30	25	26	33	76	
4016	Мельничное, д. вольн.	при р. Погожевѣ.	30	27	29	69	65	Часовня 1.
4017	Молодичи, д. вольн.	при прудѣ и колодцахъ.	33	27	10	16	37	
4018	Кобровское, д. вл.	при р. Сидѣ.	35	28	35	131	136	Часовня 1.
4019	Васильевское, д. вольн.	при колодцахъ.	35	29	13	26	36	
4020	Павлово, д. вольн.	при рудѣ Погожева.	35	29	7	30	37	
4021	Зеленое, д. вольн.	при р. Сидѣ.	36	30	27	64	69	Часовня 1.
4022	Михайлово, д. вольн.	при р. Сидѣ.	40	38	43	88	94	Часовня 1.
4023	Хотыничское, д. вольн.	при безымянномъ ручьѣ.	38	31	10	43	42	
4024	Косыно, д. вольн. и вл.	при безымянномъ ручьѣ.	45	33	16	44	47	
4025	Житово, д. вольн.	при р. Погожевѣ.	30	28	35	40	61	Часовня 1.
4026	Артюховское, д. вл.	при колодцахъ.	40	36	34	123	106	
4027	Думино, д. вл.	при безымянномъ ручьѣ.	40	36	34	109	122	
4028	Носино, д. вольн.	при ручьѣ Шитѣ.	45	36	18	39	37	Часовня 1.

91% - +

3995 Гарусово, сц. вл. при прудѣ и колодцахъ. 29 25

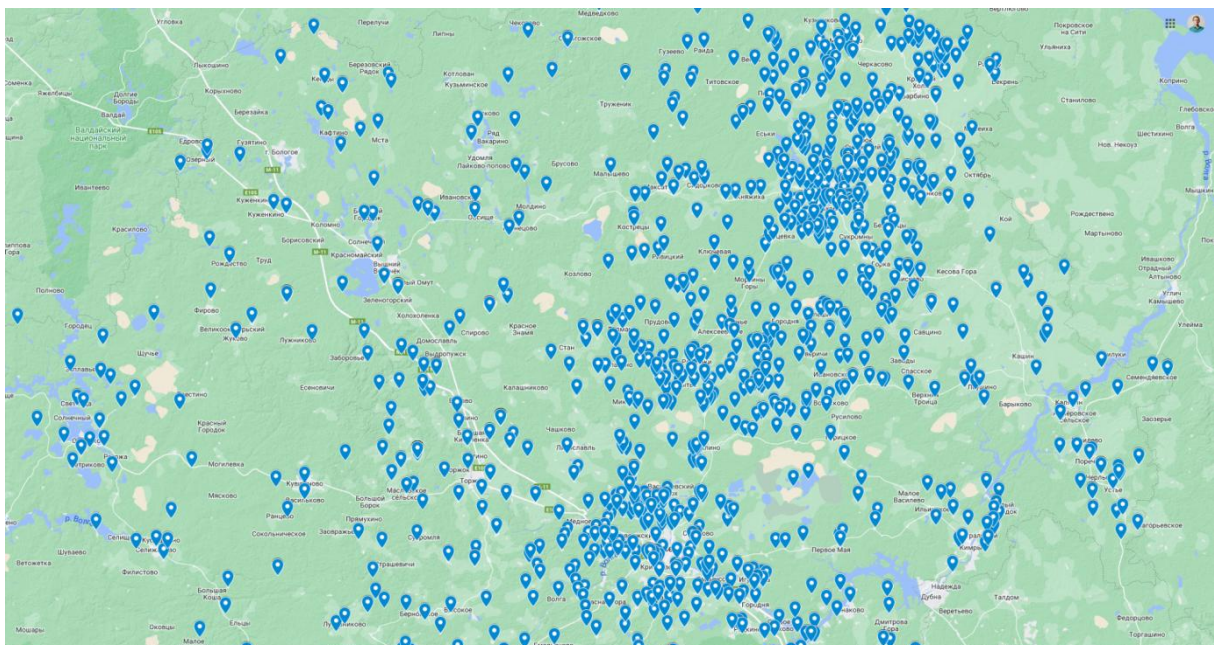
Приложение 2

Код программы на языке *Python* для автоматизированного получения координат населённых пунктов из списка

```
from geopy.geocoders import Nominatim
# Функция для получения координат по названию населённого пункта
def get_coordinates(place_name): geolocator = Nominatim(user_agent=
"geo_app") location = geolocator.geocode(place_name) if location: latitude =
location.latitude longitude = location.longitude return latitude, longitude else:
return None
# Чтение названий населённых пунктов из файла
with open('input.txt', 'r') as file: places = file.read().splitlines()
# Получение координат для каждого населённого пункта и сохранение в файл
with open('output.txt', 'w') as file: for place in places: coordinates =
get_coordinates(place) if coordinates: latitude, longitude = coordinates
file.write(f'{place}: {latitude}, {longitude}\n')
print(f'Координаты для {place} успешно сохранены.') else: file.write
(f'Не удалось найти координаты для {place}\n')
print(f'Координаты для {place} не найдены.') print('Программа завершена.')
```

Приложение 3

Визуализация населённых пунктов посредством веб-ГИС *Google Maps*



В. Н. Владимиров¹, Е. А. Брюханова², Н. В. Неженцева³,

¹ доктор исторических наук, профессор;

^{2,3} кандидат исторических наук, доцент;

^{1,2,3} Алтайский государственный университет, Барнаул, Россия

e-mail: ¹ vnapple@yandex.ru, ² brelenochka@mail.ru, ³ neshenzewan@mail.ru

ORCID: ¹ 0000-0002-0555-4999, ² 0000-0002-6974-5065, ³ 0000-0002-6700-7739

ИНФОРМАЦИОННАЯ СИСТЕМА ПО ИСТОРИЧЕСКОЙ СТАТИСТИКЕ АЛТАЯ: ОСОБЕННОСТИ ФОРМИРОВАНИЯ КОМПЛЕКСА РЕГИОНАЛЬНЫХ ДАННЫХ

Аннотация. Представлены особенности сбора и публикации статистических материалов по территории Алтая во второй половине XIX – начале XX в. Выявленные сведения позволили разработать архитектуру базы данных «Историческая статистика Алтая», которая позволяет аккумулировать данные об основных организаторах и участниках статистических обследований, способах публикации их результатов, выявить наиболее востребованные тематики статистических изданий, сформировать связи между источниками и публикациями. Создание информационной системы направлено на формирование комплекса региональной статистики Алтая и будет способствовать расширению информационных возможностей проведения исторических исследований в регионах России.

Ключевые слова: база данных, исторический источник, Алтай, XIX–XX вв.

Статистические материалы являются одними из наиболее востребованных исторических источников. Тем не менее выявление таких материалов является довольно трудоёмким процессом, особенно на уровне регионов. Для формирования комплекса региональной статистики требуется не только изучение истории статистической науки, структуры статистических органов и методов сбора статистических данных в целом, но и учёт различий в административно-территориальном делении в разные исторические периоды, влияющих в т. ч. на сбор и публикацию статистических данных.

Особенностью территории Алтая до конца XIX в. была двойная подчинённость в связи с существованием Алтайского (горного) округа, входившего в состав Томской губернии и одновременно находившегося в управлении Кабинетом Е. И. В. Статистическая отчётность округа подавалась в Кабинет Е. И. В., за исключением вопросов, входивших во всеподданнейший губернаторский годовой отчёт, статистические сведения собирались начальниками заводов и специалистами и являлись частью делопроизводственной документации. С созданием в 1883 г. Главного управления Алтайского (горного) округа и в 1894 г. статистического бюро при его земельной части статистические материалы по территории Алтая стали собираться не только в рамках обязательного учёта, но и по собст-

венным исследовательским программам. Так, за 1894–1917 гг. исследования Алтайского статбюро были опубликованы в многотомных изданиях по населению Горного Алтая (1900–1903 гг.), крестьянским и инородческим хозяйствам (1898–1901 гг.), а также арендного хозяйства (1896 г.) и мест водворения переселенцев (1894 г.) [1, с. 89–90].

В начале XX в. Переселенческое управление при Главном Управлении Землеустройства и Земледелия развернуло масштабную деятельность по сбору разнообразных данных по Томскому переселенческому району, включая территорию Алтая, которая рассматривалась как одно из наиболее привлекательных мест для переселения. Статистический комитет при Томском переселенческом бюро с 1905 г. представлял ежегодные обзоры, в 1915–1917 гг. издавал собственный журнал «Статистико-экономические бюллетени», а также публиковал результаты научно-статистических исследований, связанных с переселенческим делом, например, об экономическом положении переселенцев Барнаульского уезда (1913 г.), по кирпичному и кожевенному промыслам Барнаульского и Бийского уездов (1913 г.) и др.

Ещё одна особенность Томской губернии и Алтая обусловлена отсутствием земств в Сибири, роль которых довольно успешно исполнялась общественными организациями. Ярким примером является «Общество любителей исследования Алтая», реорганизованное в 1902 г. в Алтайский подотдел Западно-Сибирского отдела Русского географического общества и занимавшееся сбором разнообразных сведений о территории Алтая. Основным печатным органом Общества был «Алтайский сборник» (11 томов за 1894–1910 гг.), в котором были опубликованы результаты однодневной переписи г. Барнаула за 1895 г., данные по начальным школам Алтайского округа (1894 г.), материалы трёх подворных переписей (1897, 1901, 1905 гг.) переселенческого селения с. Чистюнька Барнаульской волости и др. Кроме того, сведения о населении и территории Алтая встречаются в изданиях Вольного экономического общества, Русского технического общества и др.

Ещё одним участником сбора статданных были научные круги, сотрудничавшие с государственными учреждениями и общественными организациями в области статистических обследований. Так, при Императорском Томском университете в начале XX в. действовал статистический семинарий, члены которого под руководством профессора М. Н. Соболева проводили обследования кустарных промыслов Барнаульского и Бийского уездов. Результаты обследования были опубликованы в «Трудах статистико-экономического отделения Юридического кабинета при Императорском Томском университете» в 1909 г. [2, с. 35].

С одной стороны, статистические материалы по территории Алтая собирались в соответствии с общегосударственной программой текущей статистики и специальных обследований, с другой, – имели свою специфику, обусловленную управлением частью территории Кабинетом Е. И. В., активной вовлечённостью региона в переселенческие процессы, а также значительным интересом к статистике общественных и научных организа-

ций. Эти особенности обеспечили довольно широкий спектр изучаемых и публикуемых сведений в ведомственных и общественных изданиях, но привели к разрозненности статистических данных.

С целью формирования комплекса региональной исторической статистики нами была создана информационная система «Историческая статистика Алтая» (URL: statalt.hist.asu.ru). Она разработана под управлением СУБД *MySQL* и состоит из справочников и четырёх основных модулей: хранилища, источники, метки и документы [3].

Справочники являются системообразующим элементом, отражают логику базы данных и позволяют выстраивать связи между остальными модулями. В настоящее время созданы справочники видов хранилищ, категорий источников и меток. Все справочники являются динамическими и могут дополняться.

Модуль «Хранилища» создан для систематизации мест хранения и представления статистических материалов. В настоящее время в справочник внесены следующие типы хранилищ: архив, музей, библиотека, учреждения статистики, электронный ресурс. Модуль «Хранилища» позволяет сформировать перечень мест хранения статистических материалов по Алтаю и получить представление об их доступности, в т. ч. в цифровом формате.

Основным разделом информационной системы является модуль «Источники», аккумулирующий все выявленные нами материалы по статистике Алтая. В базе данных в настоящее время представлены следующие категории источников:

- архивный фонд и архивное дело – фонды статистических органов, первичные документы статистических обследований, неопубликованные рукописи исследований и другие материалы;
- исторический документ – законодательные акты, инструкции, положения и другие документы, регулирующие проведение тех или иных статистических мероприятий, а также регламентирующие деятельность статистических органов;
- статистический источник – опубликованные статистические данные;
- издание и статья – публикации, основанные на статистических материалах.

Отдельный интерес представляет «Коллекция» – это опция информационной системы, позволяющая объединять несколько источников в группу по определённым признакам. Первоначально она создавалась для внесения сведений об архивных и музейных материалах, имеющих уровневую структуру. Но в процессе развития информационной системы появилась и была реализована идея создания номинативных (объединяющих однотипные или серийные издания) и тематических коллекций. Примером номинативной коллекции является коллекция «Обзоры Томской губернии», включающая серию соответствующих выпусков 1879–1912 гг., имеющих однотипное библиографическое описание. Тематическая коллекция представлена «Кустарными промыслами Томской губернии», куда

внесены исследования разных лет специалистов Томского губернского кустарного комитета, посвящённые промыслам Барнаульского и Бийского уездов.

Ещё одной важной функцией базы данных является возможность «связывания» источников. Функция «Связи» актуальна для изданий, статей и статистических источников и представляет собой систему ссылок, во-первых, на те публикации и архивные материалы, которые послужили основой для данного источника, а во-вторых, на те, в которых упоминается сам источник.

Таким образом, модуль «Источники» направлен на составление списка опубликованных и архивных источников, формирование библиографических связей и их последующий анализ.

Несмотря на то, что форма ввода данных источника является довольно подробной, для создания качественного многоаспектного и функционального поиска в информационной системе этого недостаточно – необходим некий набор характеристик, позволяющий унифицировано описывать источники. Для этого был создан модуль «Метки». На этапе разработки были созданы следующие категории меток: административно-территориальная единица, издатель/исследователь, исторический период и статистическая тема. Каждый источник сопровождается теми метками, которые характеризуют его наиболее полно.

Следующим разделом информационной системы является модуль «Документы», представляющий собой отдельный блок для размещения датасетов – массивов статистических данных, переведённых в цифровой вид и готовых для дальнейшей обработки. Датасеты создаются по наиболее востребованным, а также малоизученным или труднодоступным статистическим источникам. В настоящее время в электронный формат переведены статистические таблицы опубликованных материалов переписей населения 1897, 1920 и 1937 гг. по городам и некоторым уездам/округам Алтая. Перевод статистических данных в электронный формат позволил расширить возможности работы с источником и провести источниковедческий анализ. Все датасеты будут доступны пользователям на сайте, в т. ч. для скачивания, и могут быть использованы как в научно-исследовательской, так и в образовательной деятельности.

В целом создание регионального информационного ресурса по комплексу статистических материалов по истории Алтая XIX – первой половины XX в., включающего не только статистические издания, но и материалы со статистическими данными из различных источников и хранилищ, направлено на расширение информационных возможностей проведения исторических исследований в регионах России.

Работа выполнена при поддержке Российского научного фонда, проект № 22-28-20287 «Историческая статистика Алтая онлайн: репрезентация статистических источников в цифровом пространстве».

Список источников

1. Скопа В. А. История развития статистики и статистических учреждений Томской губернии в 1835–1919 гг. / В. А. Скопа. Барнаул: БГПУ, 2009. 209 с.
2. Могильницкая К. И. Очерки становления высшего образования в Сибири / К. И. Могильницкая. Томск: НТЛ, 1997. 52 с.
3. Владимиров В. Н. Информационная система «Историческая статистика Алтая» / В. Н. Владимиров, Е. А. Брюханова, Н. В. Неженцева // *Quaestio Rossica*. 2022. Т. 10. № 5. С. 1 884–1 896.

В. М. Дамдынчап,
кандидат исторических наук, доцент
Тувинский государственный университет, Кызыл, Россия
e-mail: damdvera@yandex.ru
ORCID: 0000-0001-8477-7578

ПРОСТРАНСТВЕННАЯ РЕПРЕЗЕНТАЦИЯ И АНАЛИЗ СОЦИОДЕМОГРАФИЧЕСКОГО ПРОСТРАНСТВА ТЕРРИТОРИИ ТУВЫ СОВЕТСКОГО ПЕРИОДА НА ОСНОВЕ ГИС-ТЕХНОЛОГИЙ

Аннотация. В статье проанализирована динамика демографических процессов в советской Туве и их взаимосвязи с социально-экономическими изменениями в рассматриваемый период на основе статистических данных Всесоюзных переписей населения 1959, 1970, 1979, 1989 гг. Для анализа были составлены сводные таблицы численности населения, мужчин и женщин, по национальному составу по районам Тувы. Были созданы административные карты Тувинской автономной области до 1961 г., Тувинской автономной республики до 1991 г. в растровом формате. Работа над проектом осуществлена в приложении *ARGIS* и проведён пространственный анализ расселения.

Ключевые слова: демографические показатели, геоинформационные технологии, расселение населения, советская модернизация, национальный состав, половой состав

Одним из аспектов исследования социального развития и города, и села является изучение расселения, представляющего собой процесс распределения и перераспределения населения по территории его обитания. Понятие «расселение населения» включает и сеть поселений, и те функциональные территориальные взаимосвязи в этой сети, которые позволяют рассматривать её как определённую систему передвижения людей (переселение, миграция).

В советский период в Туве сложился аграрно-промышленный комплекс. В 1970-е гг. наблюдался рост городского населения, увеличение количества посёлков городского типа, был создан пгт Каа-Хем в 1974 г. Сельское население сосредоточилось в районных центрах, уменьшилось количество жителей небольших сёл в республике.

Рассмотрим пространственный каркас расселения в советский период с учётом половой структуры и национального состава. Расчёт произведён с помощью программы *QGIS* на основе административных карт советского периода. Количество административно-территориальных районов варьировалась от 13 до 16. На основе созданных карт и составленных таблиц был осуществлён пространственный анализ демографических показателей.

В составленных таблицах население разделено на городских и сельских жителей. Всего в советский период в Туве было пять городов: Кызыл, Туран, Чадан, Шагонар, Ак-Довурак. В 1956 г. был основан пгт Хову-Аксы – рабочий посёлок при комбинате «Тувакобальт». В 1979 г. был основан пгт Каа-Хем – райцентр Кызылского района.

Численный состав населения в советский период растёт, гендерный состав по районам примерно одинаков. В советский период с развитием различных отраслей хозяйства увеличилось количество поселений, в 1989 г. было 5 городов, 3 пгт и 450 сельских населённых пунктов [1, с. 39].

Анализ показал, что в советский период происходит увеличение количества городских жителей: в 1959 г. – 33 % городского и 67 % сельского населения, в 1989 г. – 47 % городского и 53 % сельского населения. Основная отрасль экономики – отгонное скотоводство, но постепенно происходят изменения в структурном составе населения, животноводов становится меньше по сравнению с рабочими и служащими.

По национальному составу в таблицах выделены четыре группы: тувинцы, русские, украинцы и прочие национальности. Первые три составляли численное превосходство, хотя регион был многонациональным: в 1959 г. в Туве проживали представители более 20 национальностей, в 1970 и 1979 гг. – уже более 70, в 1989 г. – более 80 [1, с. 41].

Данные, приведённые на картограмме (рис. 1), показывают, что распределение населения по национальностям на территории Тувы было следующим: русские, украинцы жили в основном в городских поселениях, поскольку основная масса тувинцев занята в животноводстве, – в Кызыле как административной столице и Ак-Довураке как промышленном центре. В 1959 г. русского населения было больше в других городах – в Туране, Шагонаре. Постепенно количество тувинцев, проживающих в городах, увеличивалось, обусловлено это развитием системы образования, увеличением количества образованных жителей.

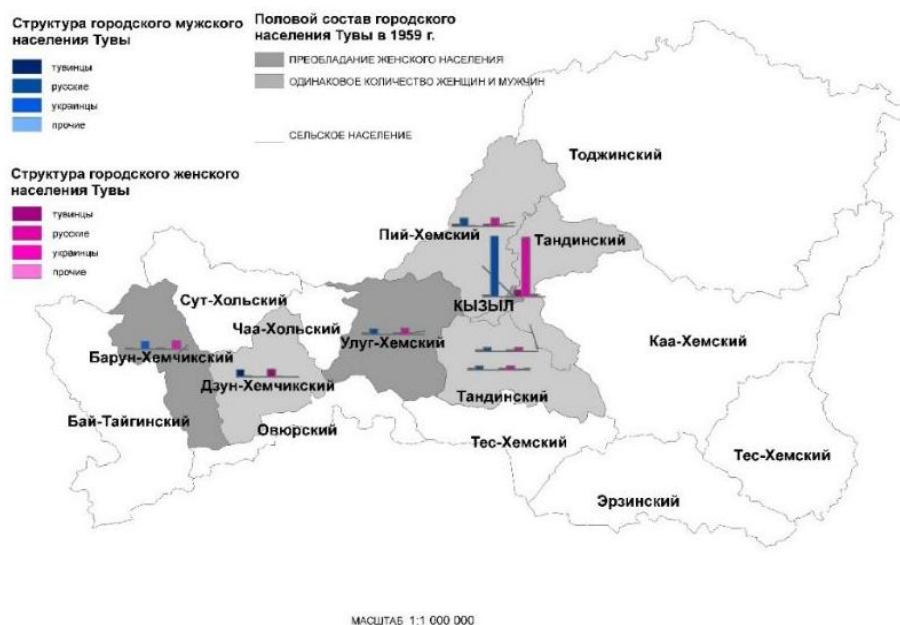


Рис. 1. Структура городского населения в 1959 г.

На картограмме показано, что с 1944 по 1961 гг. существенно вырос г. Кызыл, будучи столицей Тувы за счёт как внутренней миграции, так и внешней. Здесь развернулось строительство промышленных предприятий. В целом количество русского населения составляло 79,5 % городских жителей [1, с. 48], большая часть приходилась на Кызыл, 70 % русского населения представляли приезжие специалисты, которые пополняли рабочие кадры. В других городских поселениях численность населения была примерно одинакова, чуть больше в Туране – 5 646 чел. Тувинцев в Кызыле в 1959 г. было всего 483 чел. По гендерному составу большой диспропорции не наблюдалось, как было в целом в СССР. В г. Шагонаре и пгт Кызыл-Мажалык преобладало женское население – 51,5 и 50,9 % соответственно [1, с. 40]. Эти два поселения были административными центрами и колхозными центрами животноводческого направления.

Распределение сельского населения Тувинской автономной области рассмотрим на картограмме (рис. 2). Из неё видно, что население росло за счёт естественного прироста и миграции специалистов из других регионов. Сельских жителей в Туве было в два раза больше, чем городских. Активно увеличивалось население районов, где первыми появились колхозы и проживало русское население: Каа-Хемский, Тандынский, Улуг-Хемский. В советский период русское население появилось во всех районах области, хотя в период ТНР они были сосредоточены только в четырёх. Интересно, что в Барун-Хемчикский район для работы в новых колхозах, предприятиях приезжали специалисты с Украины. В этот период началось организационное оформление колхозных центров, административных центров, шло строительство новых сёл в районе. В труднодоступных районах – Овюрском, Эрзинском и Монгун-Тайгинском – численность русского населения и представителей других национальностей была существенно меньше.

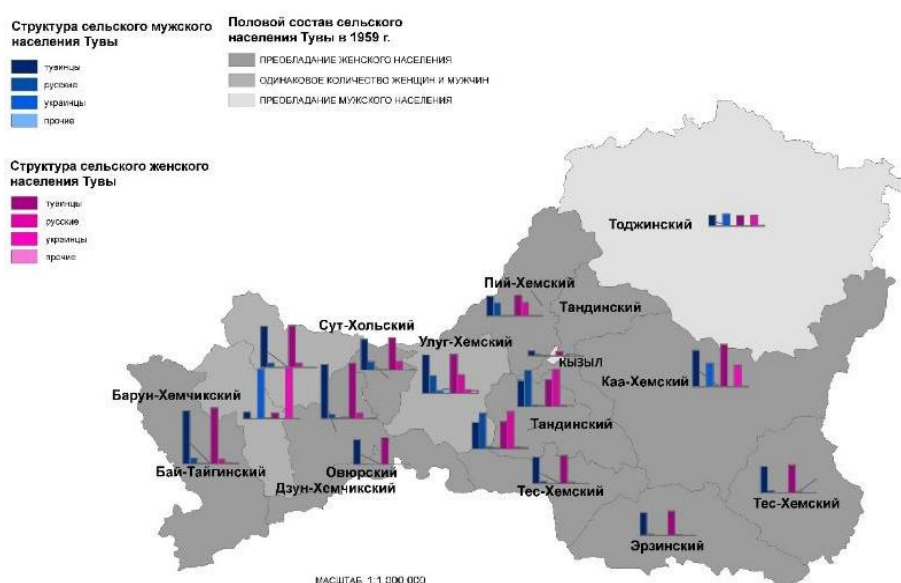


Рис. 2. Структура сельского населения в 1959 г.

Как показывают данные картограммы по половому составу населения, больших диспропорций между мужским и женским населением не наблюдается. В трёх районах – Барун-Хемчикском, Сут-Хольском и Улуг-Хемском – равное количество (по 50 %) мужчин и женщин. В период Великой Отечественной войны тувинские добровольцы были в большинстве своём выходцами из районов, где проживали русские, выбирали тех, кто понимал русский язык и говорил на нём. Поэтому здесь не было отрицательных демографических последствий. А эти районы – западные, здесь русского населения было мало. В двух первых районах развито отгонное скотоводство, в последнем развивалось земледелие.

На следующей картограмме (рис. 3) мы видим изменения в распределении городского населения. Через 20 лет количество городского населения увеличилось в 1,5 раза. Лидирующее положение сохраняется за Кызылом (51 683 чел.), затем идёт г. Ак-Довурак и пгт Кызыл-Мажалык (14 036 чел.), г. Чадан (7 589 чел.), пгт Хову-Аксы (4 632 чел.), г. Туран (4 539 чел.), г. Шагонар (4 512 чел.) (см. прил.) [3]. Национальный состав меняется в сторону увеличения количества тувинцев в городских поселениях. За это время подготовлены уже национальные кадры, получившие среднее профессиональное и высшее образование. Так, в Кызыле количество тувинцев в 1970 г. выросло в 13 раз по сравнению с 1959 г. Русское население представлено во всех районах, меньше всего было в Овюрском районе – 74 чел. Много украинцев было в Шагонаре, Чадане, где они были заняты в строительстве и добывающей промышленности. В г. Ак-Довураке на комбинате «Туваасбест» работало только 25 % тувинцев [2]. Гендерный состав населения изменился в районах, где развивалась добывающая промышленность, – Барун-Хемчикском (г. Ак-Довурак), Дзун-Хемчикском (г. Чадан), Тандынском (пгт Хову-Аксы). На горно-обогатительных комбинатах «Туваасбест», «Тувакобальт» и Чаданском угольном разрезе была потребность в мужском труде. В городах Туране и Шагонаре сохраняется небольшое преобладание женщин.

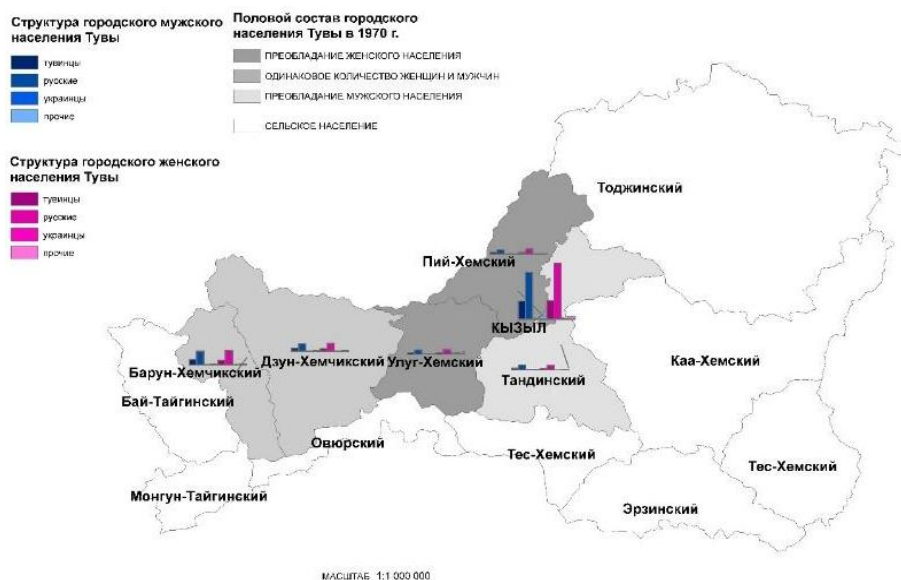


Рис. 3. Структура городского населения в 1970 г.

На картограмме (рис. 4) представлены изменения в расселении сельского населения. К 1970 г. количество сельчан – 143 873 чел. – в 1,7 раза больше городского. В два раза увеличилось население двух районов: Дзун-Хемчикского и Улуг-Хемского, поскольку была проведена административная реформа. В Дзун-Хемчикском районе увеличение произошло за счёт присоединения территории Сут-Хольского района, в Улуг-Хемском – за счёт Чаа-Хольского района. Население Тандынского района увеличилось за счёт строительства комбината «Тувакобальт», для обеспечения его работы были созданы укрупнённые совхозы, куда были направлены специалисты из других районов и регионов Сибири.

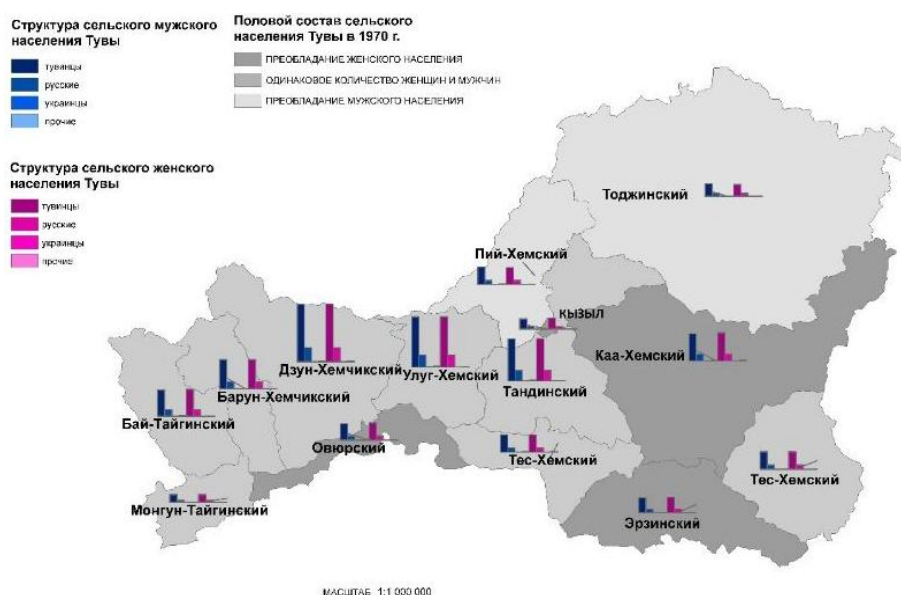


Рис. 4. Структура сельского населения в 1970 г.

По национальному составу во всех районах стали преобладать тувинцы, что обусловлено ростом образованности, развитием непродовольственных отраслей в сельской местности.

Заметно увеличилось количество русских в труднодоступных районах – Овьюрском, Монгун-Тайгинском и Эрзинском. Эти районы приграничные, здесь были образованы стационарные пограничные отряды. В районных центрах развивались образовательные, культурные и медицинские учреждения, где большинство специалистов составляли русские и представители других национальностей.

В десяти районах наблюдается почти одинаковое количество мужчин и женщин, что обусловлено ростом благосостояния, системы здравоохранения и образования. Только Тоджинский район с преобладанием мужского населения, поскольку здесь были созданы рыболовецкие хозяйства, в которых работают в основном мужчины.

Данные следующей картограммы (рис. 5) показывают, что к 1979 г. структура городского населения по половому составу изменилась, увеличилось количество женщин (разница с количеством мужчин – примерно 2 %). Существенное изменение произошло в пгт Хову-Аксы, мужское на-

селение уменьшилось с 56 до 48 %, а женское увеличилось с 46 до 52 %. Скорее, такое положение связано не с увеличением женщин на основном комбинате «Тувакобальт», а с ростом посёлка. К 1979 г. здесь были введены в строй районная больница на 120 мест, вторая школа, второй ясли-сад, две общественные столовые на 110 мест, организован свой строительный трест «Тувакобальтстрой» [5, с. 195]. Во всех указанных предприятиях и организациях преимущественное количество женщин.

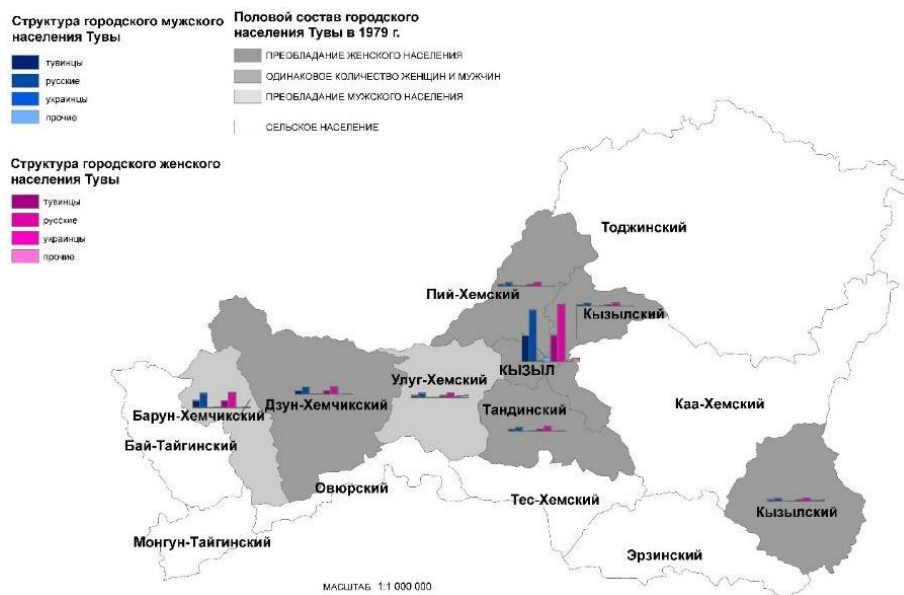


Рис. 5. Структура городского населения в 1979 г.

Показатели картограммы (рис. 6) свидетельствуют о том, что количество сельского населения в Тувинской АССР ненамного увеличилось по сравнению с городским населением – всего на 11 323 чел. [4]. Это период наступления демографической стагнации. Период больших строек завершён, миграционный поток уменьшился, и рост количества населения в районах происходил за счёт естественного прироста.

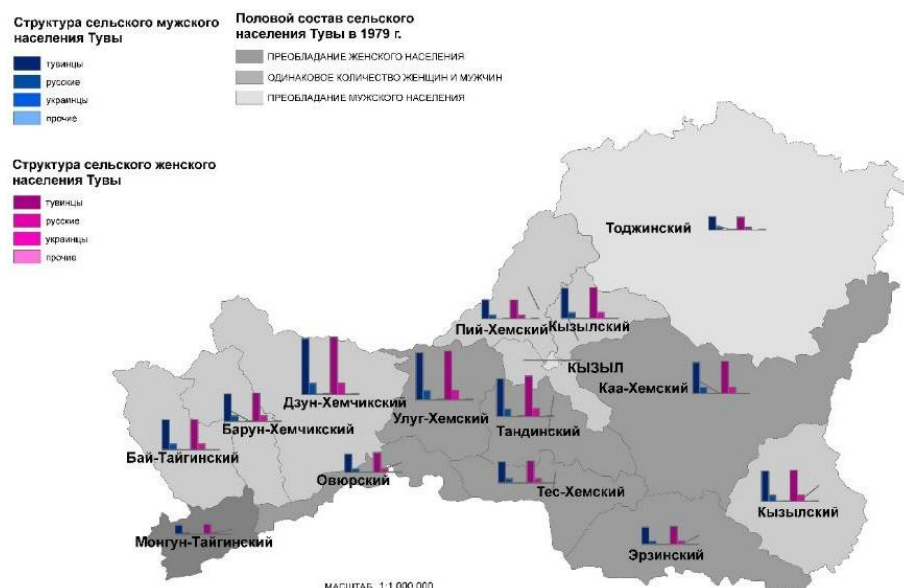


Рис. 6. Структура сельского населения в 1979 г.

В трёх районах – Дзун-Хемчикском, Тандынском и Улуг-Хемском – сохраняется двойное превышение населения, чем в других районах, но в 1970-е гг. количество и здесь уменьшилось. Сказались застойные изменения в регионе в целом. Многие уехали на большие стройки того времени – например, строительство БАМа. Из всех районов уменьшение произошло в одном – Барун-Хемчикском – вероятно, из-за смены работы в сельскохозяйственной отрасли на промышленную работу на комбинате «Туваасбест» г. Ак-Довурак. Увеличение женского населения во многих районах, где раньше мужчин и женщин было поровну, связано с процессами уменьшения количества мужчин в силу социальных причин: увлечения алкогольными напитками, возрастания смертности среди мужчин.

Следующая картограмма (рис. 7) показывает, что в распределении населения больших изменений не произошло. Последнее десятилетие советского периода самое кризисное. Темпы развития замедлились, но всё же продолжалось строительство новых микрорайонов в Кызыле, Ак-Довураке, Чадане. Численность городского населения республики увеличилась лишь в 1,3 раза. В этот период урбанизация достигла максимального уровня. Росла внутренняя миграция из села в город. Почти вдвое увеличилось население г. Шагонара, обусловлено это переселением из старого города и посёлков, которые находились в зоне затопления Саяно-Шушенской ГЭС. По плану переселения жители сёл должны были переезжать в новые сёла, а жители старого Шагонара – в новый город. В 1982 г. началось постепенное заселение г. Шагонара, в город переселилась и часть жителей местечка Кок-Чыраа [9, с. 126–135]. Увеличилось и население пгт Каа-Хем – в 1,8 раза. Здесь была построена птицефабрика, сданы новые двухквартирные дома для жителей молочно-товарной фермы, сюда шёл миграционный поток из сёл республики, привлекала близость города.

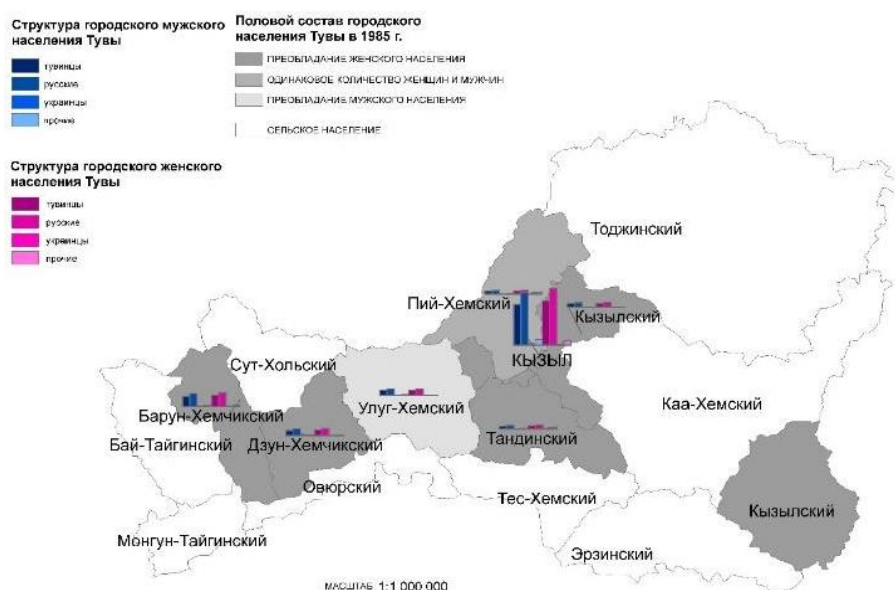


Рис. 7. Структура городского населения в 1989 г.

Структура сельского населения, как показывает картограмма (рис. 8), постепенно менялась. Больших диспропорций не наблюдается. Русские представлены во всех районах, но постепенно их количество уменьшается, преимущество сохраняется только в Кызыле. В ходе административно-территориальных реформ было образовано 15 районов вместо 14. Распределение населения в западной и центральной части Тувы примерно одинаковое.

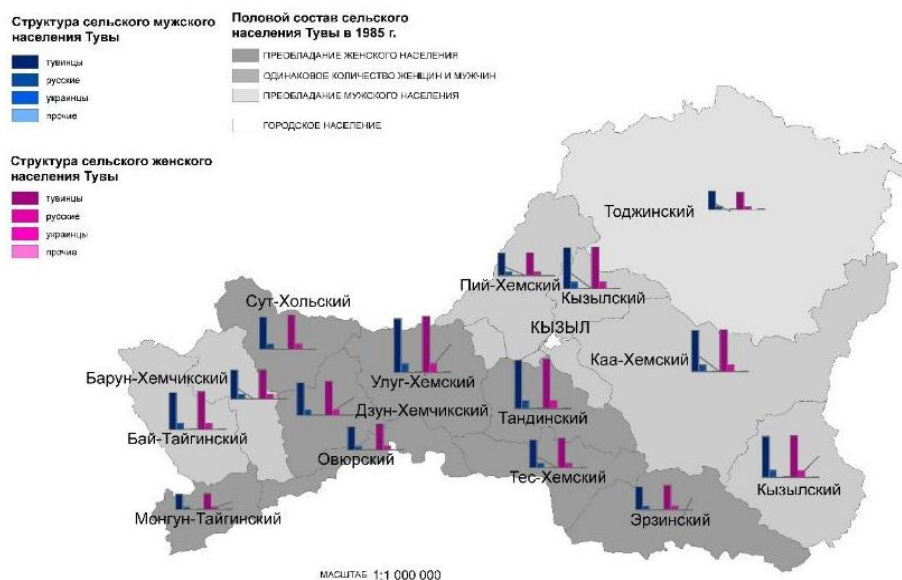


Рис. 8. Структура сельского населения в 1989 г.

В центральных и южных районах Тувы женское население увеличивается, что стало следствием кризисных социальных явлений, таких как рост алкоголизма, преступлений среди мужчин. Из всех районов преобладание количества мужчин сохранялось только в Тоджинском.

Таким образом, пространственное распределение населения свидетельствует о том, что в советский период экономических преобразований все районы Тувы были активно освоены. Были созданы экономические зоны по основным отраслям производства. В период перехода на оседлость в западных районах Тувы создавались колхозы и появилось много приезжих специалистов сельскохозяйственного направления. Постепенно с увеличением численности тувинцев соотношение было в пользу коренного населения. В Тандынском, Улуг-Хемском, Пий-Хемском, Кызылском районах русское население преобладает в 1959 г.; к 1989 г. они уже не составляют большинства. В советское время проводилась ещё и политика переселения тувинцев из западных районов в районы, где численность тувинцев была небольшой – Тоджинский, Каа-Хемский.

Русские и представители других национальностей представлены во всех районах, в основном там, где развивалась промышленность. В сельской местности их количество постепенно сокращается, поскольку отгонное животноводство сохранялось в Туве и им занимались преимущественно тувинцы. В Тоджинском районе, где тувинцы занимались отгонным оленеводством, проживало достаточно большое количество русского

населения – представители староверов, приезжие, работавшие в леспромысле и рыболовецком хозяйстве. В западных и южных районах Тувы тувинцы занимались традиционным хозяйством, переход на оседлость изменил и их расселение – люди перебирались в районные центры.

Таким образом, приведённое на картограммах распределение населения позволяет увидеть, что активно развивались районы с промышленными предприятиями, где была создана вся инфраструктура для более благоприятного развития. Количество населения постоянно увеличивалось, менялся образ жизни тувинцев под влиянием как объективных, так и субъективных факторов.

Тем не менее оставались такие районы, как Каа-Хемский, Тоджинский, Монгун-Тайгинский, в которых сохранялось традиционное отгонное животноводство и они не были привлекательны для населения. Поэтому это районы с небольшим количеством населения.

Список источников

1. 60 лет вхождения Республики Тыва в состав РФ. Кызыл, 2004. 50 с.
2. Ак-Довуракский асбестовый – в строй! // Тувинская правда. 1964. № 210 (5941).
3. Всесоюзная перепись населения 1970 г. URL: demoscope.ru/weekly/ssp/rus70_reg2.php.
4. Всесоюзная перепись населения 1979 г. Численность наличного населения РСФСР. URL: demoscope.ru/weekly/ssp/rus79_reg1.php.
5. История Тувы: в 3 т. Т. 3 / под общ. ред. В. А. Ламина. Новосибирск: Наука, 2016. 455 с.
6. Итоги Всесоюзной переписи населения 1959 г. РСФСР. М., 1963; Итоги Всесоюзной переписи населения 1970 г. Т. 4. М., 1973.
7. Народное хозяйство Тувинской АССР. Кызыл, 1981. 216 с.
8. Народное хозяйство Тувинской АССР. Кызыл, 1990. 233 с.
9. Саая С. М. Из истории строительства новых населённых пунктов Улуг-Хемского района в связи с затоплением территории в 1973–1985 гг. (по материалам Национального архива Республики Тыва) / С. М. Саая // Архив: история и современность: матер. межрег. НПК (Кызыл, 2020) / отв. ред. Б. В. Мунге; ред.-сост.: М. М.-Б. Харунова, Р. Н. Базыр; сост. Д. Д. Монгуш. Кызыл: Аныяк, 2020.
10. Юбилейный статистический сборник к 100-летию единения России и Тувы: стат. сб. Кызыл: Тывастат, 2014. 208 с.

Е. В. Данилов,
аспирант

Московский государственный университет им. М. В. Ломоносова, Москва, Россия
e-mail: random.pc.user@gmail.com

ОПЫТ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ АНАЛИЗА ПРОЦЕССОВ ОБРАЩАЕМОСТИ ЗА МЕДИЦИНСКОЙ ПОМОЩЬЮ В РОССИИ В КОНЦЕ XIX ВЕКА

Аннотация. В тексте исследуются возможности применения статистических методов при изучении феномена обращаемости за медицинской помощью в Российской империи в конце XIX в. Показано, что выбранные автором факторы, такие как численность врачей и фельдшеров, грамотность и земский статус в некоторой степени оказывают влияние на распределение обращаемости. Кроме того, данные на уровне губерний значительно более чувствительны, чем данные на уровне уездов.

Ключевые слова: *Российская империя, здравоохранение, земство, медицина, врач*

Обращаемость населения за медицинской помощью – это важное и заслуживающее исследовательского интереса явление. Его можно рассматривать как некоторый статистический показатель того, как часто люди обращаются за медицинской помощью, что является совершенно верной интерпретацией этого явления. Но это определение лишь частично охватывает содержательную сущность рассматриваемого явления. Обращаемость за медицинской помощью зависит от огромного количества факторов – от социальных до экономических и политических; от сугубо индивидуальных – до факторов, характерных для социальных групп и социумов [1]. Изучение этого феномена может пролить свет на множество аспектов того, как устроено общество: от социальных и культурных предрассудков до особенностей социальной политики тех или иных государств и обществ.

Одним из методов изучения обращаемости за медицинской помощью является статистический анализ. В данном тексте предложена попытка отразить опыт исследования процессов, связанных с обращаемостью населения за медицинской помощью в Российской Империи в конце XIX – начале XX в. Методы статистического анализа за последние несколько десятилетий все чаще используются в исторических исследованиях, что позволяет исследователям взглянуть на изучаемые проблемы с применением несколько иной оптики. Так, например, исследования Пола Грегори предложили иной, более позитивный взгляд на процессы экономического развития в Российской империи накануне Первой мировой войны [2].

В данном исследовании была произведена попытка исследовать процессы обращаемости за медицинской помощью в России на рубеже XIX–XX вв. Исследование затрагивает период 1892–1895 гг.; во многом

в тексте предлагается сравнить результаты, полученные в итоге анализа обращаемости для этих срезов. Анализ и сравнение производятся по таким характеристикам, как статистический показатель обращаемости, а также влияние на него таких факторов, как грамотность, земский статус и численность врачей и фельдшеров – для последнего применяются методы регрессионного и корреляционного анализа. Выбор этих характеристик связан с тем, что они, во-первых, доступны для исследователей, а во-вторых, эти факторы так или иначе отражают разные аспекты того, из чего состоит феномен обращаемости. Так, численность врачей и фельдшеров во многом объясняют принимающую способность медицинских институтов, т. е. характеризуют состояние медицинской инфраструктуры и то, насколько она приспособлена для работы с пациентами. Грамотность – намного более сложная характеристика, она относится уже не к медицинской части, а к части населения – субъекта, который обращается за медицинской помощью; в исследовательской литературе достаточно часто поднимается вопрос о влиянии грамотности на то, как люди воспринимают изменчивость своего здоровья и какие действия они (не) предпринимают для его поддержания. Фактор земств в данном случае является не менее важной характеристикой в силу следующих причин: во-первых, земства в основном занимались вопросами местного благоустройства, в т. ч. касающихся общественного здоровья (строительство лечебниц, найм врачей, поддержание гигиены); однако влияние земств не ограничивалось только на организационном аспекте. Земство – это местное самоуправление, это в некотором смысле комьюнити, которое так или иначе способствует выработке определённого коллективного подхода в отношении целого ряда социальных вопросов, в т. ч. касающихся здоровья и медицины (влияние коллективных факторов относительно моделей здорового образа жизни и обращаемости за медицинской помощью также дебатировалась в работах по социальной медицине).

Т. е., даже при включении в работу относительно небольшого количества исследуемых параметров появляется достаточно непростой интерфейс обращаемости за медицинской помощью. Некоторые из них характеризуют сторону медицинских институтов, некоторые – обращающихся за медицинской помощью; иные так или иначе можно определить как к первым, так и ко вторым. Далее предлагается попытка изложить то, как эти факторы объясняют обращаемость за медицинской помощью в Российской империи на рубеже XIX–XX вв.

Первый этап работы заключался в сборе данных. К сожалению, данные по обращаемости населения за медицинской помощью фактически не оцифрованы и не доступны в виде онлайн-датасетов, поэтому сначала были произведены ручной сбор данных и преобразование их в форме электронных таблиц. Источниками информации послужили отчеты Медицинского департамента и Управления главного врачебного инспектора МВД, а также материалы Первой всеобщей переписи населения 1897 г., составленные Н. А. Тройницким. Затем последовал этап статистического анализа.

Срезы 1892 и 1895 гг. являются особенными в силу того, что для них есть в наличии данные по обращаемости за медицинской помощью не только на уровне губерний (что характерно для последующих срезов), но и на уровне отдельных уездов. В связи с этим было принято решение не только сравнить результаты анализа этих срезов как на погубернском, так и на поуездном уровне. Разница значительная, т. к. в первом случае число исследуемых объектов (губерний) составляет 60, а во втором (уездов) – 587. Однако также следует заметить, что различие заключается не только в количестве исследуемых объектов. Если для погубернского уровня в наличии данные по обращаемости как во всей губернии, так и в городах и сельской местности, то для поуездного уровня такого различия нет. В таком случае выходит, что исследование также поможет выявить, в какой степени исследуемые факторы (грамотность, земский статус, численность врачей и фельдшеров) одинаково объясняют распределение обращаемости в разных регионах Европейской России.

Результаты исследования следующие. Первый этап заключался в исследовании распределения обращаемости за медицинской помощью в Европейской России на поуездном уровне в 1892 и 1895 гг. Результаты таковы, что наилучшим образом из выбранных параметров это распределение объясняет комбинация факторов численности врачей и фельдшеров и медицинского статуса – в пределах 0,24 и 0,31 для 1892 и 1895 гг. соответственно. Более того, мы видим, что наиважнейшим фактором является фактор земского статуса – именно он оказывает наибольшее влияние на распределение исследуемого параметра.

Второй этап заключался в исследовании распределения обращаемости на погубернском уровне. Основное отличие этого этапа заключалось в том, что в нём использовались данные, разделённые не по отдельным уездам, но по принципу «обращения в городах / сельской местности» в пределах целой губернии. Результаты этого этапа значительно отличались результатов предыдущего этапа. Так, в целом уровень объяснения распределения обращаемости в целом стал выше (пиковый R^2 – 0,47 для комбинации факторов численности врачей и фельдшеров и земского статуса).

В целом получается, что в силу разделения данных погубернского уровня по принципу «город / сельская местность» эти данные гораздо более сенситивны для анализа в выбранной комбинации факторов, нежели данные поуездного уровня. Это может говорить о том, что принципы оказания медицинской помощи значительно отличались на уровне городов и провинции. В итоге выбранные факторы в некоторой степени объясняют распределение обращаемости за медицинской помощью в конце XIX в. Впрочем, в начале следующего века степень этого объяснения возрастает ещё больше, как и в целом численные показатели обращаемости [3], что говорит о том, что интенсивность развития медицинских институтов и их взаимодействия с населением страны было как минимум не в нисходящей динамике.

Таблица 1

Результаты регрессионного анализа обращаемости в Европейской России,
поуездный уровень. 1892, 1895 гг.

Показатель	1892			1895		
	1	2	3	4	5	6
Грамотность, %	0,126*** (0,042)			0,232*** (0,041)		
Городское население, %		0,170*** (0,042)			0,235*** (0,041)	
Медицинские сотрудники			0,282*** (0,041)			0,414*** (0,038)
Земства (1 – есть, 0 – нет)	0,404*** (0,042)	0,404*** (0,042)	0,424*** (0,041)	0,382*** (0,041)	0,381*** (0,041)	0,421*** (0,038)
Число объектов	439	439	439	469	469	469
R^2	0,17	0,18	0,24	0,19	0,19	0,31

Примечания. Зависимая переменная – обращаемость за медицинской помощью в уездах Европейской России. В скобках приведены стандартные ошибки коэффициентов.

*** $p < 0,01$, ** $p < 0,05$, * $p < 0,1$.

Таблица 2

Результаты регрессионного анализа обращаемости в Европейской России,
погубернский уровень. 1895 г.

Показатель	Город		Сельская местность	
	1	2	3	4
Грамотность, %	0,292*** (0,142)		0,070 (0,148)	
Медицинские сотрудники		0,453*** (0,138)		0,446*** (0,128)
Земства (1 – есть, 0 – нет)	0,409*** (0,142)	0,322*** (0,138)	0,507*** (0,148)	0,736*** (0,128)
Число объектов	39	39	39	39
R^2	0,27	0,37	0,20	0,47

Примечания. Зависимая переменная – обращаемость за медицинской помощью в уездах Европейской России. В скобках приведены стандартные ошибки коэффициентов.

*** $p < 0,01$, ** $p < 0,05$, * $p < 0,1$.

Список источников

1. Mackian S. A Review of Health Seeking Behaviour: Problems and Prospects. Health Systems Development / S. Mackian. Manchester: University of Manchester, 2002.
2. Gregory P. R. Russian National Income, 1885–1913 / P. R. Gregory. Cambridge University Press, 1982.
3. Данилов Е. В. Динамика и факторы обращаемости населения за медицинской помощью в Европейской России начала XX в. / Е. В. Данилов // Историческая информатика. 2022. № 4. С. 65–83. DOI: 10.7256/2585-7797.2022.4.39251.

С. А. Саломатина,

кандидат исторических наук

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

e-mail: ssalomatina@hist.msu.ru

ORCID: 0000-0003-0748-6229

РАССТОЯНИЯ МЕЖДУ РЕГИОНАЛЬНЫМИ РЫНКАМИ РОССИЙСКОЙ ИМПЕРИИ ВО ВТОРОЙ ПОЛОВИНЕ XIX ВЕКА: СЕТЕВОЙ И СТАТИСТИЧЕСКИЙ АНАЛИЗ

Аннотация. В исследовании анализируется статистика коммерческих переводов между отделениями Государственного банка Российской империи за 1868, 1878, 1888 и 1898 гг. Этот денежный оборот отражает оборот товаров и услуг между регионами. В центре внимания находится малоизвестное явление 1890-х гг. – заметное увеличение товарно-денежных оборотов в пределах ближних расстояний (примерно в радиусе 400 км от конкретного центра), тогда как в предыдущие десятилетия «ближние связи» были намного слабее. Это явление анализируется методами сетевого и статистического анализа, которые подтверждают, что топология товарно-денежных потоков 1890-х гг. резко отличается от предыдущих десятилетий. Причины этого бурного развития на местах нуждаются в дальнейшем исследовании.

Ключевые слова: *банковские переводы, межрегиональные платежи, региональные финансовые рынки, сетевой анализ, геоинформационный анализ, Российская империя*

Это исследование посвящено малоизвестному аспекту регионального развития Российской империи 1890-х гг. – резкому усилению экономических связей на близких расстояниях на региональных рынках (в радиусе 300–400 км от центра региона). В предыдущие десятилетия, наоборот, «ближние связи» были намного менее заметны. Эта динамика свидетельствует о важных изменениях на региональном уровне, что расширяет наши представления о глубине экономического развития страны.

Все расчёты выполнены по статистике коммерческих переводов между отделениями Государственного банка Российской империи (Госбанк) для 1868, 1878, 1888 и 1898 гг. [1–3; 4, с. 57–81]. В тот период Госбанк был крупнейшим коммерческим банком страны с самой обширной сетью отделений (47 в 1868 г. и 122 в 1898 г.), поэтому переводы по поручениям клиентов между отделениями отражают денежные потоки на большей части территории страны. Поскольку денежный поток противонаправлен потоку товаров и услуг, по этим данным можно оценивать именно товарно-денежные потоки.

Это исследование является продолжением статьи, в которой на географических картах показано, что топология платёжной сети 1898 г. значительно отличается от более ранних временных срезов [5]. Для иллюстрации на рис. 1 приводятся крупнейшие потоки по европейской части страны для 1878 г., а на рис. 2 – для 1898 г.

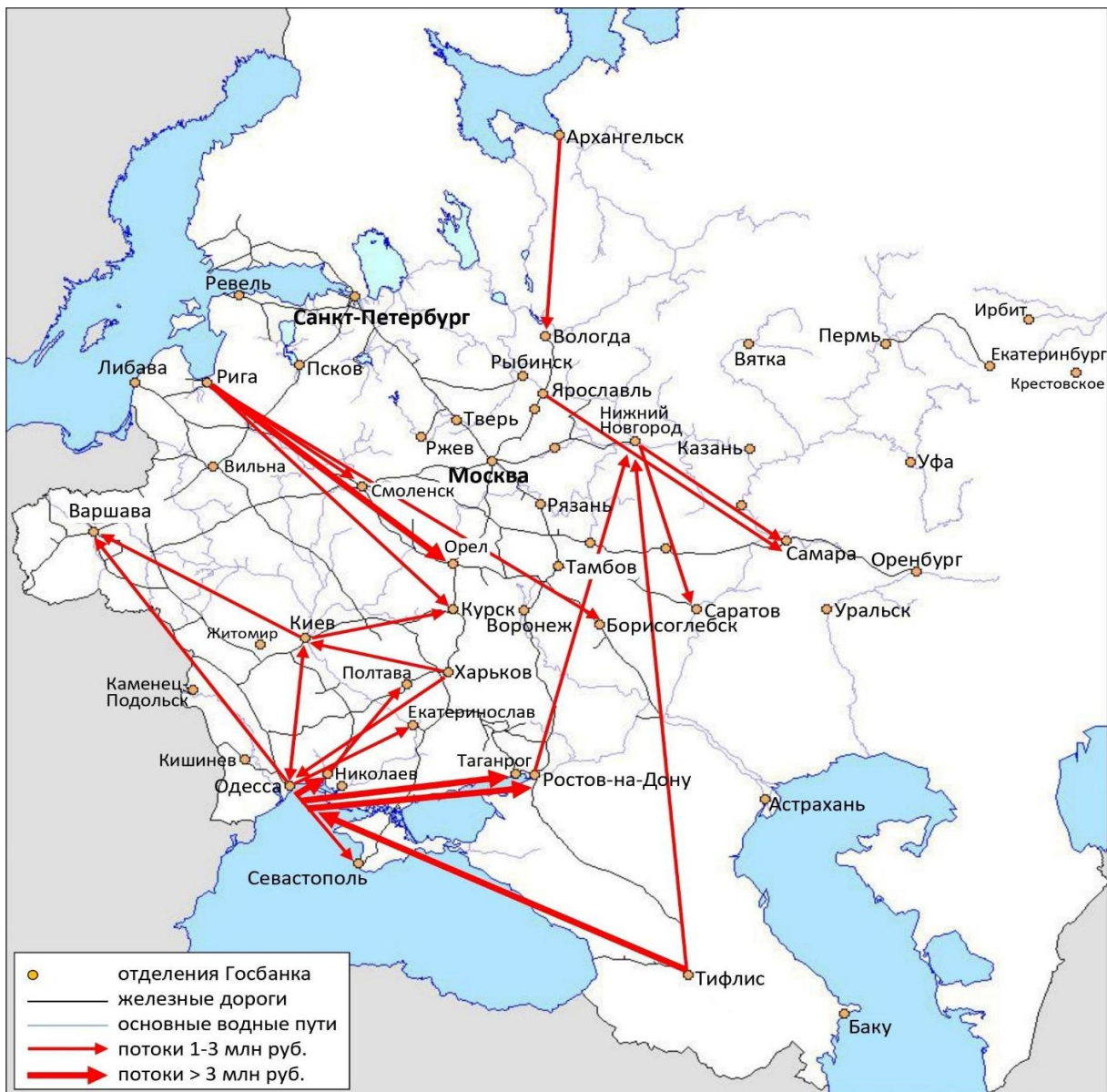


Рис. 1. Переводы между отделениями Госбанка в европейской части Российской империи и на Урале в 1878 г., за исключением Санкт-Петербурга и Москвы, денежные потоки объёмом более 1 млн руб. По Варшаве есть только данные о переводах из отделений Госбанка в Польский банк

В качестве гипотезы, объясняющей разницу этих двух карт, можно предположить, что к 1898 г. стало намного больше крупных ближних связей, чем раньше. Вопрос можно поставить шире – есть ли взаимосвязь между размером денежного потока и расстоянием между этими населёнными пунктами.

Для ответа на этот вопрос используется база данных (более 10 тыс. записей), единицей учёта в которой является связь, или денежный оборот между двумя отделениями, рассчитанный как годовая сумма переводов в обоих направлениях. В базу данных добавлены расстояния в километрах по прямой между каждой парой отделений. Исследование ограничено только европейской частью страны и Уралом, чтобы исключить потоки Сибири и Средней Азии, подчинявшиеся другим закономерностям. Временные срезы, как и в предыдущей статье, – 1868, 1878, 1888 и 1898 гг.



Рис. 2. Переводы между отделениями Госбанка в европейской части Российской империи и на Урале в 1898 г., за исключением Санкт-Петербурга и Москвы, денежные потоки объёмом более 1 млн руб.

Отдельные связи сгруппированы в региональные рынки, или платёжную сеть каждого отделения. Это сеть типа «звезда», в которой из центра (отделение) денежные потоки (связи) направлены к другим отделениям. Для каждого рынка выделены столичные связи (с Санкт-Петербургом и Москвой), ближние связи (в пределах 400 км) и дальние связи (более 400 км). Для каждого рынка рассчитывались, во-первых, доля ближних связей, потому что чем выше эта доля, тем более регионально ориентированным является конкретный рынок, и, во-вторых, коэффициент корреляции между денежными оборотами и расстояниями на этом рынке.

В результате получается, что денежные обороты и расстояния связаны следующим образом. Во-первых, на рынке будет работать принцип «чем меньше расстояние, тем больше обороты», или отрицательная корреляция примерно $-0,4$, если есть достаточно много ближних связей, примерно 50 % в денежном выражении. Во-вторых, если на рынке растёт чис-

ло дальних связей, то коэффициент корреляции будет двигаться из отрицательных значений к нулю. В-третьих, даже если на рынке много дальних связей, существенная положительная корреляция, например, 0,3 (чем дальше расстояние, тем больше оборот) невозможна, потому что на дальних расстояниях обороты в подавляющем большинстве небольшие. Редким исключением из этого являются крупные обороты 1878 и 1888 гг. в западном направлении по железным дорогам из Центрального Черноземья (на Ригу, Либаву, Вильно и близкие к ним центры).

Взаимосвязь оборотов и расстояний представлена на рис. 3. Это диаграмма рассеяния, объектами на которой являются экономико-географические районы, используемые для публикации региональных данных в отчётах Госбанка. Каждый район состоит из нескольких отделений, т. е. из нескольких рынков. В данном случае важен не сам перечень районов, а их расположение на диаграмме рассеяния. По оси X показана доля ближних связей (в %), а по оси Y – коэффициенты корреляции между оборотами и расстояниями.

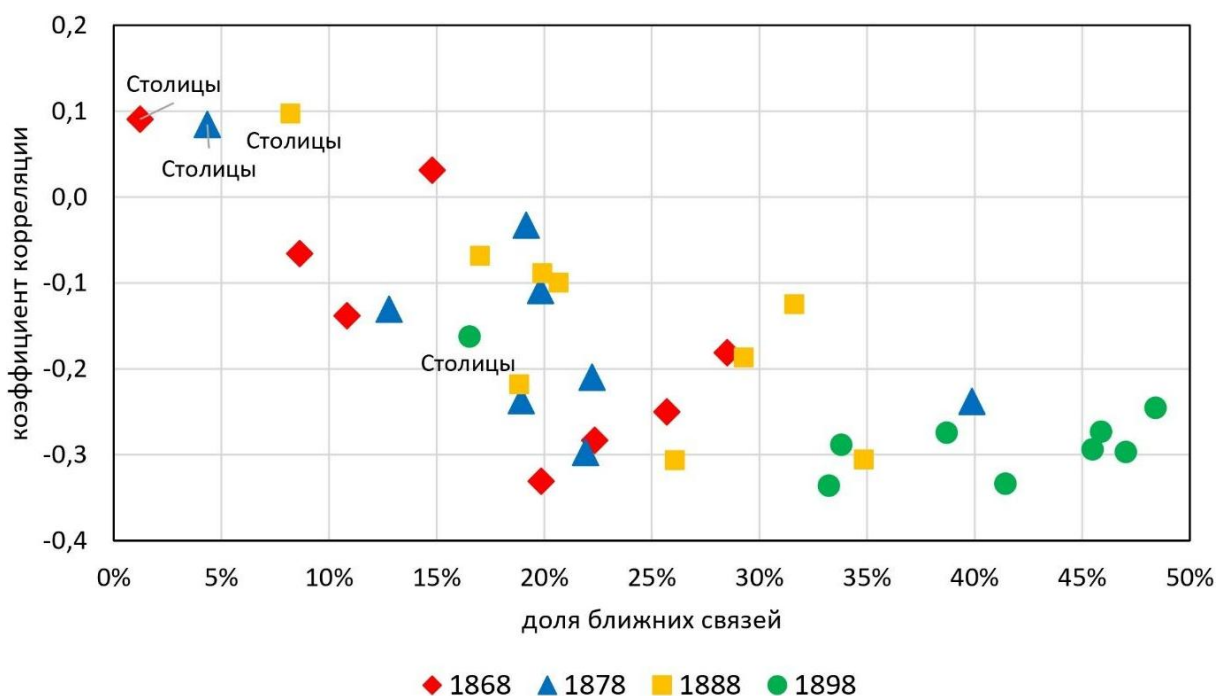


Рис. 3. Корреляция денежных оборотов и расстояний по экономико-географическим районам европейской части Российской империи в зависимости от доли ближних связей на региональных рынках в 1868, 1878, 1888 и 1898 гг.

Из диаграммы рассеяния следует, что в 1898 г. во всех районах европейской части страны и на Урале увеличились обороты на ближних расстояниях и усилилась отрицательная корреляция между оборотами и расстояниями в диапазоне от $-0,34$ до $-0,25$. Таким образом, действительно можно утверждать, что в 1890-е гг. структура товарно-денежных потоков значительно изменилась по сравнению с предыдущими десятилетиями. Обмен на ближних расстояниях вдруг занял существенное место в региональной экономике.

Для сравнения с региональными рынками на рис. 3 показан Столичный район (Санкт-Петербург и Москва). Это высший сегмент платёжной сети, как показано в вышеупомянутой статье – «двойная звезда» [5], т. е. рынки двух столиц и огромный денежный поток между ними. Это высший сегмент платёжной сети, на котором реализованы почти все возможные пары связей, т. е. региональные центры, как правило, имели выход на столицы. Таким образом, этот столичный сегмент сети по структуре не может быть ориентирован на местные связи. И действительно в 1868, 1878 и 1888 гг. корреляция расстояний и оборотов в столичном сегменте отсутствует. Нет её, по сути, и в 1898 г., хотя значимая тенденция в сторону отрицательной корреляции появилась и здесь ($r = -0,16$; $p = 0,028$), т. е. усиление ближних региональных связей в Центральной России проявилось в некоторой степени и здесь.

Уточнение причин усиления ближних связей в 1890-е гг. – это предмет будущих исследований, в которых предстоит ответить на следующие вопросы.

Во-первых, влиял ли рост сети отделений Госбанка на рост ближних связей? От 1868 к 1898 г. эта сеть выросла почти в два раза.

Во-вторых, повлиял ли на рост ближних связей рост сети региональных железных дорог? В 1890-е гг. растут как раз местные дороги.

В-третьих, в какой мере на изменения топологии товарно-денежных потоков в 1890-е гг. повлияла реформа железнодорожных тарифов 1888 г.? Как известно, эта реформа меняла приоритеты между дальними и ближними перевозками. В дополнение к этому, были ли заметные изменения в условиях почтовых и телеграфных переводов?

В-четвёртых, в какой мере рост ближних связей к 1898 г. – это реакция на чрезмерное сжатие этих связей в конце 1880-х гг.? Предыдущее десятилетие было наиболее сложным периодом общемировой длинной депрессии 1873–1893 гг., которая сопровождалась ещё и падением аграрных цен, что в полной мере затронуло европейскую часть Российской империи. Возможно, 1890-е гг. стали просто периодом восстановления связей предыдущего десятилетия.

И в-пятых, нельзя не учитывать общую благоприятную экономическую конъюнктуру 1890-х гг., период самого быстрого промышленного роста Российской империи. Есть много признаков, что этот рост не ограничивался тяжёлой промышленностью, а затронул более широкие слои экономики, включая процессы «на местах». В какой мере рост ближних связей 1890-х гг. отражал и этот процесс?

Исследование проведено при поддержке гранта РФФИ № 20-09-00342. Автор выражает благодарность за полезные советы А. А. Фролову, за помощь в подготовке карт Т. Я. Валетову, за помощь в подготовке данных Е. В. Данилову.

Список источников

1. Ведомость о переводах сумм по переводным билетам и телеграммам за 1868 г. // Отчёт Государственного банка по операциям за 1868 г. СПб., 1869.
2. Ведомость о суммах, переведённых по переводным билетам и телеграммам за 1878 г. // Отчёт Государственного банка за 1878 г. СПб., 1879.
3. Ведомость о суммах, переведённых по переводным билетам и телеграммам за 1888 г. // Отчёт Государственного банка, его контор и отделений за 1888 г. СПб., 1889.
4. Особые приложения к отчёту Государственного банка за 1898 г. СПб., 1899.
5. Саломатина С. А. Внутренние товарно-денежные потоки Российской империи во второй половине XIX в.: сетевой и геоинформационный анализ / С. А. Саломатина // Историческая информатика. 2022. № 4. С. 155–178. DOI: 10.7256/2585-7797.2021.4.37027.

Н. В. Солощенко,

кандидат исторических наук, преподаватель

МИРЭА – Российский технологический университет, Москва, Россия

e-mail: soloshchenko-natalya@yandex.ru

ORCID: 0009-0004-8353-6114

ОБРАЗ «НОВОГО РАБОЧЕГО» ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ СССР В ГОДЫ ПЕРВОЙ И ВТОРОЙ ПЯТИЛЕТОК В ОТРАЖЕНИИ ПУБЛИЦИСТИЧЕСКИХ МАТЕРИАЛОВ (ОПЫТ СЕТЕВОГО АНАЛИЗА И КОНТЕНТ-АНАЛИЗА)

Аннотация. В докладе рассматривается проблема формирования «нового рабочего» пищевой промышленности СССР в период индустриализации. Целью данного исследования является выявление и изучение характерных черт пропагандистского образа «нового рабочего» пищевой индустрии в отражении публицистических очерков, посвящённых истории предприятий и трудовому пути ударников и стахановцев в годы первой и второй пятилеток. Автор использует сетевой анализ в программах *UCINET* и *NETDRAW* и контент-анализ в программе *MAXQDA* в качестве основных методов изучения исторических источников. На основании нескольких этапов работы с оцифрованным текстовым массивом в докладе делается вывод о ведущей роли таких качеств, как коллективизм, профессиональные знания и опыт, молодость, активное участие в социалистическом соревновании и работе комсомольских ячеек в портрете «нового рабочего» пищевой отрасли, который пыталась сформировать советская пропаганда в 1928–1937 гг. в публицистических материалах.

Ключевые слова: *сетевой анализ, контент-анализ, MAXQDA, UCINET, новый советский человек, пищевая промышленность СССР*

В период индустриализации СССР пищевая промышленность была крупнейшей отраслью промышленности группы «Б» по числу крупных фабрик и заводов (4 191 предприятие в 1935 г.) [1, с. 67] и второй в той же группе по численности занятого персонала (798,9 тыс. чел. в 1935 г.) [1, с. 67].

Для оптимального выполнения плановых показателей в годы первых пятилеток на пищевых производствах требовалось создать дополнительные элементы мотивации труда и продумать действенные механизмы социальной мобилизации трудящихся. Одним из них стало формирование собирательного образа «нового советского человека», занятого в производстве продуктов питания, – «нового рабочего» пищевой промышленности СССР.

Рассматриваемая модель пропаганды доводилась до рядовых сотрудников фабрик и заводов разными способами. Весьма показательным примером являются в данном случае популярные очерки об истории предприятий и достижениях рабочих пищевой промышленности, печатавшиеся

издательствами «Профиздат» [2; 3], «Молодая гвардия» [4], «Пищепромиздат» [5] и некоторыми другими в годы первой и второй пятилеток. Эти публицистические произведения были рассчитаны на чтение рабочими коллективами. Как правило, очерки издавались в виде небольших по объёму и формату брошюр. Максимальное число страниц в них обычно не превышало 80. Простым языком, ёмко и образно, они описывали этапы истории развития производства, жизненного пути и профессионального роста ударников, стахановцев, участников социалистического соревнования. Некоторые очерки были построены в форме воспоминаний или автобиографий, в других успехи фабрики или конкретного рабочего описывались от третьего лица. Публикации были направлены на формирование позитивного образа «нового советского человека» труда, на внедрение в массовое сознание персонала фабрик и заводов тщательно отобранных установок и взглядов на взаимоотношения с коллективом и администрацией предприятий, на труд и его результаты, мотивы и стимулы деятельности, ответственность перед товарищами и страной.

Проблема формирования «нового советского человека» многократно становилась предметом исследования в работах российских [6–9] и зарубежных авторов [10–12]. Тем не менее актуальность научных работ, посвящённых анализу форм, техник и способов пропаганды, социального конструирования, не вызывает сомнений в современных реалиях общественной жизни. В настоящее время растёт потребность в использовании исследователями новых компьютеризованных методов изучения исторических источников по этой проблематике. В этой связи весьма полезным станет применение к указанным публицистическим материалам таких популярных методов анализа нарративных источников, как контент-анализ и сетевой анализ. Оба этих метода успешно используются историками для работы с разнообразными историческими материалами [13–19].

Целью данного исследования является изучение характерных черт пропагандистского образа «нового рабочего» пищевой промышленности СССР в годы первой и второй пятилеток на примере публицистических очерков, посвящённых предприятиям и рабочим этой отрасли. Основными методами исследования являются сетевой анализ в программе *UCINET* и его визуализация в программе *NETDRAW*, а также контент-анализ в программе *MAXQDA*.

Этапы работы с текстами очерков включали в себя, во-первых, сбор материала и его традиционный источниковедческий анализ. Всего было изучено 50 официально опубликованных в публицистических очерках биографий рабочих-пищевиков. Их тексты были оцифрованы и распознаны для удобства дальнейших этапов работы.

Во-вторых, был проведён полноценный контент-анализ в программе *MAXQDA*, который включал в себя работу со словарём текстов, формирование системы категорий и индикаторов, лексический поиск и разметку полученного массива, анализ частот встречаемости категорий по отдельности и анализ частот их совместной встречаемости. Категории включали

в себя наиболее важные черты портретов «новых рабочих», отражённые в очерках. Всего было создано 15 категорий. Среди них по относительным частотам встречаемости особо выделяются *участие в социалистическом соревновании, молодость, профессиональные знания и опыт, комсомольская работа*.

В-третьих, полученные матрицы частот совместной встречаемости категорий были импортированы в программу *UCINET*. На основании матриц были построены графы, отражающие семантические связи главных черт собирательного образа «нового рабочего» пищевой промышленности. На данном этапе сначала была изучена структура графов в первоначальном виде. В ходе анализа были выявлены связи категорий, которые не были очевидными. Так, например, категория *молодость* неоднократно связывалась в текстах публицистических материалов с категорией *профессиональные знания и опыт*. В очерках указывалось на глубокие познания и практический опыт молодёжи в организации производственных процессов на пищевых предприятиях в годы первой и второй пятилеток. Вполне очевидно, что и в данном, и в прочих случаях желаемые черты «новых рабочих» выдавались пропагандой за действительные.

Затем полученные в ходе работы сети были перестроены с применением параметра *betweenness* («посредничество»). С учётом этого показателя связующая роль в образе «нового рабочего» пищевой промышленности принадлежала такому качеству, как *коллективизм*. С его помощью происходила «увязка» прочих личных черт, профессиональных навыков, политической и общественной активности в единый образ, который пыталась сформировать пропаганда.

Таким образом, проведённый анализ показал, что в изученных источниках наибольшее значение придавалось преодолению индивидуализма, вовлечению рабочих в активное участие в социалистическом соревновании и повышению активности молодёжи. Особое внимание к молодёжи не было случайным. Именно на молодое поколение в годы первых пятилеток возлагались основные задачи строительства развитой индустрии. По мнению пропаганды, сформировать новые качества рабочего было проще из более «пластичного материала», каким являлись юноши и девушки, не обременённые богатым жизненным опытом и стажем работы на пищевых предприятиях до революции. В рассматриваемый период эти «новые советские рабочие» становились ценным ресурсом, на который могла опереться страна, чья экономика проходила сложные этапы развития в рамках мобилизационной модели.

Я сердечно благодарю своего научного руководителя, члена-корреспондента РАН, доктора исторических наук, профессора Л. И. Бородкина и коллектив кафедры исторической информатики исторического факультета МГУ им. М. В. Ломоносова за знания, терпение, поддержку, доверие и развитие.

Список источников

1. Труд в СССР: стат. справ. / под ред. А. С. Попова. М.: ЦУНХУ Госплана СССР: В/о Союзоргучёт, 1936. 386 с.
2. Выборнова А. И. Путь в жизнь / А. И. Выборнова. М.: Профиздат, 1931. 30 с.
3. Подкопаев Б. И. За сырьём для второй пятилетки / Б. И. Подкопаев. М.: Профиздат, 1931. 39 с.
4. Платанова Л. И. На Бабаевке / Л. И. Платанова. М.: Молодая гвардия, 1931. 31 с.
5. Раськин Р. С. Стахановцы-кондитеры / Р. С. Раськин. М.–Л.: Пищепромиздат, 1936. 62 с.
6. Дискурсы «новый человек» и «советский народ» как общность в социокультурном пространстве СССР: сб. ст. Ставрополь, 2022. 276 с.
7. Кириченко С. В. Советские идеология и культура в 1920–30-х гг.: воспитание «нового человека» / С. В. Кириченко // Вестник ТГУ. История. 2020. № 65. С. 20–26.
8. Поршнева О. С. «Новый человек» как феномен советского проекта индустриального развития (1920 – начало 1930-х гг.) / О. С. Поршнева // Индустриальное развитие региона и мира: история и современность: матер. Всеросс. НК (Екатеринбург, 2018). С. 231–240.
9. Человек советский: за и против | Homo Sovieticus: Pro et Contra: моногр. / под общ. ред. Ю. В. Матвеевой, Ю. А. Русиной. Екатеринбург: УрФУ, 2021. 412 с.
10. Bernstein S. Raised under Stalin. Young Communists and the Defense of Socialism / S. Bernstein. Cornell University Press, 2017. 268 p.
11. Sharafutdinova G. Was There a "Simple Soviet" Person? Debating the Politics and Sociology of "Homo Sovieticus" / G. Sharafutdinova // Slavic Review. 2019. Vol. 78. No. 1. Pp. 173–195.
12. Soboleva M. The Concept of the "New Soviet Man" and its Short History / M. Soboleva // Canadian-American Slavic Studies. 2017. Vol. 51. No. 1. Pp. 64–85.
13. Алеткина Е. Ю. Информационный потенциал многотиражных газет предприятий нефтяной промышленности Урало-Поволжья в 1960–80-е гг. как источников по изучению мотивации труда: контент-анализ писем в газеты / Е. Ю. Алеткина // Историческая информатика. 2023. № 2. С. 1–18.
14. Алеткина Е. Ю. Отражение проблем мотивации труда в письмах работников нефтяной промышленности Урало-Поволжья в 1956 г. (контент-анализ территориально-отраслевых газет «Нефтяник Татарии» и «Нефтяник Башкирии») / Е. Ю. Алеткина // Историческая информатика. 2022. № 4. С. 135–154.
15. Горецкая Е. М. Память о ГУЛАГе в воспоминаниях заключённых: гендерный и хронологический аспекты контент-анализа / Е. М. Горецкая // Исторический журнал: научные исследования. 2023. № 3. С. 1–18.
16. Гребенченко И. В. Проект «Союз-Аполлон» в советских (российских) и американской газетах: контент-анализ / И. В. Гребенченко // Историческая информатика. 2023. № 1. С. 90–101.
17. Бородкин Л. И. Сетевой анализ в исторических исследованиях: микро- и макроподходы / Л. И. Бородкин // Историческая информатика. 2017. № 1. С. 110–124.
18. Гребенченко И. В. Контент- и сетевой анализ воспоминаний советских космических конструкторов: круг коммуникаций / И. В. Гребенченко // Клио. 2021. № 8 (176). С. 31–39.
19. Ермошин А. Д. Сетевой анализ просопографической базы данных об архитекторах Московского метрополитена 1935–1991 гг. / А. Д. Ермошин // Историческая информатика. 2017. № 4. С. 130–142.

А. В. Стальмакова,
аспирант

Балтийский федеральный университет им. И. Канта, Калининград, Россия
e-mail: angelinasaenko08@gmail.com
ORCID: 0000-0003-1285-2342

ПРОБЛЕМЫ СОВЕТСКОЙ ИСТОРИИ НА СТРАНИЦАХ КАЛИНИНГРАДСКОЙ ПАРТИЙНОЙ ПРЕССЫ В ПЕРИОД ПЕРЕСТРОЙКИ

Аннотация. Предпринимается попытка через призму региональной прессы проследить трансформацию отношения к советской истории в период перестройки. С помощью контент-анализа проанализирована генеральная совокупность главной областной газеты «Калининградская правда», выделены основные исторические явления и события. Сделан вывод о незначительном интересе к ключевым сюжетам общесоюзной истории, преимущественном сохранении предыдущего дискурса в трактовке Октябрьской революции, Гражданской и Великой Отечественной войн. Кардинальное изменение отношения к дискуссионным проблемам истории СССР произошло лишь в конце перестройки – в 1989–1991 гг. Это было вызвано необходимостью реагировать на письма читателей, активно участвующих в обсуждении «белых пятен» советской истории, критически относящихся к политическим практикам сталинизма и в целом к историческому наследию советского периода.

Ключевые слова: контент-анализ, советский народ, перестройка, пресса, Ленин, сталинизм, Калининградская область

Провозглашение политики «гласности» в период перестройки качественно повлияло на информационный дискурс, принятый в советском обществе. Особенно важным стало обращение прессы к историческому нарративу о недавнем прошлом страны, который подвергся существенному пересмотру. Новые и переосмысленные старые сюжеты способствовали формированию нового типа исторической памяти, заметно повлиявшей на массовое общественное сознание. Вслед за газетой «Московские новости» во многих центральных изданиях был взят на вооружение лозунг о необходимости заполнения «белых пятен» истории, т. е. обнародования тех исторических фактов, которые фальсифицировались или скрывались властью [4, с. 30].

По мнению М. В. Новиковой, к процессу пересмотра советской истории в годы перестройки правомерно применение термина «историческая политика»: «Пласт публикации по исторической проблематике был вписан в определённый идеологический контекст и служил цели дать историческое обоснование необходимости реформирования советского общества» [3].

Парадоксальность ситуации отмечает А. Юрчак: «Дискурс перестройки и гласности вскрыл множество неизвестных фактов, заставил критически взглянуть на многие явления советского прошлого, в то же время способствуя созданию новых мифов о советском прошлом. В результате невероятной увлечённости новыми публикациями в общественном языке появились новые темы, сравнения и идеи, которые привели к изменению доминирующего дискурса и сознания» [4, с. 43].

Вместе с изданием запрещённых ранее литературных произведений Ахматовой, Гумилева, Солженицына и других авторов возросло число научных и публицистических работ о политических репрессиях и последствиях культа личности Сталина. Эта проблематика заняла прочное место среди самых актуальных тем в периодической печати.

В Калининградской области интерес к прошлому имел свои особенности, в основном он концентрировался на региональной истории довоенного времени. Если в других регионах СССР (например, Горьковской области) «досоветская» история в основном понималась как дореволюционная, то для истории Калининградской области рубеж проходил в 1945 г., когда после завершения Великой Отечественной войны северная часть бывшей Восточной Пруссии вошла в состав СССР. В процессе формирования на присоединённой территории «новой исторической общности» – «советского народа» [1, с. 4] – произошёл отказ от изучения довоенной истории, выразившийся в признании бывшей провинции Германии «разбойничьим оплотом немецкой военщины и реакции» и провозглашением политики «изгнания прусского духа» [2, с. 14].

Особую роль в формировании и продвижении официального исторического дискурса играла главная газета области, орган областного комитета партии и облисполкома «Калининградская правда». Цель данного исследования – проследить изменения в исторической политике, отражённые на страницах региональной прессы. Из генеральной совокупности 1 772 публикаций по исторической проблематике, напечатанных в «Калининградской правде» с 1985 по 1991 гг., 609 (34 %) посвящены советскому периоду (1917–1991 гг.).

Контент-анализ газеты производился в программе MAXQDA 2020. На основе сгенерированного частотного словаря были выделены следующие смысловые категории, отражающие главные явления и события советской истории: **Октябрьская революция и Гражданская война, сталинизм, Великая Отечественная война, послевоенное восстановление личности, практики коммеморации**. Дополнительно были проанализированы авторство статей и отношение к историческим темам и личностям.

В первую очередь необходимо отметить, что в отличие от центральной прессы, в которой доминировали «трудные вопросы» истории 1920–30-х гг., в «Калининградской правде» по-прежнему центральное место (81 %) занимала тема Великой Отечественной войны, что объясняется не только инерционной политикой редакции, но и особым значением этого события для создания самой западной российской области. С другой сто-

роны, самые острые полемические темы, относящиеся к трактовке революционного наследия и оценке эпохи сталинизма в совокупности не превышали 10 % от общего числа публикаций в газете по советской истории.



Рис. 1. Темы советской истории, отражённые в «Калининградской правде»

Немалая часть публикаций касалась ключевых фигур советской истории, газета публиковала информацию о вождях революции, героях Великой Отечественной войны, космонавтах. Здесь лидировало имя Владимира Ильича Ленина – всего 113 раз. Его упоминание почти исключительно в положительном контексте касалось всех ключевых исторических проблем СССР (см. рис. 2). К концу перестройки стали появляться сообщения, косвенно отражающие изменение отношения к советским вождям: «На днях в посёлке Новостроево Озёрского района был разобран памятник В. И. Ленину. Кто это сделал, пока неизвестно. Как сообщила нам специалист Новостроевского сельского совета Л. И. Лазарева, фигура разборная. У неё были сняты голова, плечи и рука. Оставалась часть туловища с ногами. На сегодня всё приведено в прежний вид. Памятник вождю в сборе вновь стоит, как и стоял, в сквере у отделения связи» [5].

Результаты контент-анализа публикаций в газете «Калининградская правда» по советской истории позволили сделать следующие выводы.

1. Советский период занимал лишь одну треть в общей совокупности публикаций по исторической тематике, тогда как основной интерес в годы перестройки оказался смещён в сторону довоенной немецкой истории края, которая ранее находилась под негласным запретом. При этом более половины всех статей по советской истории являлись спущенными сверху по разрядке материалами ТАСС или перепечатками из центральной партийной прессы.

2. Материалы по наиболее острым политическим сюжетам советской истории, которые в это время волновали общество, для калининградской прессы не являлись приоритетными, занимая лишь 10 % всех публикаций по советской истории, что объяснялось, во-первых, довольно консервативной политикой редакции, а во-вторых, тем, что основное внимание областной партийной газеты было сосредоточено на Великой Отечественной войне, с которой было связано рождение Калининградской области.

3. Важной составляющей публикаций по истории СССР были фигуры В. И. Ленина и И. В. Сталина. В газете сохранялся положительный образ Ленина как основателя советского государства, возврат к идеям которого происходил в годы перестройки. Личность Сталина упоминалась нейтрально в контексте событий Великой Отечественной войны и резко негативно в связи с культом личности и репрессиями.

Список источников

1. «Советский народ» на крайнем западе России (1945–1991 гг.): теория и практика формирования «новой исторической общности» в Калининградской области. Калининград: БФУ им. И. Канта, 2022. 235 с.

2. Костяшов Ю. В. Изгнание прусского духа: как формировалось историческое сознание населения Калининградской области в послевоенные годы / Ю. В. Костяшов. Калининград: КГУ, 2003. 161 с.

3. Новикова М. В. Историческая проблематика в советской газетной периодике в годы перестройки: сравнительный анализ центральной и региональной прессы (на материалах Нижегородской (Горьковской) области): автореф. дисс. / М. В. Новикова. 2018.

4. Юрчак А. Это было навсегда, пока не кончилось. Последнее советское поколение / А. Юрчак. М.: Новое литературное обозрение, 2020. 664 с.

5. Вновь стоит, как и стоял // Калининградская правда. 25.09.1991. С. 4.

А. А. Стороженко,
кандидат исторических наук, доцент
Тувинский государственный университет, Кызыл, Россия
e-mail: storozhencko.ale@yandex.ru
ORCID: 0000-0001-9986-8966

СТАРООБРЯДЧЕСКИЕ МИГРАЦИИ В 30–50-Е ГОДЫ XX ВЕКА В СИБИРЬ: ОПЫТ ПРИМЕНЕНИЯ ЦИФРОВЫХ ТЕХНОЛОГИЙ

Аннотация. Благодаря современным информационным технологиям – технологии баз данных, сетевому анализу – появляется возможность более эффективно и предметно проследить особенности размещения населения, формирования направлений хозяйственной деятельности в зависимости от природно-климатических и социально-исторических условий на основе массовых статистических источников. В то же время следует отметить, что идущая апробация иных методологических подходов на более общих процессах развития локальных сообществ эмпирически может оказаться достаточно продуктивным выходом для завершения принципов системного подхода изучаемого предметного поля. В настоящей работе предложен вариант реконструкции процесса складывания конфессионального пространства современной Енисейской Сибири в исторической ретроспективе на основе методов и инструментов *Digital Humanities*. Она заняла адекватное место в источниковедческой системе познания староверов-часовенных.

Ключевые слова: старообрядцы, Сибирь, переселения, сетевой анализ, базы данных, моделирование, конфессиональная миграция

История складывания старообрядческих общин часовенного согласия в Сибири, их внутренней структуры, культуры и религиозной жизни давно привлекает внимание исследователей. Изучение переселенческих вопросов неизбежно вызывает как минимум упоминание о роли старообрядческого населения в этом процессе, ведь в большинстве сибирских регионов староверы были первопроходцами и первыми покорителями новых земель. Актуальность изучения процесса переселения в Сибирь старообрядцев Пермско-Вятского Прикамья состоит в необходимости исторической реконструкции и анализа практик социокультурного освоения территории Сибири конфессиональными сообществами.

Скитские поселения, являющиеся ведущим элементом конфессионального сообщества старообрядцев «енисейского меридиана», пережившие за последние 70 лет немало структурных коллизий, нуждаются в систематизации и визуализации и по количеству, и по структуре размещения. Последнюю трудно адекватно представить без технологии сетевого анализа, позволяющей уточнять локализацию как целых согласий, так и отдельных общин, проследить зависимость, преимущественно кустовую,

скитских и мирских поселений, наглядно видеть природосообразный характер поселений, различия религиозных практик.

Имеющиеся в распоряжении автора архивные материалы (похозяйственные книги, посемейные списки) и опубликованные биографические списки позволяют систематизировать материал в виде источникориентированной базы данных, создание которой не просто его упорядочивает, но и позволяет делать более глубокие обобщения.

Моделирование процесса конфессиональной миграции старообрядцев Пермско-Вятского Прикамья / Молотовской области в Сибирь в середине XX в. сделано на основе организации данных массовых статистических источников, опубликованных документов при помощи СУБД *Access* и последующего анализа посредством запросов. Этот подход для дальнейшего комплексного изучения трансформации социокультурного ландшафта является перспективным направлением изучения миграций как таковых.

Созданная база данных «Конфессиональные миграции старообрядцев» (далее – БД КМС) содержит сведения о более чем 1 500 старообрядцах, переселившихся в период с 1925 по 1945 гг. с территории преимущественно южных районов современного Пермского края, а также соседних уральских и сибирских областей на территорию Луговатского сельсовета Енисейского района Красноярского края. Она содержит сведения о поселенческой структуре конфессиональных мигрантов, половозрастном составе населения Сибири во второй четверти XX в., основным направлениям и объёмам переселений. Есть возможность создать социокультурный коллективный портрет конфессионального переселенца и произвести пространственное картографирование сети населённых пунктов.

Пофамильный (посемейный) список содержит сведения о половозрастном составе конфессиональных мигрантов, семейном и конфессиональном статусе, семейной роли и конфессиональном сане. База также содержит сведения о местах выхода переселенцев и местах вселения, даты въезда. Геокодирование позволяет представить процесс формирования сети мирских и конфессиональных поселений. На этой основе возможна визуализация самой трудно поддающейся анализу сферы бытования конфессиональных общин – процессу складывания конфессиональных центров – скитов часовенного согласия.

Именно такая модель расселения и взаимодействия может быть понимаема в современной лексике как «сетевая». Подобный кустовой тип размещения населения характеризуется наличием стохастических для внешнего наблюдателя «узлов» различной направленности. При этом конфессиональный вектор как самой миграции, так и обустройства и жизнедеятельности становится стержневым, факторным для воспроизводства специфической «сети». В зависимости от внутренних рефлексий и внешних вызовов возможно произвольное перемещение как информационных потоков, так и их носителей – человеческого ресурса.

Проанализированные массивы данных были подвергнуты дальнейшему социально-сетевому анализу, позволившему визуализировать направления миграционных потоков, выделить наиболее значимые маршруты, оценить их объёмы и пофамильный состав как отдельных кланов, так и семейно-родственных корпораций. Технология сетевого анализа позволяет уточнить направления, объёмы и динамику конфессиональной миграции старообрядцев на макроисторическом уровне и в дальнейшем проследить процесс складывания конфессиональных центров уже на уровне микроанализа.

П. С. Воронина,

администратор Гуманитарного инновационного парка, научный сотрудник
Петрозаводский государственный университет, Петрозаводск, Россия
e-mail: psvoronina@yandex.ru
ORCID: 0000-0003-1899-3808

БАЗА ДАННЫХ «КАРБЕЖЕНЦЫ: ЛЮДИ И СУДЬБЫ (1920-Е ГОДЫ)»: КОНЦЕПЦИЯ И СТРУКТУРА

Аннотация. С 2015 г. на базе Исследовательской лаборатории локальной и микроистории Карелии Петрозаводского государственного университета (руководитель – кандидат исторических наук, доцент Ирина Александровна Чернякова) ведётся разработка и заполнение базы данных, посвящённых карбеженцам (карельским беженцам). Наполнение базы во многом диктует источник – сохранившиеся в фондах Национального архива Карелии анкеты, опросные листы и протоколы карельских крестьян, принявших решение вернуться в СССР из Финляндии после вынужденной миграции в 1921–1922 гг. В данной статье представлена структура базы, концепция её создания и наполнения, которые беспрерывно дополняются и совершенствуются.

Ключевые слова: *Карельская трудовая коммуна, карбеженцы, карельская авантюра, карельское восстание, база данных*

В 1921–1922 гг. в северных волостях Карельской трудовой коммуны происходили события, причины и характер которых до сих являются предметом дискуссии. В отечественной историографии они определяются по-разному: карельская авантюра, белофинская авантюра, карельский мятеж, карельское восстание [3; 4; 6; 7]. Однако, вне зависимости от особенностей событий 1921–1922 гг. и точки зрения, с которой они рассматриваются, как непосредственные участники, так и исследователи сходятся во мнении, что конфликт стал причиной массовой миграции обитателей карельских деревень.

Председатель СНК КАССР Эдвард Гюллинг в очерке, посвящённом 10-летию установления советской власти в Карелии, подчеркнул: «белые частью насильственно, частью напугиванием [так – авт.] увели с собой» около 10–12 тыс. чел. [5, с. 67], а исследователь С. С. Хесин утверждал, что «во всём Ухтинском районе из 17 600 чел. ... осталось 7 358» [8, с. 144]. По оценкам НКВД, в Финляндию ушло более 12 тыс. чел. [10]. Катастрофическая ситуация сложилась в Тихтозерской волости: волисполком приводил данные о том, что до восстания в волости проживало 1 420 чел., из них в Финляндию бежало 1 400 [2, с. 61]. Так, например, согласно данным проводившегося зимой 1922 г. обследования населённых пунктов волости, в деревне Хирвисалма, «состоящей из восьми дворов, осталась только слепая старуха в возрасте 76 лет, без всяких средств к существованию». Уточнялось, что жители деревни «добровольно перешли

на сторону бандитов и увезли с собой всё имущество, а также и весь скот» [14, л. 28]. Полностью обезлюдила крупная деревня Охта, где до конфликта насчитывался 21 двор [14, л. 79]. В волостном центре – деревне Тихозеро – комиссией были обнаружены только «один старый старик, две старых старухи и 11-летний мальчик», которые «не могли по старости и молодости перейти к бандитам» [14, л. 39].

Стремясь вернуть население на покинутые территории, советское правительство приняло решение объявить бежавшим на территорию Финляндии карельским крестьянам амнистию. Для координирования этой работы была создана отдельная структура – Комиссия по амнистированию карбеженцев (карельских беженцев)¹³ [1, с. 31]. Впоследствии амнистия несколько раз продлевалась. Была разработана и процедура амнистирования, включающая в себя заполнение карбеженцами и сотрудниками спецслужб специальных документов: анкет, опросных листов, протоколов.

Сегодня анкеты и опросные листы карбеженцев хранятся в Национальном архиве Республики Карелия в составе фонда Р-689 «Президиум Верховного Совета РК (1921–1994)». До 1990 г. данные дела хранились под грифом «секретно».

При первичном просмотре стало понятно, что провести анализ данного комплекса источников – более 7 000 архивных листов – возможно лишь исключительно с использованием информационных технологий. Однотипность документов, одинаковость времени и обстоятельств их возникновения позволили определить, что основным инструментом их анализа с целью создания коллективного портрета карбеженцев станет база данных.

В основу карточки БД легла анкета образца 1923 г., состоящая из четырёх информационных блоков:

- 1) блок А «Общие вопросы»;
- 2) блок Б «Вопросы, касающиеся вооружённых выступлений в Карелии в 1921 и 1922 гг. (зимой)»;
- 3) блок В «Вопросы, касающиеся проживания просителя в Финляндии»;
- 4) блок Г «Вопросы, касающиеся возвращения в Карелию».

Анкета заполнялась на русском или финском языках в трёх экземплярах. При этом жёны, состоящие в браке свыше 6 мес., а также дети до 18 лет вносились в анкеты мужа и отца. Все лица, указанные в анкете, а также сам подаватель должны были представить четыре свои фотографии. Этот пункт не касался детей до 14 лет.

Карточка БД была разработана в программе *FileMaker Pro*. К её преимуществам можно отнести простоту представления данных, возможность внесения изменений как в структуру самой таблицы (включение или удаление новых граф и слоёв), так и в уже внесённую информацию (исправление ошибок, дополнения, комментарии). Слои карточки были созданы

¹³ Карбеженцами или карельскими беженцами называли жителей карельских деревень, которые ушли в Финляндию после событий 1921–1922 гг.

в соответствии с обозначенными выше блоками и дополнены вопросами из опросного листа. Например, только опросный лист содержал вопрос об уровне образования и национальности. Помимо этого, в отдельные слои вносятся информация из протоколов допросов и эвакуационных свидетельств.

При разработке БД и формировании принципов её заполнения было принято решение действовать в рамках источник-ориентированного подхода: информация в карточку вносится так, как она указана в источнике. При этом, если в разных документах на один и тот же вопрос представлены разные ответы, то эти различия фиксируются в карточке с обозначением, из какого документа почерпнута приведённая информация. Такой подход позволяет сопоставлять данные из различных документов и проводить их проверку.

Зачастую такие различия проявляются в именах и датах рождения. Например, уроженка д. Лайдосалмы Олангской волости Лукерья Андреевна Краскова в опросном листе указана как Евгения [12, л. 278–281], а в документах Родиона Андреевича Мاستинена из той же деревни представлен разный год рождения: 1877 г. в анкете и 1872 г. в опросном листе [13, л. 284–285]. Фиксируются также все варианты имён и фамилий карбеженцев. Так, у Василия Осиповича Архипова из д. Кизреки Кестеньгской волости была финская фамилия Хаапала [11, л. 374], а Александр Ларионович Анциферов из д. Ушково Маслозерской волости имел второе имя и фамилию – Сантери Мартинен (*Santteri Martinen*) [11, л. 363].

Помимо этого, особое внимание уделяется различным пометам, выявленным на документах. Их содержание вносится в отдельное поле. Так, на опросном листе Ивана Васильевича Архипова из Кизреки помета гласит, что «значится Архипов Иван Васильевич из д. Сенозеро Кестеньгской волости в списках бандитов как ушедших вместе с бандитами в Финляндию» [11, л. 430]. Приведённый пример свидетельствует, что анкеты карбеженцев проходили тщательную проверку.

В отдельный слой карточки вносятся тексты протоколов допросов. Это довольно трудоёмкий процесс, однако именно данные нарративы позволяют уточнить как причины миграции, так и последовательность действий карбеженцев.

Михаил Иванович Орехов (ошибочно записанный как Арехов) из д. Костамуксы указал в анкете, что причиной выезда в Финляндию стало тяжёлое материальное положение [11, л. 346–350]. Согласно протоколу допроса, основным фактором, побудившим покинуть родную деревню как самого Михаила Ивановича, так и всю его семью, стал голод, причём уход происходил в разное время, в т. ч. до начала событий осени 1921 – зимы 1922 г.: «В 1921 г. в нашей местности стал ощущаться недостаток в хлебе, а посему я решил идти на заработки в Финляндию. В мае месяце 1921 г., числа не помню, я и осуществил поставленную перед собой задачу, т. е. перешёл госграницу на сторону Финляндии на участке против своей деревни. О своём намерении уйти в Финляндию я в известность никого

не ставил. Остальные члены семьи остались дома, за исключением брата Семёна, который ушёл в Финляндию за много раньше меня, тоже не поставив никого в известность. Осенью 1921 г. приходил в Финляндию мой отец. В июле месяце ушёл из дому старший брат Тимофей и во время Каравантюры [в] 1921 г. пришли в Финляндию мать и обе сестры. Остальных членов моей семьи уйти в Финляндию также побудил голод» [там же].

Голод, а также отсутствие работы стали причиной ухода Марии Романовны Андреевой из д. Мулдусельги Сямозерской волости в сентябре 1923 г. В протоколе допроса она отметила, что «в Финляндию ушла ввиду того, что дома не было хлеба и одежды, а работать никуда не принимали. В известность своих родителей об уходе в Финляндию я не ставила, и они об этом не знали, т. к. я пошла в лес за ягодами, а вместо того ушла в Финляндию» [1, л. 199–201].

Приведённые свидетельства вступают в некое противоречие с данными, представленными на VI Всекарельском съезде советов в 1926 г., где подводились итоги амнистии. Согласно анализу анкет, в зависимости от причины ухода в Финляндию беженцы были поделены на три группы: покинувшие Карелию в порядке военной мобилизации (сюда отнесли руководителей и активистов восстания); совершившие добровольный переход по «неведению сути и значения Советской власти» (они были обозначены как «масса» и составляли, как можно предположить, основную часть репатриантов); вынужденные покинуть Карелию из-за, как отмечено, насильственных действий белогвардейцев. По этой причине, по мнению представителей власти, могли уйти в Финляндию лишь представители беднейшего крестьянства [1, с. 21]. Однако даже единичные примеры из анкет карбеженцев показывают, что причины миграции имели более сложный характер, включая тяжёлое экономическое положение в крае и вызванную в связи этим бытовую неустроенность населения.

Для более глубокого анализа биографий карбеженцев особое внимание уделяется поиску и внесению в базу информации из других источников: это дополнительные комплексы анкет, данные метрических книг, переписей, таможенных книг, различных обследований. Приведём один пример.

Егор Андреевич Красков 1842 г. р. из д. Лайдосалмы возвращался из Финляндии осенью 1923 г. С ним, по данным анкеты, следовала жена Мария Аверкиевна, 65 лет, и дочь Марфа, 45 лет [12, л. 274]. При этом в материалах переписи 1897 г. в д. Лайдосалме Олангской волости Кемского уезда Архангельской губернии зафиксировано домохозяйство государственного крестьянина Егора Андреева Краскова. На момент переписи хозяину было 49 лет. Основным занятием он указал рыболовство, был неграмотным, принадлежал к филипповскому согласию. Вместе с ним проживала работница Мария Аверкиева Карпова (40 лет) и двое её незаконнорождённых детей – Назар (19 лет) и Марфа (16 лет), немые от рождения [9, л. 152]. Поскольку анализ других переписных листов по Олангской волости не показал настолько полного совпадения в именах, то пола-

гаем возможным идентифицировать описанных выше карбеженцев и крестьян, зафиксированных в переписи, как одних и тех же людей.

Отметим также, что возможности программы *FileMaker Pro* позволяют прикреплять к карточке БД изображения. Как правило, это фотографии карбеженцев, которые прилагались к анкетам, а также изображения из других источников. Такой подход позволяет не нарушать аутентичности источника при его анализе.

Исследование выполнено за счёт гранта Российского научного фонда № 23-28-10260 (URL: rscf.ru/project/23-28-10260), проводимого совместно с Республикой Карелия с финансированием из Фонда венчурных инвестиций Республики Карелия (ФВИ РК).

Список источников

1. Воронина П. С. Амнистирование карельских беженцев 1923 г. / П. С. Воронина // CARELiCA: науч. эл. журнал. 2018. № 1/2018 (19). С. 16–25. URL: [carelica.petrSU.ru/CARELiCA/_1_2018_\(19\)/voronina.html](http://carelica.petrSU.ru/CARELiCA/_1_2018_(19)/voronina.html).
2. Воронина П. С. Карельские беженцы в Ухтинском уезде: региональный аспект (1920-е гг.) / П. С. Воронина // CARELiCA: науч. эл. журнал. 2021. Вып. 1. № 25. С. 56–66. URL: [carelica.petrSU.ru/CARELiCA/_1_2021_\(25\)/voronina.html](http://carelica.petrSU.ru/CARELiCA/_1_2021_(25)/voronina.html).
3. Гардин Е. С. Славная страница в летописи побед Советской Армии: Ликвидация белофинской авантюры 1921–1922 гг. в Карелии / Е. С. Гардин // На рубеже. 1952. № 2. С. 68–77.
4. Гусев К. В. К истории карельского мятежа (по материалам комиссии по реабилитации при Президенте РФ) / К. В. Гусев // Отечественная история. 1996. № 6. С. 71–84.
5. Гюллинг Э. А. Десять лет Карельской Автономии / Э. А. Гюллинг // Десять лет Советской Карелии. Петрозаводск: ЦИК АК ССР, 1930. URL: elibrary.karelia.ru/book.shtml?id=9310&cType=1.
6. Килин Ю. М. Карелия в политике советского государства / Ю. М. Килин. Петрозаводск: ПетрГУ, 1999.
7. Седякин А. И. Ликвидация белофинской авантюры в Карелии / А. И. Седякин // Красный страж. 1937. № 2.
8. Хесин С. С. Разгром белофинской авантюры в 1921–1922 гг.: военно-политический очерк / С. С. Хесин. М.: Воениздат, 1949.
9. ГААО (Государственный архив Архангельской области). Ф. 6. Оп. 19. Д. 25.
10. НАРК (Национальный архив Республики Карелия). Ф. Р-689. Оп. 1. Д. 57/601.
11. НАРК. Ф. Р-689. Оп. 14. Д. 1/1.
12. НАРК. Ф. Р-689. Оп. 14. Д. 3/9.
13. НАРК. Ф. Р-689. Оп. 14. Д. 4/11.
14. НАРК. Ф. Р-1932. Оп. 1. Д. 1/8.

А. А. Акашева,
кандидат исторических наук, доцент
Нижегородский государственный университет им. Н. И. Лобачевского,
Нижний Новгород, Россия
e-mail: annakasheva@yandex.ru

НАБОРЫ ДАННЫХ, ЛОКАЛЬНЫЕ ПРОЕКТЫ, ОНЛАЙНОВЫЕ РЕСУРСЫ ОБ ИЗДАНИЯХ ПЕРВОЙ ВСЕОБЩЕЙ ПЕРЕПИСИ НАСЕЛЕНИЯ 1897 ГОДА РОССИЙСКОЙ ИМПЕРИИ

Аннотация. В докладе представлен опыт и результаты изучения опубликованных изданий переписи населения 1897 года в электронных онлайн-каталогах и фондах 89 библиотек России. Делается вывод о том, что изначальная источниковедческая задача, реализованная на базе информационных и сетевых технологий, открыла возможности дальнейшего изучения материалов переписи в междисциплинарном разрезе – на стыке социокультурной и социальной истории России, истории отечественной статистики и книговедения.

Ключевые слова: перепись 1897 года, история России, история статистики, история книжного дела, электронная каталогизация, веб-приложения, ленты времени, электронные таблицы, базы данных, наборы данных, графический векторный редактор, метод компьютерного монтажа изображений, Лев Манович

С 2019 г. я занимаюсь вопросами анализа опубликованных материалов переписи 1897 года как исторического источника. Этот источникориентированный подход открыл путь для изучения проблемы распространения статистической информации в дореволюционной России. В докладе я презентую ресурсы, которые были созданы мною в ходе источниковедческого анализа переписи. Они могут представлять интерес для специалистов, занимающихся схожими вопросами применительно к своим историческим документам.

На первом этапе (2019 г.) мы собрали, упорядочили и систематизировали всю информацию об изданиях переписи [1] и о местах их хранения в 89 библиотеках РФ с помощью СУБД *MS Access*. Мы анализировали, насколько полно представлены тома переписи в библиотеках, в частности нас интересовали проблемы их электронной каталогизации, дублетности записей, качество оцифровки изданий [2; 3]. Наиболее ценная с практической точки зрения информация о доступности конкретного тома в конкретной библиотеке оформлена в виде электронной таблицы *MS Excel*, в которую мы выгрузили данные из БД и представили здесь [4]. В ней указаны прямые ссылки на все имеющиеся полнотекстовые версии томов во всех региональных и федеральных библиотеках России по состоянию на 2022 г.

В 2021 г. была создана интерактивная лента времени [5], в которой мы разместили все издания переписи по дате их выхода в свет. Мы выделили этапы публикации переписи, которые до этого в историографии отсутствовали, и выдвинули предположение о том, что результаты переписи могли быть опубликованы не за 9 лет, а за 3,5 года [6]. В этом же таймлайне мы разместили ссылки

на все оцифрованные на сегодняшний момент копии переписных томов, хранящихся в РНБ и РГБ.

Наконец, в текущем году, используя метод компьютерного монтажа изображений Льва Мановича, который он применил к анализу обложек популярного американского журнала *Time* [7], мы собрали в своём локальном проекте обложки всех изданий переписи с помощью программы векторной графики *InkScape*. Благодаря этому мы смогли выявить три типа обложек и, соответственно, три типа изданий, которые предназначались для определённого круга читателей-потребителей статистической информации: для императорской фамилии; для высших правительственных учреждений и сановных лиц империи; для широкой публики [8].

Этот проект подвёл нас вплотную к идее о том, что опубликованные тома переписи нужно рассматривать в свете вопросов востребованности и доступности статистических сведений среди населения страны до революции. Именно эта тема является перспективной, т. к. она находится на стыке источниковедения, социальной и социокультурной истории России, истории отечественной статистики и книговедения. Новым импульсом к её развитию служит применение информационных технологий и аналитических возможностей, предоставляемых различными сервисами визуализации в вебе.

При поддержке Программы академического стратегического лидерства «Приоритет-2030» Министерства науки и высшего образования Российской Федерации.

Список источников

1. Справочники по истории дореволюционной России: библиографический указатель / под ред. П. А. Зайончковского. 2-е изд. М.: Книга, 1978. С. 103–104.
2. Акашева А. А. Исторический источник в цифровой среде – издания первой всеобщей переписи населения 1897 г. на сайтах ведущих библиотек России / А. А. Акашева, А. С. Кулакова // Историко-культурное наследие в цифровом измерении: матер. Междунар. НК (Пермь, 2021). Пермь: ПГНИУ, 2021. С. 66.
3. Акашева А. А. Вопросы электронной каталогизации и сканирования фондов региональных и федеральных библиотек России. На примере Первой всеобщей переписи населения 1897 г. / А. А. Акашева // Румянцевские чтения – 2022: матер. Междунар. НПК в 2 ч. Ч. 1. С. 5–9.
4. Акашева А. А. Издания Первой всеобщей переписи населения Российской империи 1897 г.: эл. табл. / А. А. Акашева, А. С. Кулакова. Нижний Новгород: ИМОМИ ННГУ, 2023. URL: imomi.unn.ru/about/personalii/361-akasheva-anna-anatolevna.
5. Акашева А. А. Перепись 1897 г. Ход издания: интерактивный таймлайн / А. А. Акашева. Версия 1.1; релиз 27.06.2023. Нижний Новгород: ИМОМИ ННГУ, 2022. URL: clck.ru/34npgw.
6. Акашева А. А. Изучение хода публикации материалов переписи населения Российской империи 1897 г. с помощью веб-приложения «TimelineJS» / А. А. Акашева // Новые концепции и технологии исследований в междисциплинарных областях истории и культуры: сб. ст. междунар. науч. симпозиума. Пермь: ПГНИУ, 2022. С. 5–12.
7. Manovich L. Cultural Analytics / L. Manovich. Cambridge: The MIT Press, 2020. P. 224. URL: manovich.net/index.php/projects/cultural-analytics.
8. Акашева А. А. Книжные издания всероссийской переписи населения 1897 г. как первый опыт широкомасштабной публикации данных в государственной статистике. Источниковедческое изучение средствами культурной аналитики / А. А. Акашева // История. 2023. № 6.

УДК 930.2:004

Л. Н. Водолажская

Крымский федеральный университет им. В. И. Вернадского, Симферополь, Россия

e-mail: larvodol@gmail.com

ORCID: 0000-0002-2588-7002

ТРАНСДИСЦИПЛИНАРНЫЙ КУРС «ЦИФРОВЫЕ И ЕСТЕСТВЕННО-НАУЧНЫЕ МЕТОДЫ В АРХЕОЛОГИИ»

Аннотация. В статье рассматривается новый трансдисциплинарный курс «Цифровые и естественно-научные методы в археологии», разработанный на кафедре информатики Физико-технического института Крымского федерального университета и внедрённый в учебный процесс в 2022/23 уч. г. Показана его актуальность, а также связь с одной из приоритетных отраслей экономики – туристической индустрией и сопряжёнными областями. В статье описывается структура курса, его содержание и темы. Отмечено значение освоения данного факультативного курса для более эффективного обучения студентов-гуманитариев и для внедрения цифровых методов в гуманитарные: археологические и культурологические исследования. Подчёркивается, что программа трансдисциплинарного курса была разработана сотрудником, имеющим второе высшее – историческое/археологическое образование, а также опыт работы в музеях и археологических экспедициях, что позволило разработать программу курса, максимально ориентированную на актуальные практические задачи.

Ключевые слова: трансдисциплинарный, методы, цифровые, гуманитарные, естественно-научные, археология, музей, программное обеспечение

В последние годы трансдисциплинарный подход, предполагающий интеграцию знаний и методов из нескольких дисциплин для решения сложных социальных, исторических и т. п. проблем, привлекает значительное внимание в гуманитарных исследованиях и гуманитарном образовании. Под трансдисциплинарным подходом в гуманитарных исследованиях обычно понимается интеграция социальных, гуманитарных наук, искусства и т. п. В рамках *Digital Humanities* под трансдисциплинарностью понимается интеграция различных гуманитарных дисциплин с информатикой. Цифровые гуманитарные науки – *Digital Humanities* – характеризуются применением компьютерных методов в гуманитарных областях, включая историю и археологию [1; 2]. Однако в современном информационном обществе в последнее время в различных областях, включая гуманитарные исследования и гуманитарное образование, стала возрастать роль комплексных – цифровых и естественно-научных методов. Развитие технологий и доступ к большим объёмам данных, полученных в т. ч. с по-

мощью естественно-научных методов, создают новые возможности для исследования и понимания гуманитарных проблем. Комплексное использование цифровых и естественно-научных методов позволяет расширить границы гуманитарных дисциплин, обогатить их анализ и предоставить новые инструменты для исследования сложных явлений и проблем. Интеграция цифровых и естественно-научных методов в рамках гуманитарных исследований порождает качественно новую трансдисциплинарность *Digital Humanities* и открывает уникальные возможности для более глубокого понимания истории и культуры нашей цивилизации.

Современные исследования историко-археологических памятников и музейных предметов невозможны без применения как новейших цифровых технологий, так и естественно-научных методов. С помощью современного оборудования получают большие массивы данных, характеризующих различные параметры объектов исследования – например, археологических находок и музейных предметов, являющихся классическими объектами гуманитарных исследований. Полученные данные нуждаются в обработке и анализе цифровыми методами. Таким образом, цифровые гуманитарные науки, имеющие отношение к истории, археологии и музейному делу, во многих случаях одновременно связаны и с естественно-научными методами. Следовательно, для более эффективного внедрения цифровых методов в археологические и культурологические исследования необходимо внедрять их одновременно с естественно-научными методами. Для такого внедрения необходим качественно другой уровень образования археологов и культурологов, которые, обладая большим объёмом гуманитарных знаний и в лучшем случае базовым объёмом знаний в области математики и информатики, не имеют достаточной подготовки в области естественных наук.

Республика Крым является регионом с большим туристическим потенциалом. Туристическая деятельность и деятельность в области туристической индустрии в целях развития внутреннего туризма, к которым относится деятельность библиотек, архивов, музеев и прочих объектов культуры, значится в перечне приоритетных отраслей экономики. Характерной особенностью туристического сектора Крыма является наличие большого количества историко-археологических памятников и музеев на полуострове. Всего в Крыму насчитывается около 400 музеев. Первую по численности группу музеев составляют исторические и историко-археологические – около 70 %, вторую – около 20 % – краеведческие музеи. Из них 27 крупных государственных музеев, из которых 15 (8 заповедников, 7 музеев) находятся в республиканском ведении и 12 – в муниципальном. Во всех исторических и краеведческих музеях республиканского ведения хранятся археологические коллекции. Кроме того, в Крыму находятся историко-археологические музеи-заповедники:

- музей-заповедник «Херсонес Таврический»;
- Восточно-Крымский историко-культурный музей-заповедник;
- музей-заповедник «Судакская крепость»;

- историко-археологический музей-заповедник «Неаполь Скифский»;
- историко-археологический музей-заповедник «Калос Лимен»;
- Бахчисарайский историко-культурный и археологический музей-заповедник.

В Крымском федеральном университете им. В. И. Вернадского (КФУ) готовят высококвалифицированных специалистов – историков, археологов и культурологов для исследования этих памятников и работы в музеях, экскурсионных бюро и других организациях туристического сектора экономики.

Хороший уровень исследований историко-археологических памятников и музейных предметов в принципе невозможен без применения естественно-научных методов и новейших цифровых технологий. Однако до последнего времени в КФУ отсутствовал учебный курс, в рамках которого студенты изучали бы как цифровые, так и естественно-научные методы в историко-археологических исследованиях и музейном деле.

С целью ознакомления студентов с применением цифровых и естественно-научных методов в гуманитарных областях и туристической сфере на кафедре информатики Физико-технического института КФУ был разработан и в 2023 г. внедрён в преподавательский процесс новый междисциплинарный факультативный курс – «Цифровые и естественно-научные методы в археологии». Он был предназначен как для студентов-гуманитариев, так и для студентов, обучающихся по естественно-научным направлениям. Программа курса была разработана сотрудником кафедры информатики, имеющим второе – историческое/археологическое образование, опыт работы в музеях и археологических экспедициях.

В рамках курса рассматриваются основные задачи, которые необходимо решать в процессе историко-археологических исследований, а также естественно-научные методы и соответствующая приборная база, используемые в процессе решения этих задач. Студентов знакомят с цифровыми методами обработки информации, получаемой с помощью этих приборов, а также с задачами цифровой археологии и программным обеспечением, используемым для их решения. Дополнительно более подробно рассматриваются цифровые астрономо-геодезические методы в археологических исследованиях, а также цифровые технологии и программное обеспечение, используемое в музеях и для реконструкции объектов культурного наследия.

Курс состоит из разделов, включающих следующие темы (табл. 1).

Таким образом, в рамках разработанного курса студенты получают представление о реальных задачах в области археологии и музейного дела, а также современных как цифровых, так и естественно-научных методах и технологиях, применяемых для их решения.

Успешное внедрение в учебный процесс этого факультативного курса позволило начать разработку курса более широкой направленности, ориентированного не только на археологические, но и на гуманитарные исследования различного профиля.

Темы разделов

Раздел	Темы, дидактические единицы
Информатика и цифровые методы и технологии в гуманитарных исследованиях	Информатика, её методы и технологии. Области и разделы информатики. Прикладная информатика в конкретных научных областях. Цифровые гуманитарные науки. Историческая информатика и цифровая история. Направления исторической информатики и цифровой истории. Цифровая археология как составная часть исторической информатики и цифровой истории
Археологические объекты исследований	Археология как историческая наука. Хронологические периоды в археологии. Типы археологических памятников. Памятники подводной археологии. Археологические источники. Типы предметных источников (артефактов). Примеры археологических памятников Крыма различных археологических периодов
Цифровые методы в археологических исследованиях	Цифровая археология и цифровые технологии в археологии. Основные задачи цифровой и вычислительной археологии. Программное обеспечение, используемое при решении задач цифровой и вычислительной археологии
Естественно-научные методы в археологических исследованиях	Взаимосвязь современных цифровых и естественно-научных методов. Типы естественно-научных методов в археологических исследованиях и задачи, которые решают с их помощью. Полевые и лабораторные исследования в археологии с помощью комплексных – цифровых и естественно-научных методов. Оборудование, используемое в прикладных комплексных археологических исследованиях. Примеры исследований археологических памятников и артефактов с помощью цифровых и естественно-научных методов
Цифровые методы в астрономо-геодезических исследованиях археологических памятников	Программное обеспечение, используемое в астрономо-геодезических исследованиях археологических памятников (археoaстрономические исследования). Примеры применения астрономо-геодезических методов в исследованиях древних памятников Северного Причерноморья

Список источников

1. Бородкин Л. И. Digital History: применение цифровых медиа в сохранении историко-культурного наследия? / Л. И. Бородкин // Историческая информатика. 2012. Т. 1. № 1. С. 14–21.

2. Володин А. Ю. Digital Humanities (цифровые гуманитарные науки): в поисках самоопределения / А. Ю. Володин // Вестник ПГНИУ. Сер.: История. 2014. № 3 (26). С. 5–12.

В. В. Жила, М. А. Михайлова, Е. А. Ермолаев, Я. В. Кочнева
Сибирский федеральный университет, Красноярск, Россия

ЦИФРОВОЕ ГУМАНИТАРНОЕ ОБУЧЕНИЕ КАК ОСНОВА РАЗРАБОТКИ УЧЕБНЫХ ПРОГРАММ

Аннотация. Цифровые гуманитарные проекты и методы становятся всё более распространёнными в академической среде гуманитарных наук. Цифровые проекты позволяют студентам использовать новые технологии, создавать осязаемые научные результаты. Лаборатория образования и развития цифровых исследований (*LEADR*) – новая инициатива в области цифровых гуманитарных наук. Использование подобных цифровых проектов как части учебной программы гуманитарных наук является необходимым, чтобы научить студентов не только предметному контенту, но и необходимым навыкам получения информации, её анализу, медиаграмотности.

Ключевые слова: *цифровые гуманитарные проекты, LEADR, медиаграмотность, анализ*

В то время как многие представители высшего образования призывают к большему (и лучшему) использованию цифровой педагогики для студентов, сфера цифровых гуманитарных наук продолжает переживать значительный рост [3; 7]. Всё большее число университетов предлагают дополнительные специализации в области цифровых гуманитарных наук и аналогичных областей, и ещё больше университетов включают разработку цифровых проектов и критику существующих цифровых проектов в курсовые работы. Лаборатория образования и развития цифровых исследований стремится применить гуманитарное обоснование для развития у студентов критического анализа и развития веб-взаимодействия при создании мультимодальных проектов, визуализации данных и т. д.

LEADR представляет собой инициативу по разработке учебных программ, разработанную в результате сотрудничества факультетов истории и антропологии. Само пространство было спроектировано так, чтобы способствовать сотрудничеству и открытости в децентрализованном образовательном пространстве. Почти вся мебель является передвижной, что позволяет переключаться между организацией семинаров и групповой проектной работой, а также децентрализованными презентационными пространствами. *LEADR* также полностью оснащён большими экранами для групповой работы, машинами, камерами, аудио- и световыми комплектами, 3D-принтерами, микрокомпьютерным оборудованием и другими технологиями, облегчающими инновационную цифровую работу. Сотрудники *LEADR* (два штатных преподавателя и три аспиранта факультетов истории и антропологии) помогают преподавателям в разработке заданий, разрабатывают и проводят обучающие и дискуссионные занятия, ориенти-

рованные на цифровые технологии, оказывают помощь в разработке руководств по выставлению оценок. Лаборатория открыта почти 50 ч в неделю. Компонент *LEADR* для курсов аналогичен использованию лабораторных работ во многих университетах – это возможность для студентов взять содержание и методы, которые они прочитали и обсудили, и применить их на практике, где они могут создавать и экспериментировать.

Структура, используемая в *LEADR*, состоит из четырёх гибких целей обучения, которые можно объединить и развивать в соответствии с потребностями курса и желаемыми результатами преподавателя и партнёров. Цели: 1) информационная грамотность; 2) цифровая грамотность; 3) грамотность в области данных; 4) вычислительный анализ.

Студентам *LEADR* часто приходится разрабатывать цифровые проекты, включающие цифровые первоисточники, в т. ч. цифровые коллекции и веб-сайты, основанные на повествовании. Использование цифровых объектов может варьироваться от нескольких фотографий, используемых для иллюстрации основных понятий в письменной форме, до создания выставок, основанных на больших коллекциях оцифрованных документов, артефактов и фотографий. Поскольку обнаружение и оценка первоисточников является важнейшим компонентом почти всех проектов, обучение информационной грамотности в сочетании с *LEADR* часто начинается с особого внимания к поиску, оценке и пониманию цифровой информации.

Критическое производство и потребление цифровых медиа становится гораздо более эффективным при понимании технических принципов и теоретических возможностей медиа. Цифровое общение важно не только в дисциплинарном смысле, поскольку исследования в области истории и антропологии всё чаще производятся в цифровом формате, но и как необходимый навык для участия в жизни гражданского общества. В *LEADR* студенты учатся писать для веб-аудитории различными способами: определять аудиторию и пользователей, использовать соответствующую риторику, структурировать повествование для эффективной работы в сети, интегрировать мультимедиа и использовать *HTML/CSS* или систему управления контентом (*CMS*) для публикации и распределения стипендий. Критическая оценка «цифровости» объекта способствует лучшему использованию возможностей средств массовой информации, а также пониманию внешних факторов: этические и юридические права, конфиденциальность, постоянство, непостоянство, приём аудитории и обратная связь.

Метаграмотность, расширение традиционных информационных навыков играют ключевую роль в обучении студентов создавать и ремиксовать цифровой контент [5]. Авторы расширили традиционные понятия, включив в них «совместное производство и обмен информацией в коллективной цифровой среде», что особенно важно, т. к. это «требует постоянной адаптации к новым технологиям и понимания критического мышления и размышлений, необходимых для участия в этих пространствах» [5].

Технологии меняются, и проприетарные платформы, которые упрощают производство и публикацию, требуют постоянных денежных вложений и, вероятнее всего, исчезнут, заменив новые. Это означает, что необходимо сосредоточить внимание на общении с использованием открытых, устойчивых инструментов, которые обеспечивают полную свободу и владение контентом, включая производство контента в плоских *HTML*-файлах, *CMS* с открытым исходным кодом.

Поскольку термины «открытые данные» и «большие данные» становятся всё более распространёнными, способность получать доступ, уточнять, манипулировать, критически оценивать и обмениваться данными становится всё более важным навыком. Обучение часто происходит на общеуниверситетских семинарах, кратких семинарах для обучения исследователей методам курирования данных или, возможно, на семинарах для преподавателей, сотрудников и аспирантов [1]. Подобные занятия дают возможность устанавливать стандарты метаданных на ранних этапах исследовательского процесса, документировать обработку данных и использовать устойчивое форматирование среди других практик курирования, но часто они оторваны от контента и практик сообщества [2]. Вместо этого преподавание принципов работы с данными на курсах позволяет уделять больше внимания использованию данных с полным учётом контекста и потенциальных вопросов, которые могут быть заданы к данным.

Информационная грамотность в отношении данных – это основа, предназначенная для интеграции в курсы и процессы последипломных исследований в контексте предметных данных и практики сообщества на основе предметной области. *DIL (Data Information Literacy)* состоит из 12 основных этапов, которые включают в себя: введение в базу данных, форматы данных; обнаружение и сбор данных; управление данными и организация, преобразование данных и обеспечение качества совместимости; метаданные; практику курирования данных и повторного использования; сохранение данных; анализ данных; визуализацию данных; этику, включая цитирование данных.

Целевое сообщество, занимающееся информационной грамотностью, обычно нацелено на преподавателей, академических сотрудников, аспирантов, научных сотрудников [2; 6]. Важно, чтобы эти ценные навыки и базовые понятия широко преподавались на уровне бакалавриата, среди студентов, в частности гуманитарных направлений. Навыки работы с данными и грамотность приобретают решающее значение, поскольку такие термины, как «большие данные», становятся всё более и более распространёнными в темах научных исследованиях, а организации всё чаще делают данные открытыми во имя прозрачности. Таким образом, без навыков доступа к данным и их оценки могут возникнуть серьёзные препятствия для их успешного использования и анализа.

Использование этой структуры, как отмечают исследователи, в течение учебного года привело к ряду изменений в способах концептуализации, структурирования и преподавания курсов. В целом эти изменения способствовали большему количеству партнёрских курсов, более захватывающим и увлекательным заданиям и мероприятиям, оценке, которая фокусируется на грамотности и размышлениях, а не на конечном продукте.

На административном уровне структура позволит добиться этих улучшений несколькими способами.

Во-первых, перенос фокуса задания с результата на процесс и практику. Существует ряд инструментов, которые предлагают элегантные конечные интерфейсы, но мало возможностей для настройки и ограничивают фактическое изучение публикаций, методов исследования и лежащих в их основе технологий. В первую очередь следует рассмотреть область, метод и технические цели, после чего можно соответствующим образом определить процесс и ожидания. Может возникнуть соблазн сформулировать цели обучения, а затем перейти к самому простому решению, чтобы получить конечный результат, отказываясь или частично отказываясь от целей обучения в процессе. Структура помогает нам помнить об этом, а также формирует критерии оценивания, в которых демонстрация методов и навыков важнее быстрого достижения презентабельного результата.

Во-вторых, эта структура полезна для привлечения студентов, которые плохо знакомы с цифровыми исследованиями, или даже тех, кто скептически относится к их ценностям. Данный подход позволяет намного легче выйти за рамки разговоров, в которых слишком много внимания уделяется результатам или технологиям, и вместо этого сосредоточиться на обучении, которое происходит в ходе разработки проекта. Этот тип рефрейминга может быть очевиден для многих, кто занимается цифровыми гуманитарными науками, но может иметь большое значение для того, чтобы сделать цифровые методы более интересными, ценными и доступными для тех, кто является новичками в этой области.

В-третьих, структура служит отличным способом для сотрудничества между преподавателями, студентами, аспирантами, специалистами в области образовательных технологий. Научный диалог между специалистами в предметной области с ограниченными цифровыми навыками и программистами, которые имеют небольшой опыт работы в предметной области, развивается быстрее. Так, стороны могут гораздо легче определить, где их научный опыт может быть ценным для процесса обучения.

Список источников

1. Carlson J. Determining Data Information Literacy Needs: a Study of Students and Research Faculty / J. Carlson, M. Fosmire, C. C. Miller et al. // *Libraries and the Academy*. 2011. No. 11 (2). Pp. 629–657.

2. Carlson J. Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers / J. Carlson, L. Johnston // *The Journal of Academic Librarianship*. 2015. No. 15 (2). Pp. 113–159.

3. Hirsch B. D. *Digital Humanities Pedagogy: Practices, Principles and Politics* / B. D. Hirsch // Open Book Publishers. 2012. 450 p.
4. Locke B. T. *Digital Humanities Pedagogy as Essential Liberal Education: a Framework for Curriculum Development* / B. T. Locke // *Digital Humanities Quarterly*. 2017. No. 11 (3). Pp. 15–24.
5. Mackey T. P. *Metaliteracy: Reinventing Information Literacy to Empower Learners* / T. P. Mackey, T. E. Jacobson. 2014. 248 p.
6. Shorish Y. *Data Information Literacy and Undergraduates: a Critical Competency* / Y. Shorish. 2015. No. 22 (1). Pp. 97–106.
7. Thomas W. *The Other End of the Scale: Rethinking the Digital Experience in Higher Education* / W. Thomas, E. Lorang // *Educause Review*. 2014. No. 49 (5). Pp. 59–67.

З. Р. Ишанходжаева,
доктор исторических наук, профессор
Национальный университет Узбекистана им. М. Улугбека, Ташкент, Узбекистан
e-mail: zamira.i@rambler.ru

ЦИФРОВЫЕ ТЕХНОЛОГИИ В ВЫСШЕМ ИСТОРИЧЕСКОМ ОБРАЗОВАНИИ В УЗБЕКИСТАНЕ

Аннотация. Статья показывает поэтапный процесс цифровизации высшего гуманитарного образования в Узбекистане, использование цифровых технологий в исторической науке, анализируется опыт, достижения и перспективы развития цифровизации гуманитарных дисциплин.

Ключевые слова: образование, технологии, междисциплинарный профиль, цифровая история, цифровая гуманитаристика, электронная платформа, электронные ресурсы, 3D-реконструкция, визуализация, виртуальная реальность, дополненная реальность

Активное внедрение ИТ-технологий в высшее образование в последние годы стало наиболее востребованным, т. к. на рынке труда во всех странах мира усилился запрос на специалистов данного направления. Это актуализирует проблему подготовки и привлечения высококвалифицированных специалистов междисциплинарного профиля, что является одной из сложных задач не только для работодателя, но и государств в целом [1]. В Узбекистане реализуются комплексные меры по активному развитию цифровой экономики, а также широкому внедрению современных информационно-коммуникационных технологий во все отрасли и сферы, прежде всего в государственное управление, образование, здравоохранение и сельское хозяйство. В этом русле была одобрена стратегия «Цифровой Узбекистан – 2030», принятая 07.10.2020, разработанная Министерством по развитию информационных технологий и коммуникаций при участии заинтересованных министерств и ведомств, представителей бизнес-сообщества и научных кругов, а также зарубежных экспертов [2]. Также приняты отраслевые программы по цифровизации конкретных направлений народного хозяйства. Утверждённая в рамках стратегии «Дорожная карта» на 2020–2022 гг. предусматривала развитие четырёх ключевых сфер – электронного правительства, цифровой индустрии, цифрового образования, цифровой инфраструктуры [3].

Сегодня в Узбекистане высшее образование характеризуется устойчивыми тенденциями к информатизации и цифровым ресурсам, активным слиянием образовательных и информационных технологий на всех уровнях, формированием на этой основе принципиально новых инновационных, интегрированных технологий обучения. Внедрение цифровых техно-

логий и искусственного интеллекта в историческое образование является актуальной задачей, решающей вопросы перехода к информационной цивилизации [4]. Использование математических методов в исторических исследованиях убедительно показано и в трудах известных зарубежных учёных, таких как Л. И. Бородкин и др. [5].

Создание информационной образовательной среды обеспечивает доступность и качество системы образования. В этой связи как один из примеров можно рассмотреть электронную образовательную платформу Национального университета Узбекистана *Hemis*, которая предназначена для размещения учебных материалов для студентов всех направлений бакалавриата очной, заочной и вечерней форм обучения, а также магистратуры. Коллективом университета разработаны, созданы и внедрены в электронный учебный процесс все необходимые учебные материалы – лекции, презентации, задания в виде тестов и индивидуальных тем самостоятельных работ, тематики курсовых работ и рефератов, итоговых выпускных работ для каждого студента, вопросы семинарских занятий, видеоролики, список литературы и др. Например, по предмету «Научное наследие мыслителей Востока» на платформе размещены тексты лекций, презентации, тестовые задания по темам курса, планы семинаров, тематика самостоятельных работ, видеозаписи и список литературы. При помощи оцифрованных документов, миниатюр и фрагментов трудов учёных разъясняется роль учёных-энциклопедистов в становлении образования и науки эпохи Возрождения и их плодотворное влияние на развитие средневековой научной мысли в странах Востока и Запада. Таким образом, при изучении данной дисциплины используются цифровые ресурсы – исторические карты, миниатюры, тексты произведений средневековых учёных, исторические фото- и видеоматериалы, а также студентам предлагается составлять таблицы, диаграммы на основе данного учебного материала [6].

Одним из основных факторов, определяющих успех цифровых технологий в историческом образовании, является способность использовать их для более глубокого понимания материала, а также для развития критического мышления. В контексте исторического образования цифровые технологии и искусственный интеллект могут быть использованы для создания интерактивных занятий и учебных пособий, которые могут помочь преподавателям и студентам лучше понимать богатое культурно-историческое наследие, его значение и связь с современностью. Динамичное и бурное развитие эпохи цифровизации насытило эту среду многочисленными гаджетами, что расширило само понятие «интерактивное обучение». На сегодняшний день интерактивными можно назвать и подходы в обучении, и формы, и технологию, и методы, и средства обучения [7].

Один из ярких примеров использования цифровых технологий и искусственного интеллекта в историческом образовании – создание 3D-моделей исторических мест и событий. Такие модели могут воссоздавать архитектуру, жизнь и повседневные аспекты времени, что даёт студентам возможность «перенестись во времени и пространстве» и получить

более глубокое понимание того, что происходило в конкретный исторический период. На сегодняшний день имеются 3D-реконструкции многих археологических памятников историко-культурного наследия Узбекистана. Например, чат-бот *Boburnama* (URL: t.me/boburnama) функционирует на базе исторического источника – труда государственного деятеля и основателя империи Бабуридов – Захириддина Мухаммада Бабура «Бабурнамэ». Бот может не только воспроизвести рельеф местности, описанный в книжном варианте «Бабурнамэ», но и визуализировать некоторые исторические события, описанные в этом источнике, – например, такие как Битва при Анкаре 1402 г., а также внешний вид исторических личностей, упомянутых в источнике. Кроме того, цифровые технологии и искусственный интеллект могут быть использованы для создания учебных игр, которые могут повысить интерес студентов к истории и выработать у них умение переключаться между детальным и всесторонним изучением материала. Например, с помощью онлайн-игр студенты могут пройти через различные хронологические этапы развития цивилизаций, научиться решать сложные исторические задачи и делать выводы на основании полученных знаний.

Цифровизация истории является одним из приоритетных направлений в развитии исторического образования в Узбекистане. В последние годы в стране были сделаны значительные шаги в этом направлении – при участии государственных органов, образовательных институтов и частных компаний были запущены национальные проекты по цифровизации контента. Среди многих таких ресурсов можно назвать электронную платформу *Raqamli Tarix* (URL: raqamlitarix.uz) [8], подготовленную в 2022 г. на историческом факультете Национального университета Узбекистана. Платформа содержит документы, видео-, аудио- и фотоматериалы по истории репрессивной политики советской власти в 1925–1950 гг. в Узбекистане. Ещё один из таких проектов – «Виртуальный музей Ташкента», который был запущен в 2018 г. Он содержит более 2,5 тыс. экспонатов, воссоздаёт исторические здания, а также предлагает виртуальные экскурсии по разным районам Ташкента, позволяя увидеть город с разных точек. Кроме того, онлайн-коллекции национальных музеев стали доступны к просмотру для широкой аудитории [9]. Также заслуживают внимания работы архитектора Т. Нурулина – 3D-реконструкции памятников архитектуры древнего Ташкента (URL: civiltashkent.blogspot.com/?m).

Таким образом, можно сделать вывод о том, что современные информационные технологии, использование искусственного интеллекта в высшем историческом образовании расширяют возможности и методы передачи и распространения знаний, формирования на их основе необходимых компетенций, управления образовательным процессом и обеспечения доступа к качественному историческому образованию.

Список источников

1. Высшее образование в эпоху цифровизации: развивать и совершенствовать // Янги Ўзбекистон: газета. URL: yuz.uz/ru/news/vsshee-obrazovanie-v-epoxu-tsifrovizatsii-gazvivat-i-sovershenstvovat.
2. Об утверждении стратегии «Цифровой Узбекистан – 2030» и мерах по её эффективной реализации: Указ Президента Республики Узбекистан № УП-6079 от 05.10.2020 // Lex.uz: национальная база данных законодательства Республики Узбекистан. URL: lex.uz/docs/5031048.
3. Принята стратегия «Цифровой Узбекистан – 2030» // Norma.uz: инф.-прав. портал. URL: norma.uz/novoe_vzakonodatelstve/prinyata_strategiya_cifrovoy_uzbekistan_-_2030.
4. Ватунский А. А. Современные инновационные технологии в профессиональном образовании / А. А. Ватунский // КиберЛенинка: науч. эл. библиотека. URL: cyberleninka.ru/article/n.
5. Бородкин Л. И. Моделирование исторических процессов: от реконструкции реальности к анализу альтернатив / Л. И. Бородкин. СПб.: Алетейя, 2016. С. 5. URL: hist.msu.ru/about/gen_news/23015.
6. Ишанходжаева З. Р. Использование цифровых ресурсов в историческом образовании / З. Р. Ишанходжаева // Исторические исследования в контексте науки о данных. Информационные ресурсы, аналитические методы и цифровые технологии: матер. междунар. конф. (М., 2020).
7. Коротаева Е. В. Интерактивное обучение: аспекты теории, методики, практики / Е. В. Коротаева, А. С. Андрюнина // КиберЛенинка: науч. эл. библиотека. URL: cyberleninka.ru/article/n/interaktivnoe-obuchenie-aspekty-teorii-metodiki-praktiki.
8. Электронная платформа исторического факультета НУУз им. М. Улугбека. URL: raqamlitarix.uz.
9. VRmuseum: посетите музеи, не выходя из дома. URL: vrmuseum.uz/ru.

УДК 930

Е. А. Брюханова¹, М. В. Рыгалова²

^{1,2} кандидат исторических наук

^{1,2} Алтайский государственный университет, Барнаул, Россия

² Алтайский государственный технический университет
им. И. И. Ползунова, Барнаул, Россия

e-mail: ¹ elena@hist.asu.ru, ² mariya_rygalova@mail.ru

ORCID: ¹ 0000-0002-6974-5065, ² 0000-0002-3715-3516

СПРАВОЧНЫЕ ИЗДАНИЯ КАК ИСТОЧНИКИ ПО ГОРОДСКОМУ НАСЕЛЕНИЮ СИБИРИ КОНЦА XIX – НАЧАЛА XX ВЕКА

Аннотация. Рассматриваются возможности использования печатных справочных изданий в изучении населения городов Сибири конца XIX – начала XX в. Особое внимание уделяется адрес-календарям, издававшимся центральными и губернскими статистическими учреждениями практически ежегодно. Это позволяет собрать исчерпывающие персональные данные (Ф. И. О.) служащих различного уровня, проследить их карьерный путь, участие в различных общественных организациях. Справочные издания позволяют уточнить, дополнить информацию переписных листов Первой всеобщей переписи населения 1897 г., создать комплексный ресурс по городскому населению Сибири.

Ключевые слова: города Сибири, справочные издания, адрес-календарь, городское население

Сегодня справочные издания, содержащие широкий круг сведений, всё чаще становятся основным источником в исторических исследованиях. Это связано со слабой изученностью многих из них, а также техническими возможностями работы с большими объёмами данных, их анализом. Печатные издания конца XIX – начала XX в. содержат большой пласт информации по различным направлениям социально-экономического развития общества.

Одним из источников этой группы являются адрес-календари – официальные издания, которые публиковались как на общероссийском уровне, так и на уровне губерний, как правило, ежегодно в формате книжного издания. В них содержатся пофамильные сведения о разных группах городского населения Сибири, с указанием на род занятий, должность, чин.

Адрес-календарь на центральном уровне издавался в Санкт-Петербурге в типографии Сената под заглавием «Общая роспись всех чиновных особ в Российской империи». Он содержит персонифицированные сведения (Ф. И. О., звание, чин, должность) о придворных служащих, а также служащих государственных учреждений гражданского, военного и духовного ведомств центральных и губернских, данные о лицах, занимающих должности в крупных учебных, благотворительных и прочих учреждениях. В первой части издания публиковались списки чиновников общероссийских ведомств, во второй – данные по губерниям [4].

В связи с ростом государственного аппарата и развитием губернской статистики получили развитие местные издания адрес-календарей. Источниковедческий анализ позволяет сделать вывод о том, что в губернских адрес-календарях содержится более детальная информация о служащих, даже низшего звена, в то время как в центральных – об основном составе высокопоставленных лиц. Внешней особенностью источника является его нумерация: оглавление пронумеровано римскими цифрами, каждый раздел – «Календарные сведения», «Справочный отдел», собственно «Адрес-календарь» – имеет собственную, а не продолжающуюся нумерацию арабскими цифрами.

Адрес-календари содержат такие сведения о служащих, как: Ф. И. О., место работы (учреждение и структурное подразделение), должность, чин, звание, при их отсутствии сословие. Можно встретить адрес-календари, в которых отдельным разделом печатались сведения о наградах [2, с. 38]. Кроме того, издания за некоторые годы содержат более развернутые сведения. Например, в адрес-календаре Тобольской губернии за 1897 г. опубликованы данные по торговым предприятиям и их владельцам [1].

Исследователи отмечают, что адрес-календари являются высокоинформативными источниками по изучению административно-территориального деления, истории переселения, судебной системы, образования, медицины, торговли, промышленности и других сфер жизни общества [12, с. 108–109]. Кроме того, по ним можно отслеживать жизненный путь отдельных личностей, выявить информацию о работе служащих в нескольких учреждениях, участии их в научных и благотворительных обществах. В качестве примера можно привести Александра Николаевича Балакшина, купца из Кургана. В адрес-календаре Тобольской губернии за 1904 г. информация о нём встречается несколько раз. Так, он входит в общество по крестьянским делам Кургана, является почётным мировым судьёй Тобольского окружного суда, а также значится в уездном отделении Тобольского епархиального училищного совета и является наблюдателем метеорологической станции в д. Старо-Сидорова Курганского уезда [3, с. 4, 30, 62, 70]. Таким образом, в адрес-календарях можно найти сведения не только о государственных и общественных служащих, но и о владельцах торговых фирм, фабрик и заводов, участниках и попечителях благотворительных обществ и др.

Кроме непосредственно общероссийских и губернских адрес-календарей, сведения о персональном составе служащих публиковались в отраслевых ведомственных изданиях, к которым относятся «Списки военных и гражданских чинов по классам» [13], с указанием Ф. И. О., должности, чина, даты получения чина служащих военного ведомства. По военному ведомству также выходили региональные издания – например, «Справочные книжки о должностных лицах правительственных и общественных установлений Омского военного округа и степного генерал-губернаторства», в которых представлены сведения о начальственных чинах в Омском военном округе, а также о должностных лицах войсковых учреждений с указанием должности и Ф. И. О. по губерниям [15].

Ещё одним отраслевым изданием являются «Списки чинам министерства земледелия и государственных имуществ», издававшиеся непосредственно этим ведомством [14]. Поимённые списки дворян публиковались в «Дворянском адрес-календаре» [6]. Епархиальные адрес-календари издавались по инициативе региональных Епархий и публиковали списки лиц духовного ведомства (Ф. И. О., данные о сане, должности, образовании, времени пребывания на службе, а также периодически сведения о жаловании и наградах) [16, с. 36–46], а также списки получающих пособие для бедных священнослужителей и пенсии для вдов и детей священнослужителей [17, с. 21–24].

К числу ведомственных изданий относятся памятные книжки Западно-Сибирского учебного округа, в которых опубликованы списки учебных заведений по категориям, данные по служащим (должность, дата начала службы, образование, награды, жалование, вероисповедание) [8]. Таким образом, из источника можно как извлечь персональные сведения о служащих, траекторию их карьерного роста, так и сформировать представления о развитии сети учебных заведений.

Ещё одним ведомственным источником являются торгово-промышленные адрес-календари Российской империи [5], в которых опубликованы сведения о торговых предприятиях и их владельцах с разделением на отрасли промышленности, а также региональное издание Сибирского торгово-промышленного календаря. В нём можно найти сведения о владельцах торговых предприятий по городам Сибири, а также данные о градоначальствующих лицах, служащих правительственных, общественных и частных учреждений г. Томска и некоторых других городов, которые публиковались на непостоянной основе [9].

Сведения Сибирского торгово-промышленного календаря о представителях промышленности и торговли, публиковавшиеся ежегодно, позволяют проследить как развитие дел отдельных лиц, так и динамику указанных сфер в целом. Например, в г. Тара в 1897 г. Алексей Михайлович Щербаков значится как купец 1-й гильдии, имеющий винокуренный завод, склад и ренсковый погреб. Согласно торгово-промышленному календарю 1905 г. Щербаков значится уже как купец 2-й гильдии, имеющий только

винокурный завод. Кроме того, в 1897 г. в Таре было два купца 1-й гильдии, в 1905 г. – ни одного [10, с. 635; 11, с. 235].

Таким образом, центральные и губернские печатные справочные издания отражают сведения о персональном составе населения городов. Это позволяет отслеживать карьерный путь и продвижения по социальной лестнице. Списки населения различных категорий представляют интерес для создания номинативной базы данных. Создана и наполняется база данных «Городское население Сибири», в профиле которой объединены записи по персоналиям, которые идентифицированы, как один и тот же человек на основе разных источников, что позволяет собрать исчерпывающую информацию не только о должности, но и об участии в составе общественных, благотворительных учреждений, отследить включённость в социальную жизнь города. Также в базе данных имеются функции отображения людей, которые, вероятно, похожи (например, совпадают Ф. И. О.), но требуется дополнительная уточняющая информация для полной идентификации или установления отсутствия связей. В настоящее время доступна информация о населении по материалам переписных листов [7].

Дальнейшая работа со справочными изданиями предполагает синхронизацию базы данных на основе печатных источников с переписными листами 1897 г. для тех городов, по которым они сохранились, что позволит существенно расширить данные о персоналиях (например, о составе их семьи, уровне образования и пр.).

Исследование выполнено за счёт гранта Российского научного фонда № 19-78-10020. URL: rscf.ru/project/19-78-10020.

Список источников

1. Адрес-календарь Тобольской губернии на 1897 г. Тобольск: Тобольская губернская тип., 1897. 217 с.
2. Адрес-календарь Тобольской губернии на 1900 г. Тобольск: Тобольская губернская тип., 1899. 170 с.
3. Адрес-календарь Тобольской губернии на 1904 г. Тобольск: Тобольская губернская тип., 1904. 128 с.
4. Адрес-календарь. Ч. 1–2: Общая роспись начальствующих и прочих должностных лиц по всем управлениям в Российской империи на 1897 г. СПб.: Тип. Правительствующего Сената, 1897. 1 013 с.
5. Вся Россия. Русская книга промышленности, торговли, сельского хозяйства и администрации: торгово-промышленный адрес-календарь Российской империи. СПб.: Тип. А. С. Суворина, 1895–1902. URL: elibr.shpl.ru/ru/nodes/24062-1899-god-t-1-1899.
6. Дворянский адрес-календарь на 1897 г. СПб.: Тип. высочайше утверждённого товарищества «Общественная польза», 1896. 400 с. URL: viewer.rusneb.ru/ru/000200_000018_v19_rc_1600294?page=401&rotate=0&theme=white.
7. Население Сибири на рубеже XIX–XX вв. URL: person1897.histcensus.asu.ru/households.
8. Памятная книжка Западно-Сибирского учебного округа, заключающая в себе список учебных заведений с указанием времени открытия, источников содержания, размера платы за ученье, числа учащихся и личного состава служащих. Томск: Тип. П. И. Макушина, 1897. 209 с.

9. Сибирский торгово-промышленный календарь на 1896 г. Томск: Тип. П. И. Макушина, 1896. 794 с.
10. Сибирский торгово-промышленный календарь на 1897 г. Томск: Тип. П. И. Макушина, 1897. 692 с.
11. Сибирский торгово-промышленный календарь на 1905 г. Томск: Тип. П. И. Макушина, 1905. 744 с.
12. Ситникова Т. В. Адрес-календари и справочники-календари Царицына рубежа XIX–XX вв. / Т. В. Ситникова // Известия СГУ. Сер.: Филология. Журналистика. 2018. Т. 18. № 1. С. 108–112.
13. Список военным и гражданским чинам первых двух классов по старшинству. СПб.: Тип. Правительствующего Сената, 1895. 73 с.

N. A. Matkin¹, N. A. Klyuev²,

² PhD in Linguistics

^{1,2} HSE University, Perm, Russia

e-mail: ¹ namatkin@hse.ru, ² nakliuev@hse.ru

ORCID: ¹ 0000-0003-3987-4445, ² 0000-0003-4829-4469

A CITY IN A MIRROR OF SOCIAL MEDIA: AUTOMATED ANALYSIS APPROACH TO URBAN REPRESENTATIONS

Abstract. This study presents a novel algorithm for analyzing the complex images of Russian cities as portrayed in virtual publics on social media. To achieve this goal, the authors employed several methods including topic modeling, affinity index, and sentiment analysis to investigate both official and informal communications on VK, a widely used social networking site in Russia. The approach utilizes a topic modeling algorithm to process a vast corpus of data and identify implicit themes. Affinity index is calculated to assess the relevance of each theme to the city. The algorithm defines the ratio of sentiment for each topic and each city. This algorithmic approach is expected to find applications in public communication management, where it can be used to highlight urban problems and inform content policies.

Keywords: *social media, image of the city, urban representations, topic modeling, sentiment analysis*

1. Introduction

Social media has become a widely used platform for local authorities to communicate information about urban changes to citizens, and to construct an image of the city through official communities on social media. These online communities serve as intermediaries between authorities and citizens, allowing individuals to respond to received messages in the form of comments, and revealing their own attitudes toward the city agenda. As a result, a city is represented both through official communication and through a reflection on it.

Authors have conducted a study comparing urban representations in official communications of local authorities, such as posts, and in the responses, they receive, such as comments. Social media provides a vast amount of city data that is accessible for gathering and analysis [12]. The automatization of gathering and analyzing social media materials provides opportunities to reveal changes in the image of a city, and to construct a city brand based on city communities' reactions.

2. Related Works

Media representations of Russian cities are well studied in media linguistics. There are specific works for cities-millionaires, such as Yekaterinburg and Kazan [11], Krasnoyarsk [6], Chelyabinsk [7], Perm [10]. The complex media image of 15 Russian cities is proposed in [15]. Despite the abundance of works, studies of the image of the city in comparison of official communication and citizens' reactions to it are insufficient.

Many researchers conduct an analysis of urban representations using open Twitter data [3; 4]. The main purpose of these works is to integrate the automatic tool in the smart city systems. Theoretical features of using big data in smart city systems are described in [5; 8; 9]. Practical usage of these systems with NLP methods is described in [1; 13].

In [13] author presents the algorithm for automating parsing and analyzing comments with the Twitter Streaming API, which includes parsing, text preprocessing, topic modeling and visualization. This algorithm is the basis of smart city systems with automatic sentiment analysis [1].

The integration of urban systems into everyday life has become a common phenomenon in many cities around the globe. However, the complex methodology required to study the image of the city in Russian social media remains largely unexplored. This research seeks to address this gap by proposing a new algorithm that considers the unique features of the Russian language and behavior. The focus of this research is not on spatial analysis but rather on the image of the city in communication. The authors propose a complex analysis of the images of various cities, including Volgograd, Voronezh, Yekaterinburg, Kazan, Krasnodar, Krasnoyarsk, Nizhniy Novgorod, Novosibirsk, Perm, Samara, Omsk, Rostov-on-Don, Ufa, and Chelyabinsk.

3. Data and Preprocessing

The present study investigates the image of the city in Russian social media, with a focus on 14 cities. The authors propose a new algorithm to explore the communication about these cities on VK, a popular social networking site in Russia. The authors collected 26 191 publications (25 904 posts after preprocessing) from January 1st to December 31st, 2021, using the VK API. Official communications from local authorities' official publics were selected for analysis, and the Wall.get method was utilized for parsing. The authors developed a stop-word list, which includes the most frequently used hashtags, links, and nicknames. The list was used to filter out irrelevant information.

For comment parsing, the authors compiled a list of the most active local communities. The Wall.get comments method was employed to collect high-level comments, yielding a corpus of 7 360 210 comments. The authors didn't perform bot filtration, as they aimed to investigate the natural communication environment. While strong language is not always censored by public administration, users often add different symbols between the first and last letters of words. These words can serve as markers for problem assessment.

To tokenize comments, the authors used a special pattern for regular expressions `([А-Яа-яА-з]+[^\,?!.\U00010000-\U0010ffff]+[А-Яа яАз]+|[\U00010000\U0010ffff][А-Яа-яА-з]+|\d+)` that preserves emojis and original punctuation within words. The pymorphy2 module was applied for lemmatization.

4. Topic Modeling

The authors employed a Latent Dirichlet allocation model to sort unobserved groups and identify the most valuable words in the topic [2]. Embeddings for topic modeling were built using the Bag of Words. The Bag of Words metrics were chosen with an extension of the stop words list.

The Coherence Score with CV and UCI metrics was calculated to identify the optimal number of topics. The optimal value for official communication was found to be 12 topics, whereas the optimal value for citizens' comments was 25 topics. Since it is hard to identify topic in comments less than 3 tokens, they are not included. The most probable topic was calculated for each publication, thereby revealing the most relevant topics for each city with an Affinity Index:

$$\textit{Affinity Index} = \frac{P}{P} \times 100.$$

Affinity Index is a ratio of the share of cluster in city to the share of cluster in full corpus. Index was calculated for topic with probability more than 30 %. Heatmaps displaying relevant topics in official and naive communication are presented in Fig. 2 and 3, respectively. The most salient markers determine the cluster name, although these names only approximate the underlying topics. For official communication, the following clusters were identified: house and communal services, COVID, development, disease, ecology, event, improvement, repair, sport, transport, vaccine, and vote. For the citizens' response, the following clusters were separated on two groups: related and not related to the city. The first group consists of transport, road, repair, government, law, education, house and communal services (HCS), money, shop, COVID, animals, life, death, and search. The last one refers to the search for missing people. The second group consists of complement, sentiment, video, opinions, dating, time, everyday life, author, family, sentiment. Since topics are not relevant to city, they will not be included in subsequent analysis. The present research reveals that the distribution of valuable topics, which are significant for both local authorities and citizens, varies among all the cities, as evidenced by the examination of heatmaps (see Fig. 1, 2). The citizens' agenda appears to be more diverse, indicating that a city is represented in a greater number of elements in public consciousness, which may not always be under the control of the city administration. In contrast, local authorities are more inclined to construct an image of the city through their responsibilities. This disparity can be observed in different scenarios. For instance, while local authorities were concentrating on elections and census, citizens were engrossed in conversations about death and COVID.

Furthermore, this data was collected during the height of the COVID-19 pandemic, and thus the most frequently discussed themes were not city-specific. Specifically, COVID, vaccination, and disease were prevalent in official communications, whereas COVID and death were predominant topics in the comments section.

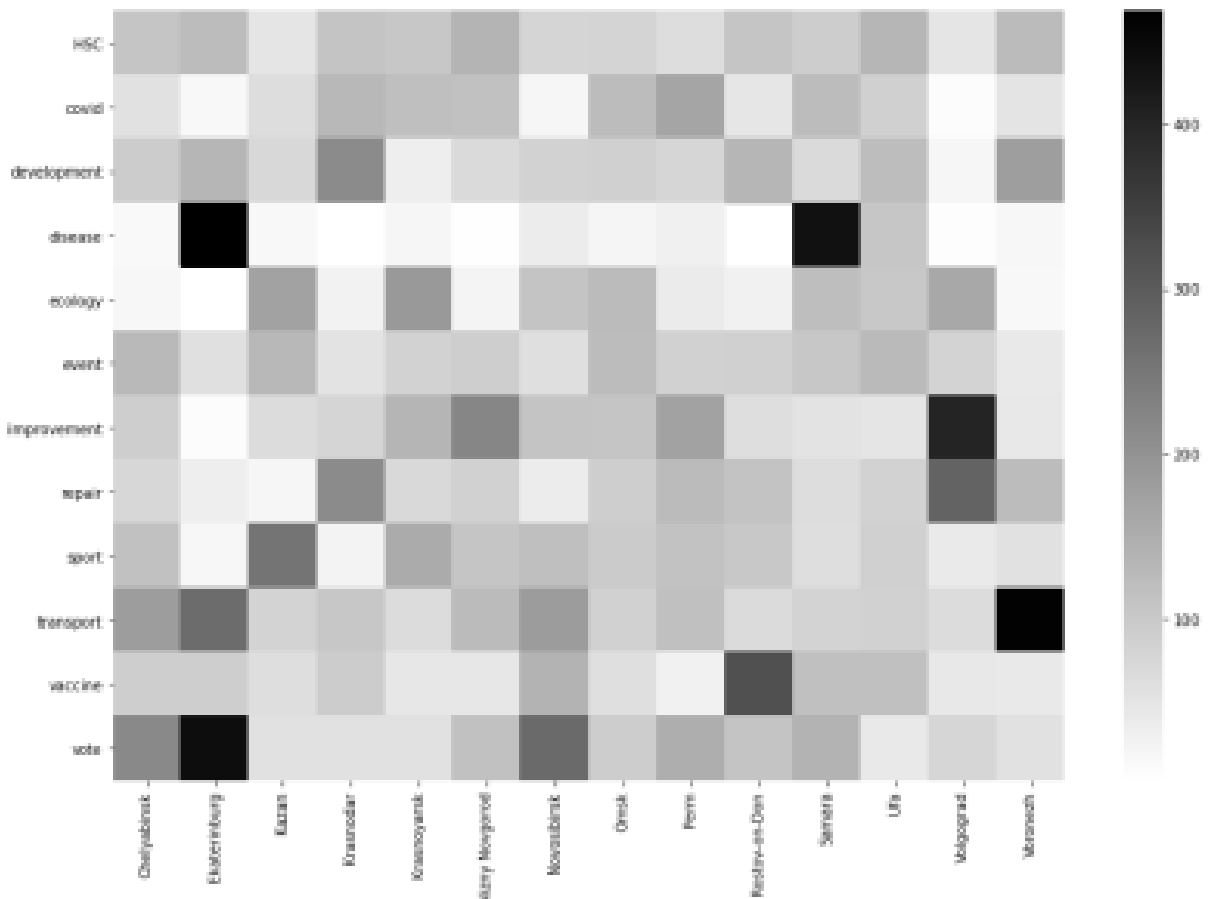


Fig. 1. Heatmap of relevant topic in official communication

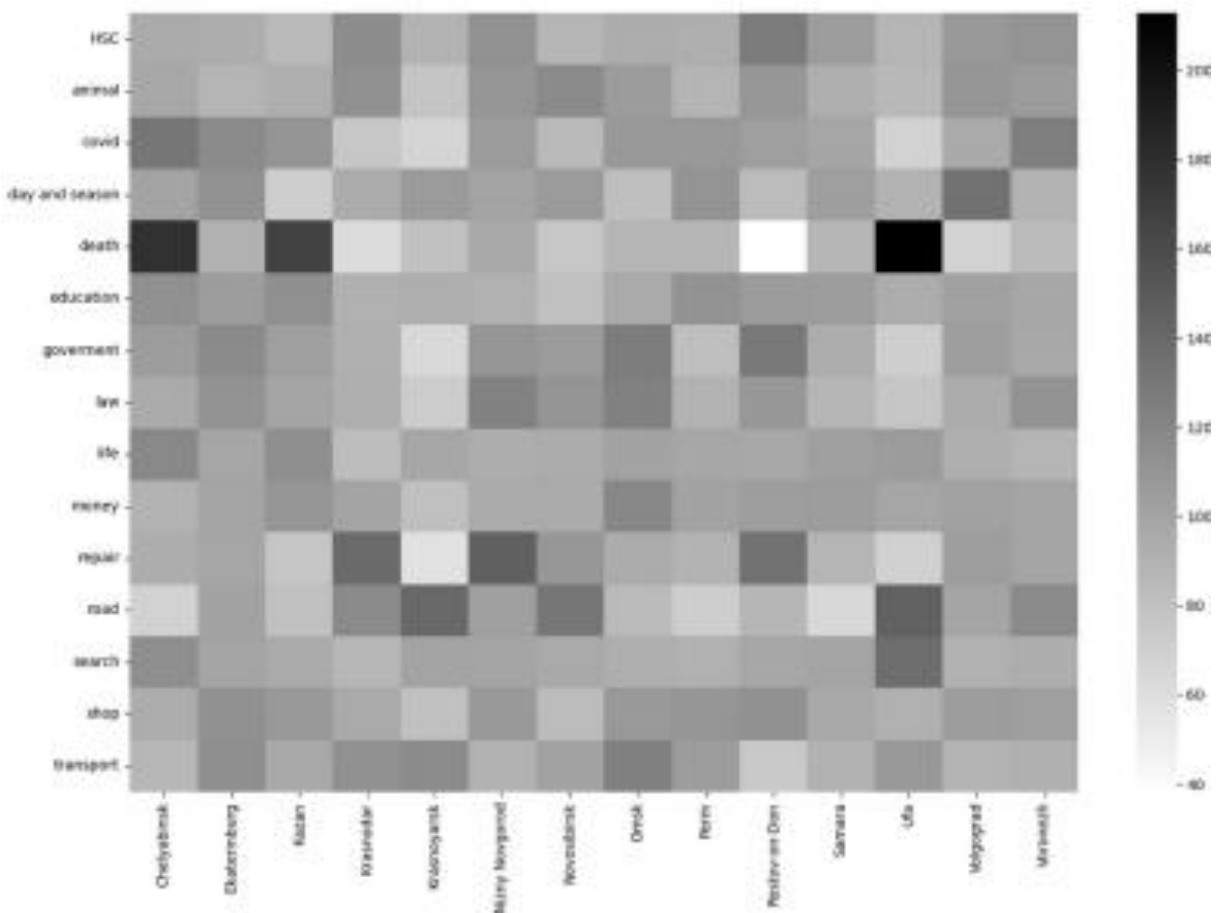


Fig. 2. Heatmap of relevant topic in citizens' comments

6. Sentiment analysis

A sentiment analysis was conducted on the topics deemed most stressful by users. Notably, formal communication was not incorporated into the analysis, given its inherent tendency towards neutrality in emotional expression. The model utilised for this analysis, known as DeepPavlov's RuBERT, was trained on "RuSentiment", a comprehensive dataset of Russian comments [14]. This model has been designed to ascertain three categories of sentiment: neutral, negative, and positive. The topics that elicited the most emotional responses (refer to Fig. 3) were those related to existential themes such as life and death, as well as quotidian elements such as daily activities and seasonal changes, inclusive of weather conditions. When localized to urban environments, the topics that predominantly evoked negative sentiments were transport, governance, and infrastructure maintenance. Conversely, the topics that were associated with positive sentiments were daily activities and seasons, death, and exploration. The positive sentiments in the latter two topics can be attributed to an expression of hope or a favorable assessment of an individual's worth. The sentiment distribution across different cities demonstrated a resemblance to affinity indices (refer to Fig. 4). These findings facilitate a nuanced understanding of the sentiments associated with specific topics within each city. Consequently, municipal administrations can employ this data to pinpoint the topics that are the most distressing and thereby trace the underlying urban issues.

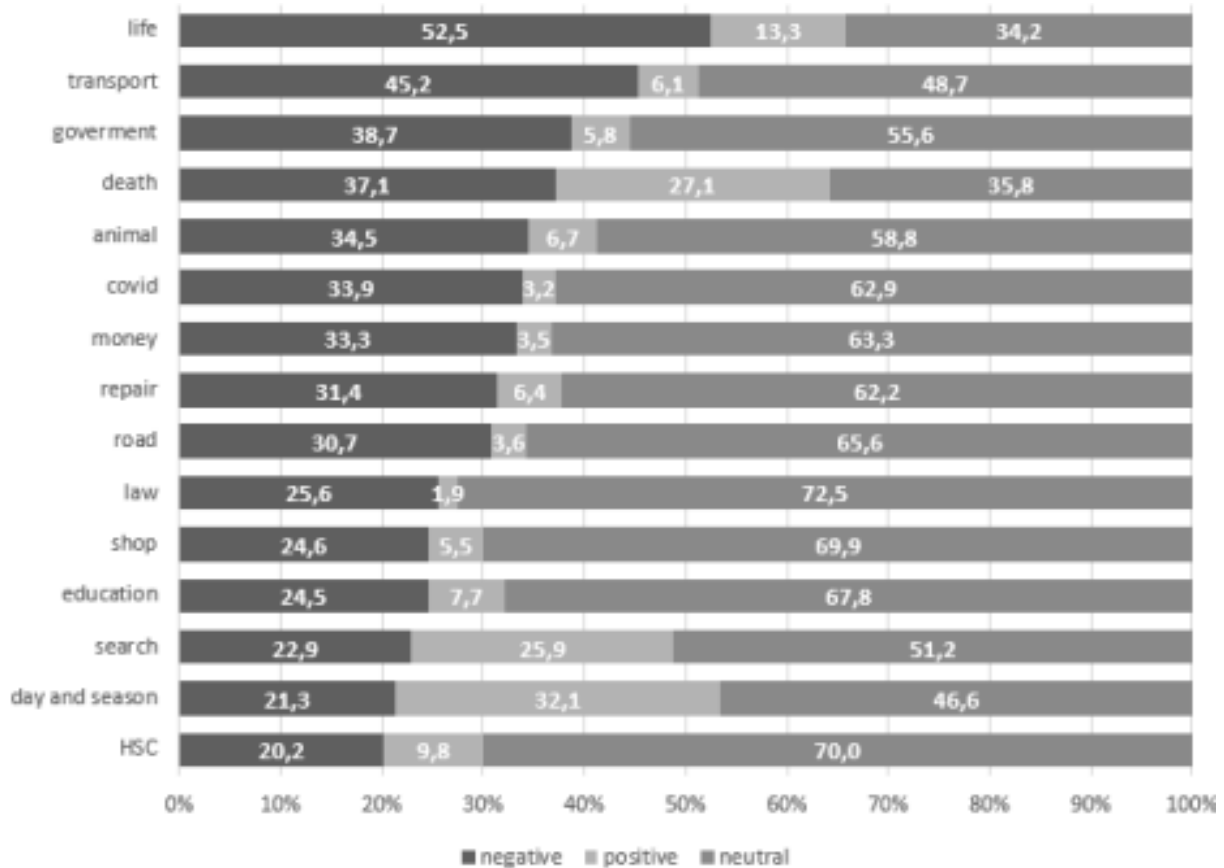


Fig. 3. Percent bar of sentiments per topic

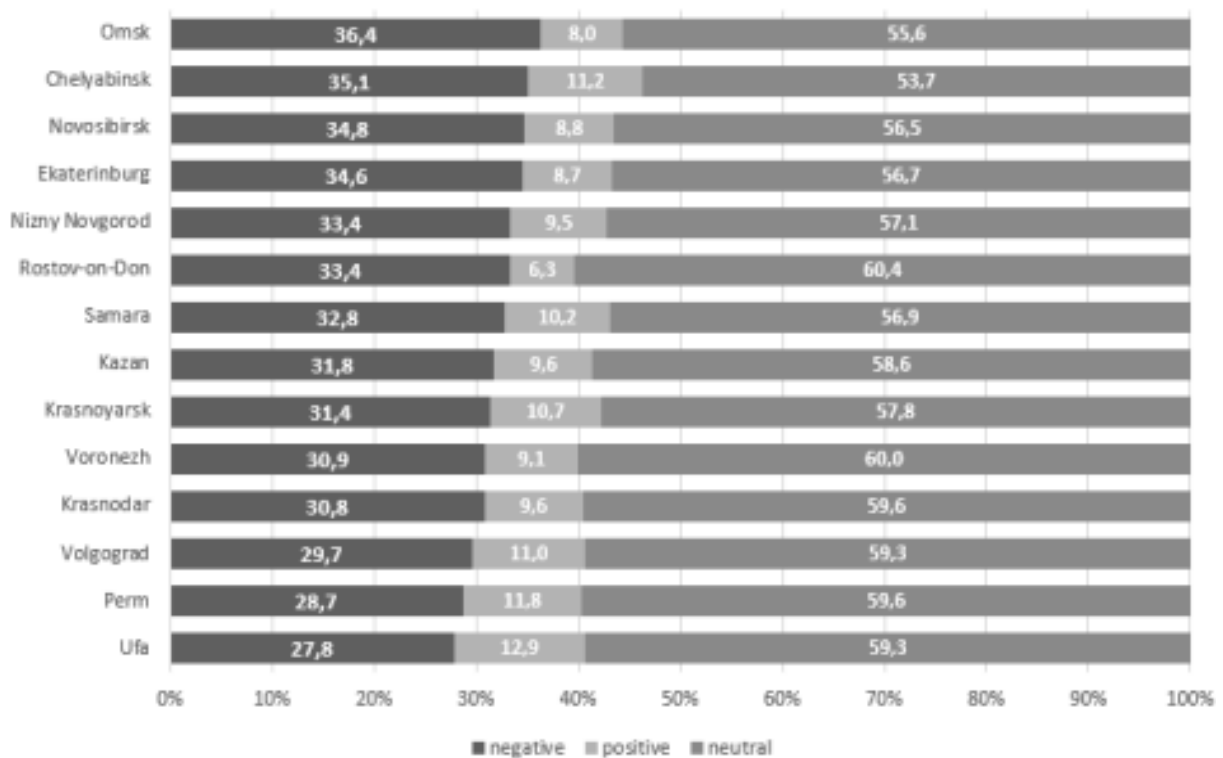


Fig. 4. Percent bar of sentiments per city

7. Conclusion

The observed variance in the distribution of topics within official communication across different cities, contrasted with the relative consistency in the distribution of topics in city communication, implies that municipal administrations tend to concentrate on particular subjects, whereas citizens are more likely to address daily challenges. This hypothesis is corroborated by sentiment analysis, which indicates that city related topics typically garner a relatively low volume of positive comments. Indeed, citizens frequently express critiques of transportation policies and governmental actions.

The primary inference drawn from these findings is that the existing content policies implemented by city administrations are suboptimal and ought to place a heightened focus on city communication within public forums. The algorithm and methodology proposed herein can be utilized to scrutinize feedback and amend content policies, potentially benefiting urban planners and policymakers endeavoring to enhance their communication with citizens.

Future research endeavors will incorporate named entity recognition in order to reconstruct the representation of city locations and mayors, in addition to evaluating public opinions about the city. Furthermore, the authors intend to update their dataset with publications from the years 2022 and 2023 in order to scrutinize the dynamic alterations in the city's image.

References

1. Alam M. Social Media Sentiment Analysis through Parallel Dilated Convolutional Neural Network for Smart City Applications / M. Alam, F. Abid, C. Guangpei et al. // *Computer Communications*. 2020. No. 154. Pp. 129–137.
2. Blei D. M. Latent Dirichlet Allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // *Journal of Machine Learning Research*. 2003. No. 3. Pp. 993–1 022.
3. Jiang B. The Image of the City out of the Underlying Scaling of City Artifacts or Locations / B. Jiang // *Annals of the Association of American Geographers*. 2013. No. 103 (6). Pp. 1 552–1 566.
4. Huang J. The Image of the City on Social Media: a Comparative Study using "Big Data" and "Small Data" Methods in the Tri-City Region in Poland / J. Huang, H. Obracht-Prondzynska, D. Kamrowska-Zaluska et al. // *Landscape and Urban Planning*. 2021. No. 206. P. 103 977.
5. Kitchin R. Data-driven, Networked Urbanism / R. Kitchin. 2015.
6. Kolmogorova A. V. Metaphorical Means for Constructing the Image of Krasnoyarsk in the Discourse of Social Media / A. V. Kolmogorova, M. I. Melnikova // *Eurasian Union of Scientists*. 2020. No. 4-9 (73). Pp. 48–52.
7. Koshkarova N. Cognitive Linguistics on the Payroll of Society: Why and How is it Important to Study Region's Media Image? / N. Koshkarova, O. Solopova // *Eurasian Union of Scientists*. 2021. No. 4 (47). Pp. 436–442.
8. Kowalik K. Social Media as a Distribution of Emotions, not Participation. Polish Exploratory Study in the EU Smart City Communication Context / K. Kowalik // *Cities*. 2021. No. 108. P. 102 995.
9. Molinillo S. Smart City Communication via Social Media: Analysing Residents' and Visitors' Engagement / S. Molinillo, R. Anaya-Sánchez, A. M. Morrison et al. // *Cities*. 2019. No. 94. Pp. 247–255.
10. Ovchinnikova I. G. Perm in the English Segment of the Web: 2012 vs. 2014 / I. G. Ovchinnikova, A. I. Zavershinskaya // *Local Discourse and the Construction of the Image of the Territory: Abstracts of the Russian Scientific Conference (Perm, 2014)*. Pp. 22–23.
11. Pavlova A. N. The City Portrait in the Federal Media: Cultural Events / A. N. Pavlova // *Vestnik VUiT*. 2017. No. 2. Pp. 156–162.
12. Popescu A. Mining Social Media to Create Personalized Recommendations for Tourist Visits / A. Popescu, G. Grefenstette // *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*. 2011. Pp. 1–6.
13. Pereira J. F. F. Social Media Text Processing and Semantic Analysis for Smart Cities / J. F. F. Pereira // *arXiv Preprint*. 2017. arXiv: 1709.03406.
14. Rogers A. RuSentiment: an Enriched Sentiment Analysis Dataset for Social Media in Russian / A. Rogers, A. Romanov, A. Rumshisky et al. // *Proceedings of the 27th International Conference on Computational Linguistics*. 2018. Pp. 755–763.
15. Vsevolodova A. N. Media Figure of Russian Cities in Federal Print Media / A. N. Vsevolodova // *Theoretical and Practical Issues of Journalism*. 2013. No. 2. Pp. 152–158.

Д. С. Мадрахимова¹,
А. В. Антощенко², Д. Ж. Корзун³ (науч. руководители),
¹ студент;

² доктор исторических наук, профессор кафедры отечественной истории;

³ кандидат физико-математических наук,

доцент кафедры информатики и математического обеспечения

¹ Национальный исследовательский университет ИТМО, Санкт-Петербург, Россия

^{2,3} Петрозаводский государственный университет, Петрозаводск, Россия

e-mail: ¹ dmadry@yandex.ru, ² antoshchenko@yandex.ru, ³ dkorzun@cs.karelia.ru

ПРИМЕНЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЭКСПЕРТНОЙ ОЦЕНКИ УСПЕШНОСТИ КАРЬЕРЫ РОССИЙСКИХ ПОРЕВОЛЮЦИОННЫХ ЭМИГРАНТОВ С УЧЁТОМ ИХ ДОРЕВОЛЮЦИОННОГО ПОЛОЖЕНИЯ В АКАДЕМИЧЕСКОМ СООБЩЕСТВЕ РОССИИ

Аннотация. В работе определяются исследовательские методики изучения механизма формирования культурно-исторической идентичности российских эмигрантов «первой волны», переживших такие травматические события, как революция 1917 г. и Гражданская война в России. Обосновываются важнейшие методы, которые позволяют соединить нарратологическое изучение коммеморативных практик с математическим анализом массовых источников, зафиксировавших изменение положения той части интеллектуальной элиты России, которая оказалась в изгнании.

Ключевые слова: *Московский университет, юбилей, русское зарубежье, эмиграция, академическое сообщество, культурно-историческая идентичность, математические методы в исторических исследованиях*

Сочетание перспективных тематических направлений исторических исследований, каким является изучение механизмов формирования исторической памяти и сохранение культурно-исторической идентичности эмигрантов, с применением математических методов и *IT*-технологий является насущной задачей современного развития исторического познания. Раскрытие механизмов формирования исторической памяти, а также выработка и усовершенствование методологических подходов к её изучению и конкретизация их в методических процедурах обработки массива данных являются одними из главных целей выполняемого научного проекта, коллектив которого осуществляет их на примере празднования в русском зарубежье 175-летнего юбилея Московского университета.

Главная цель исследовательской работы заключалась в формировании основного корпуса источниковой базы исследования – массива автобиографических анкет выпускников Московского университета и других высших учебных заведений дореволюционной России, а также в разработке основ методики, позволяющей решать намеченные в проекте задачи

и сформулировать новые исследовательские гипотезы. Анкеты отложились в фонде Р-5820 Государственного архива РФ. Они представляют собой уникальную источниковую базу, позволяющую применять не только математические методы в исторических исследованиях, но и *IT*-технологии. Однако до сих пор это собрание практически не введено в научный оборот. В отечественной историографии имеется лишь две научные статьи, в которых характеризуются или используются эти анкеты. В статье Е. В. Вьюницкой и Е. И. Пивовара [1] на основе предварительной обработки результатов анкетирования, проведённой самими членами Пражского комитета, представлены обобщённые данные, которые не в полной мере подтверждаются материалами обработанных анкет. В статье Е. В. Вьюницкой [2] предпринята попытка представить судьбы выпускников историко-филологических факультетов дореволюционных университетов, оказавшихся в изгнании. В других работах по истории празднования 175-летнего юбилея Московского университета данные анкетирования практически не используются (см., например, [3–5]).

Также целью являлась систематизация предварительно археографически обработанного массива автобиографических анкет. Помимо технической работы, связанной с созданием пакетов данных и их систематизацией в таблице, предполагалось выдвижение дополнительных гипотез исследования.

Обработка массива анкет Пражского комитета по ознаменованию 175-летнего юбилея Московского университета

Разработка принципов систематизации массива данных, охарактеризованных в статье Е. В. Вьюницкой и Е. И. Пивовара [1], первоначально предполагалась по параметрам, заданным пунктами анкеты Пражского комитета: 1) Ф. И. О., 2) год окончания вуза, 3) факультет или отделение, 4) деятельность в России, 5) деятельность в эмиграции, 6) страна проживания, что методологически обосновывалось в статье А. В. Антощенко [6]. Однако в ходе археографической обработки анкет было выяснено, что во многих анкетах есть данные о географическом перемещении в России и эмиграции. Причём последнее было более широко представлено. К тому же указание адреса в текстах ответов или на почтовых карточках, на которых прислала свои ответы значительная часть респондентов, позволило ввести дополнительные показатели: территориальное перемещение в эмиграции и не только страну, но и город проживания, что создало, в свою очередь, возможность внести качественные характеристики, разделив места проживания на «столицу» и «провинцию». Незначительная часть респондентов указали те научные степени (кандидата наук, присуждение которых предполагалось университетским уставом 1863 г.) или дипломы (I и II степени, введённые по университетскому уставу 1884 г.), которые они получали при окончании университетов. Кроме того, у определённой части респондентов можно было выяснить даты и места рождения, т. к. они были найдены в интернете. Всё это позволило увеличить

число показателей (а соответственно, столбцов в таблице) и расширить тем самым массив данных.

Для поиска связей и тенденций значение имеют данные, представляющие собой переменные категориального типа, т. е. принимающие конечное множество значений. На основе этих данных также строятся гипотезы о состоятельности карьеры респондента. К таким данным можно отнести, например, наличие у респондента диплома I, II степени или степени кандидата (4 возможных значения, включая отсутствие степени диплома или кандидата), статус оконченного учебного заведения по месторасположению (2 возможных значения – столица или провинция), статус места проживания в эмиграции (2 возможных значения – столица или провинция).

Для математической модели входными данными могут быть как текстовые описания, так и категории данных. Модели будут реализованы при помощи методов библиотеки машинного обучения *Scikit-learn* для языка *Python*, а также библиотеки глубокого обучения *TensorFlow/Keras*. Будет выполнено сравнение полученных моделей. Цель модели – определить состоятельность карьеры респондента по заданной классификации – восходящая, нисходящая, стабильная. В ходе исследования будут проверены разные типы классификаторов – преимущественно регрессионные и многослойные модели. Будут собраны их показатели, на основе которых будет выбрана лучшая модель и оптимальные параметры для неё, дающие наилучший результат. Непосредственно обучение модели планируется проводить в два этапа.

1. Вычистка текстовых данных и определение должности или сферы деятельности респондента в России и за рубежом

Данные в описании деятельности представлены в свободной форме на усмотрение респондента, поэтому для повышения точности прогноза модели необходимо удалить из текстов малоинформативные слова и фразы. Это будет выполнено с помощью библиотек *Python* для *NLP*. После этого – по оставшимся полезным тегам определить должность или сферу деятельности. Теги представляют собой фразы, которые могут быть несвязны и ошибочны с точки зрения естественного языка, но обязательно несут информативные данные, по которым можно с высокой точностью сделать прогноз. Для получения конечного результата будет разработан классификатор, принимающий на вход обработанный текст, содержащий теги о карьере респондента, и выдающий должность или сферу деятельности. Для обучения классификатора будет использован эталонный набор данных о должностях, полученный путём генерации случайных значений с помощью библиотеки *Faker* для *Python*. Набор будет включать не менее 100 000 позиций для увеличения вероятности попадания в него тегов из текста анкеты. Теги не всегда будут описывать конкретную должность, поэтому в таком случае будет определена сфера деятельности, т. е. взят непосредственно тег с не меняющими его смысл морфологическими изменениями. Например, если респондент писал «нёс военную службу», то зна-

чение «военная служба» будет сферой деятельности. Если в анкете указано «экстерн, ординатор, сверхштатный ассистент в факультетской терапевтической клинике», то должностью будут значения из набора «экстерн», «ординатор», «ассистент», которые окажутся в эталонном наборе. В итоге будет получен категориальный тип данных для описаний, т. к. число должностей и их сфер заведомо ограничено.

2. Обучение на обработанных данных

Входными данными для модели будут данные почти всех столбцов таблицы, т. е. только те данные, по которым можно сделать целевой вывод. Например, по Ф. И. О. и годам жизни такой вывод сделать нельзя. Выходными данными будут определённые моделью значения состоятельности карьеры для каждой анкеты таблицы. Чтобы говорить о корректности выданных прогнозов, необходимо получить значение не менее 90 % по всем возможным метрикам точности: *accuracy*, *precision*, *recall*, *F1-measure*. Для этого будет проведено сравнение с эталонными данными, ранее установленными исследователями. После обучения модель не будет использовать эталонные данные, вместо этого она должна давать прогноз на основе прочих данных анкеты, что позволит использовать её для других анкет, где состоятельность карьеры ещё не определена.

Экспертная оценка карьеры формируется из нескольких факторов, значение каждого из которых влияет на вероятность гипотезы о состоятельности карьеры, а значит, на её принятие или отвержение. Так, если респондент указал наличие диплома I, II степени или степени кандидата, окончил столичный университет и переехал в столицу за рубеж, то с высокой вероятностью карьера окажется восходящей и соответствующая гипотеза будет принята. Однако возможна и обратная ситуация, когда возникло противоречие. Поэтому приведённых выше категориальных данных недостаточно и необходимо анализировать содержание всей анкеты.

Дополнительные гипотезы для математической обработки анкетных данных с использованием IT-технологий

Выявление дат рождения ряда респондентов в сочетании с обязательным указанием года окончания вуза позволит рассчитать средний возраст окончания учёбы, что позволит впоследствии расчётами определить влияние возрастных характеристик респондентов на адаптацию в эмиграции. Гипотетически можно предположить, что молодые адаптировались лучше. Важным для обоснования достоверности полученных расчётных данных будет определение репрезентативности этих данных. Ещё одним важным моментом является то, что значительная часть респондентов окончила вузы в период Первой мировой войны и Революции 1917 г. в России и, как следствие, не имела возможности получить исходный опыт деятельности в соответствии с полученной специальностью. В этой связи для уточнения гипотезы важным будет ретрогрессивный расчёт «вверх» по годам от 1917 г., чтобы определить временной промежуток, когда формировался исходный опыт, значимый для адаптации.

Введение показателя места рождения важно для апробации гипотезы о влиянии географических перемещений в России для использования высшего образования как социального лифта. Гипотеза: переезд из провинции в столицу с целью учёбы в столичных вузах открывает большие перспективы по сравнению с сохранением провинциального места жизни или переездом из провинциального поселения в другое провинциальное место жительства.

Доклад подготовлен в Петрозаводском государственном университете (ПетрГУ) за счёт гранта Российского научного фонда № 23-28-00836 (URL: rscf.ru/project/23-28-00836).

Список источников

1. Вьюницкая Е. В. Анкета Пражского комитета по ознаменованию 175-летия Московского университета в его фонде в ГАРФ / Е. В. Вьюницкая, Е. И. Пивовар // Отечественные архивы. 2018. № 5. С. 37–44.

2. Вьюницкая Е. В. Выпускники Московского университета по данным анкеты Пражского комитета по ознаменованию 175-летия Московского университета / Е. В. Вьюницкая // Гуманитарий: актуальные проблемы гуманитарной науки и образования. 2019. Т. 19. № 2. С. 137–149. DOI: 10.15507/2078–9823.046.019.201902.137-149.

3. Щерблыгина И. В. Питомцы Московского университета в эмиграции / И. В. Щерблыгина. Lambert Academic Publishing, 2015. 80 с.

4. Горелова С. И. Русская эмиграция о 175-летию Московского университета / С. И. Горелова // Вопросы истории. 2005. № 8. С. 104–112.

5. Гутнов Д. А. Празднование 175-летия со дня основания Московского университета в Париже 24–26 января 1930 г. / Д. А. Гутнов // Вестник МГУ. Сер. 8: История. 2004. № 3. С. 88–101.

6. Антощенко А. В. Празднование 175-летия Московского университета в русском зарубежье (о методиках исследования) / А. В. Антощенко // Вестник ОмГУ. Сер.: Исторические науки. 2023. Т. 10. № 1 (37). С. 61–68. DOI: 10.24147/2312-1300.2023.10(1).61-68.

К. К. Косовская, Г. А. Тихонов, Д. Д. Долгозвягов
Сибирский федеральный университет, Красноярск, Россия

ЦИФРОВЫЕ ИНСТРУМЕНТЫ И РЕСУРСЫ, ПРИМЕНЯЕМЫЕ В ГУМАНИТАРНЫХ НАУКАХ

Аннотация. Эта статья посвящена методам преподавания цифровых гуманитарных наук с упором на их практическое применение в учебной среде. В статье мы рассмотрим ряд инструментов, которые помогут студентам и преподавателям начать работать с цифровыми подходами и оценить их важность для исследований. Мы рассмотрим онлайн-ресурсы для литературного анализа, простые программы для изучения медиа, архивные проекты, созданные студентами и преподавателями, которые позволяют познакомить широкую аудиторию с забытыми историями. Хотя эти инструменты не требуют практических знаний в области программирования, мы также рассмотрим ресурсы, которые пошагово обучают широко используемым языкам кодирования, таким как *Python*. В статье мы также рассмотрим методы и сервисы, которые помогут преподавателям создавать цифровые классы, основными характеристиками которых являются количественные подходы и возможности дистанционного обучения.

Ключевые слова: *цифровые гуманитарные науки, ДН-контент, дистанционное чтение, графический нарратив*

Цифровые гуманитарные науки (ЦГН) – область, которая объединяет традиционные гуманитарные науки с использованием цифровых технологий и методов исследования. Это быстро развивающаяся область, которая становится всё более популярной в академическом сообществе. Однако для того чтобы преподавать ЦГН, необходимо уметь использовать соответствующие инструменты, методы и ресурсы.

Цифровые методы и *ДН*-контент имеют огромный диапазон применения: от сложных проектов, таких как цифровые издания или корпус из нескольких работ, до очень детализированных этапов в рамках урока. Использование цифровых методов и *ДН*-контента в учебной среде может помочь студентам развить критическое мышление, умение работать с данными и технологиями, а также создавать более интерактивные уроки.

Одним из основных методов *ДН* для цифровой работы является аннотирование данных. Так, веб-сайт *Padlet* позволяет учителям создавать цифровую доску объявлений, для которой учащиеся могут размещать тексты и загружать документы, изображения, презентации и ссылки. Оба направления (*TEI* и цифровые планшеты) позволяют студентам совместно делиться работой и результатами. Одной из жизнеспособных областей применения этого метода является эстетическое чтение: вместо того, чтобы вести личный (аналоговый) дневник чтения, учащиеся могут записывать любимые цитаты, отрывки, задавать вопросы на понимание и обсуждать прочитанное. Природа этого метода восходит к принципу *ДН* о со-

вместной работе и коллективных процессах, а не индивидуальном взаимодействии с литературным произведением. С одной стороны, студенты могут познакомиться с мнением сокурсников; с другой, учителя получают обзор общего мнения о литературном произведении. Этот метод особенно хорошо подходит для домашних заданий, поскольку учащиеся могут загружать свои работы в используемое приложение или в общую облачную папку.

Самый знакомый пункт в этом списке для учителей английского языка – вики-проекты. Они стали не только хорошо изученным учебным инструментом для класса *EFL*, но и примером интеграции цифровых изданий и обучения *DH* в учебную программу. В то же время вики-проекты могут быть непростым заданием для учащихся, поскольку они требуют более процессно-ориентированного подхода к оценке, а также совместных социальных форм. Однако создание проектов, подобных вики, аналогичных веб-сайтам, таким как *Padlet*, является удобным методом *DH* для входа как для учителей, так и для студентов, поскольку он избавляет от программирования и процесса аннотации на основе кода. Таким образом, использование вики-проектов в классе *EFL* представляет собой эффективный способ интеграции цифровых изданий в учебную программу, что позволит студентам лучше понимать темы и развиваться в соответствующих областях.

В настоящее время существует множество ресурсов, которые могут быть использованы для проведения тестирования и оценки знаний, некоторые из которых доступны онлайн. Например, *Mentimeter* и *SurveyMonkey* предлагают цифровые инструменты для проведения опросов и тестирования. *Mentimeter* предоставляет учителям инструмент, похожий на презентацию *PowerPoint*, с возможностью создания облаков слов, диаграмм и обратной связи в реальном времени. Это может быть полезным для оценки знаний и понимания учеников. *SurveyMonkey*, с другой стороны, предоставляет учителям возможность создания онлайн-анкет для тестирования знаний и получения обратной связи. В отличие от *Mentimeter*, *SurveyMonkey* не предоставляет возможности мгновенной аналитики, но результаты могут быть использованы для дальнейшего анализа. Это может быть полезным для того, чтобы понимать, какие темы нуждаются в дополнительной проработке. Ещё одним отличием является то, что учителя могут проецировать свои презентации в режиме реального времени, и ответы учеников могут быть отображены на экране, что позволяет учителям видеть, какие вопросы вызывают наибольшую трудность у учеников и какие темы нуждаются в дополнительном объяснении.

SurveyMonkey предоставляет более подробные результаты, но они видны только создателю анкеты. Участие со стороны студентов осуществляется в обоих случаях либо со ссылкой, либо с кодом, который участники размещают на веб-сайте. Следует также отметить, что, как и *Kahoot*, эти веб-сайты предлагают бесплатную, но ограниченную версию. Например, бесплатная версия *SurveyMonkey* позволяет задавать десять вопросов

в одном опросе. Более того, ни один из этих веб-сайтов не собирает личные данные и не требует регистрации участников – всё, что им нужно, – это код, предоставленный учителем. Такие опросы занимают примерно от 5 до 20 мин индивидуальных занятий.

Учебная программа содержит точное описание и интерпретацию визуализированных данных, таких как диаграммы, гистограммы и т. д., но пренебрегает актуальностью для учащихся визуализировать свои собственные данные таким образом. Как и в случае с большинством других методов *DH*, работа с данными таким образом распространяет практический подход к проблемам и темам, и студенты могут также научиться анализировать статистику, создавая свой собственный набор данных. Практичность этого метода варьируется от создания собственных опросов – например, моделирования анкеты по региональным темам – до литературного и лингвистического анализа: частотности слов в художественных текстах, репрезентации пола в любимых книгах учащихся, количества приобретённых со временем словарных единиц и последующей визуализации таких данных.

Преподаватели, которые хотели бы выйти за рамки возможностей визуализации данных, приложений для работы с электронными таблицами, таких как *Microsoft Excel* или онлайн-создателей облаков слов, могут начать работать с *Gephi*. Это бесплатное программное обеспечение с открытым исходным кодом с набором функций, как для небольших, так и для крупномасштабных наборов данных, которое позволяет пользователям исследовать свои собственные данные в широком спектре визуализаций, создавать плакаты для своих визуализаций.

Стоит отметить, что учебные и внеклассные проекты, которые собирают и анализируют данные, не только дают возможность учащимся работать со статистикой, но и позволяют думать о мире с точки зрения количественной оценки выражений. Благодаря количественному анализу класс *EFL* становится более разнообразным и интересным, что, в свою очередь, дает дополнительный стимул ученикам в обучении. Такие методы не заменяют традиционный стиль преподавания, но могут предоставить ещё один способ общения. Таким образом, количественный анализ данных может стать не только эффективным, но и интересным инструментом в процессе обучения.

Стилистический анализ, основанный на дистанционном чтении, является одним из ключевых элементов методологии *DH*. Этот подход может быть полезен для учителей, желающих продемонстрировать своим ученикам, что лингвистическая проницательность литературных произведений определяется не только традиционным анализом риторических приёмов и другими герменевтическими упражнениями. Дистанционное чтение позволяет глобально анализировать текст, выявлять его особенности, сравнивать с другими текстами и группировать сходства и различия. Данный подход также может помочь ученикам развить более широкое понимание произведений, увидеть их связи с другими текстами и контекстами, а так-

же открыть новые перспективы для интерпретации их содержания. Фактически, инструменты, используемые для этого метода, не ищут слова и фразы, которые можно было бы искать при внимательном чтении, а вместо этого приходят к выводам.

Одним из таких инструментов является пакетный стило для языка программирования *RStudio*. *RStudio* – программа с графическим интерфейсом, которая позволяет пользователям получать доступ к *R*-пакетам, а не просто вводить команды в текстовом редакторе. Настройка рабочего каталога осуществляется нажатием соответствующей кнопки в *RStudio* или вводом *set* с последующим путем к папке, содержащей текстовые файлы, в командной строке. Наконец, документация на веб-сайте *stylo* представляет его функции и соответствующие команды. Учащиеся могут обнаружить стилистическое сходство в корпусе текстов, могут исследовать особенности стиля автора, сравнивая с другими. Кроме того, они могут проводить сравнительный анализ текстов и выявлять общие черты, стилистические приемы, используемые в разных текстах. Поскольку дистанционное чтение, согласно определению, не погружается в тонкости текста, а предоставляет особенности поверхностного уровня в большом количестве текстов, так, учащиеся могут познакомиться с другим подходом анализа литературного текста. С этой точки зрения *DH* в учебной среде может мотивировать студентов, которые не были заинтересованы в какой-либо форме литературного анализа.

В данной статье мы представили разнообразие инструментов и ресурсов, которые можно использовать для подготовки курса в области цифровых гуманитарных наук. *DH* предоставляет возможность для более тесного взаимодействия между преподавателями и студентами. При работе с одними и теми же инструментами и ресурсами, преподаватели и студенты могут обмениваться мнениями и идеями, что способствует более глубокому пониманию предмета исследования. Это также может привести к появлению новых идей и проектов, которые могут быть реализованы в рамках гуманитарных наук. Однако, несмотря на все преимущества *DH*, следует учитывать, что эти методы не являются универсальными и не могут заменить традиционные методы исследования в гуманитарных науках. Поэтому важно использовать *DH*-инструменты и ресурсы в сочетании с традиционными методами, чтобы получить наиболее полное представление о предмете исследования.

Таким образом, *DH* представляет собой важный инструмент для исследования и анализа различных текстов и документов в гуманитарных науках. Однако для достижения наилучших результатов необходимо использовать *DH*-методы в сочетании с традиционными методами исследования, а также расширять сотрудничество между преподавателями и студентами. Такой метод обучения открывает новые возможности для оригинального подхода к теоретическим концепциям в классе, позволяя студентам более полно и глубоко изучать материал и развивать собственные творческие способности. Кроме того, использование инновационных мето-

дов и технологий в обучении *DH* может способствовать активному вовлечению учеников в учебный процесс и содействовать более эффективному и понятному усвоению материала.

Список источников

1. Cohen M. The New Life of the New Forms: American Literary Studies and the Digital Humanities. A Companion to American Literary Studies / M. Cohen. Hoboken: Wiley, 2011. Pp. 532–548.

2. Dunst A. The Graphic Narrative Corpus: Design, Annotation, and Analysis for the Digital Humanities / A. Dunst, J. Laubrock, R. Hartel // Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. 2017. Pp. 15–20.

3. Eder M. Stylometry with R: a Package for Computational Text Analysis / M. Eder, J. Rybicki, M. Kestemont // R Journal. 2016. No. 8 (1). Pp. 107–121.

4. Moisich O. The Digital Classroom: a Digital Humanities Primer on Tools, Methods, and Resources / O. Moisich // American Studies Journal. 2020. No. 70. Pp. 165–187.

5. Ruth A. The Wiki Way of Learning / A. Ruth, L. Houghton // Australasian Journal of Educational Technology. 2009. No. 25 (2). Pp. 135–152.

6. Sofkova H. Making Room for the Transformation of Literacy Instruction in the Digital Classroom / H. Sofkova, K. Cederlund // Early Childhood Literacy. 2017. No. 17 (2). Pp. 221–253.

А. С. Дяченко¹, И. А. Кижнер²,

¹ специалист лаборатории Digital Humanities;

² кандидат культурологии, доцент,

старший научный сотрудник лаборатории Digital Humanities

^{1,2} Сибирский федеральный университет, Красноярск, Россия

e-mail: ¹ alinadyachenko12@gmail.com, ² ikizhner@sfu-kras.ru

ПОДХОД К МУЛЬТИМОДАЛЬНОМУ АНАЛИЗУ ПЕРИОДИЧЕСКОГО ИЗДАНИЯ «КУРЬЕР ЮНЕСКО»: СЕГМЕНТАЦИЯ И ИЗВЛЕЧЕНИЕ ИЗОБРАЖЕНИЙ

Аннотация. Исторические периодические издания, такие как журналы, газеты, сочетают в своем оформлении текст и изображения для достижения нужного визуального эффекта на читателя. Разделение изображений и текстов в таких документах необходимо для изучения вариативности способов представления контента, а также для подготовки корпусов изображений и текстов к мультимодальному анализу. Чтобы узнать, насколько согласованы эти два корпуса, нам необходимо создать отдельные наборы данных, а далее найти способ соединить их. В данной работе мы сосредоточимся на корпусе, охватывающем два десятилетия XX в.: 60-е и 90-е гг. Корпус включает в себя 10 тыс. изображений и более 3 млн слов из журнала «Курьер ЮНЕСКО» – главного издательского центра ЮНЕСКО. Мы выполняем сегментацию макета и извлекаем изображения с помощью набора алгоритмов. Мы сравниваем нашу модель с несколькими универсальными моделями и показываем, что необходима тонкая настройка для извлечения изображений из журнала с экспериментальным дизайном и макетом.

Ключевые слова: «Курьер ЮНЕСКО», мультимодальный анализ, кросс-модальный анализ, сегментация изображений

Введение

Мультимодальный анализ [1–3] объединяет анализ изображений и текстов, опираясь на текстоориентированные исследования [4–6] и исследования изображений [7; 8] в цифровых гуманитарных науках.

В данной работе мы готовим для анализа и сравнения исторические изображения и тексты, опубликованные на страницах журнала «Курьер ЮНЕСКО». В частности, мы представляем модель, которая даёт хорошие результаты сегментации и извлечения изображений. Работа выполнена на корпусе из 209 номеров журнала «Курьер ЮНЕСКО», выпущенных в 1960-е и 1990-е гг. Архив «Курьера ЮНЕСКО» публикуется на сайте ЮНЕСКО и включает в себя выпуски с 1947 г. по наше время. «Курьер ЮНЕСКО» содержит большое количество иллюстраций (более 10 тыс. изображений за изучаемые десятилетия), при этом синергия отношений текста и изображения высоко оценивается аудиторией [9]. Издание служило площадкой для пропаганды отношения ЮНЕСКО к тому, как воспринимать культурные каноны человечества: начиная от создания единого

культурного канона для всех до идеализированного гармоничного хора различных культурных канонов [10].

Наша работа призвана поддержать важность анализа журнала «Курьер ЮНЕСКО» как периодического издания, в котором изображения и тексты не находятся в прямых отношениях, а могут противоречить или дополнять друг друга, создавая синергию. Кроме того, дополнительную сложность добавила экспериментальная верстка в «Курьере ЮНЕСКО» с изображениями и текстами на цветном фоне. Прежде чем перейти к мультимодальному анализу [11] или кросс-модальному анализу [12–14], наша задача – подготовить отдельные корпуса текстов и изображений и соединить их для дальнейшего анализа. Наш проект основан на предварительной обработке изображений для анализа, используемых в проекте *Newspaper Navigator* [15]. Работа включает в себя оценку нескольких универсальных решений, разработанных для типичных случаев верстки исторических документов. В нашей статье тестируется подход к использованию сочетания нескольких этапов обработки страницы журнала для решения нетипичных случаев макета, которые могут стать препятствием для автоматической сегментации макета и извлечения изображений. Новый подход показывает, что специфика исторического периодического издания, направленного на создание визуального эффекта и противопоставление множества культурных канонов, может не вписываться в модели, нацеленные на универсальные дизайнерские решения. Этот подход согласуется с литературой, которая демонстрирует, как ошибочный автоматический анализ макетов в эпоху, предшествовавшую платформам распознавания, таким как *Transkribus*, препятствует текстовому анализу [16].

Цель этой статьи – определить, можем ли мы извлечь изображения из периодического издания с экспериментальным дизайном страниц. Мы разрабатываем модель, которая может выполнять сегментацию макета и извлекать изображения из Курьера ЮНЕСКО. Мы оцениваем качество модели и показываем, что она достигает точности 96 % и *F-Score* 70 %, что в три раза превышает результаты самой эффективной модели в оценочном наборе. Наконец, мы автоматизируем группировку изображений и текста на одной странице. Мы утверждаем, что универсальные модели не справляются с экспериментальным дизайном и сложностью исторических документов, а это означает, что необходима точная настройка и уникальные решения для мультимодального или кросс-модального анализа текстов и изображений.

1. Методология

1.1. Описание датасетов

Исследование проводилось на журналах «Курьер ЮНЕСКО», вышедших в 1960-е и 1990-е гг. Эти два десятилетия были выбраны для того, чтобы проверить созданный алгоритм на разных по дизайну журналах (табл. 1).

Сравнение стилистических особенностей журналов 1960-х и 1990-х гг.

Стилистические особенности журналов 60-х	Стилистические особенности журналов 90-х
Чёрно-белый формат выпуска	Цветной формат выпуска, но иногда есть чёрно белые страницы или фото
Использование чёрных и серых подложек под текст/изображения	Использование однотонных цветных или градиентных цветных подложек под текст/изображения
Преобладание чёрно-белых фотографий	Преобладание цветных фотографий

Таким образом, первый набор данных содержит 109 журналов (4 261 страниц), вышедших в период с 1960 по 1969 гг. Второй набор состоит из 110 журналов (6 042 страниц), вышедших в период с 1990 г. по 1999 г.

1.2. Извлечение изображений

Журналы «Курьер ЮНЕСКО» содержат множество изображений, в т. ч. фотографии известных личностей, исторических событий, природы, а также различные схемы и иллюстрации. Определение визуальных тенденций в изображениях, совместное изучение текстовой и визуальной частей журнала для мультимодального анализа, являются интересными направлениями исследований. Поэтому решение проблемы извлечения изображений в данном исследовании позволит нам приблизиться к поставленным будущим задачам.

Для лучшей точности извлечения изображений сначала мы обработали страницы журналов: удалили полностью текстовые страницы с помощью предобученной нейронной сети *Fast.Ai*, а также текстовый слой с оставшихся страниц с помощью *PuMuPDF*. Это позволит избежать ошибок на этапе извлечения изображений (рис. 1).



Рис. 1. Ошибки, которых удастся избежать при извлечении изображений благодаря предобработке

Далее с помощью библиотеки компьютерного зрения *OpenCV* происходит процесс поиска потенциальных границ изображений с обработанных страниц журнала. Для этого были подобраны значения контрастности, размытия и цветовых пятен (рис. 2).



Рис. 2. Процесс поиска границ с помощью *OpenCV*

Результатом работы библиотеки *OpenCV* является папка с изображениями, однако многие из них являются слишком маленькими и незначительными, поэтому мы удалили изображения длина или ширина которых меньше 150 пикселей. Также в результат извлечения попадет текст на цветных подложках, который требуется удалить из результата по извлечению визуальной части журнала. Для этого мы снова использовали *Fast.Ai*, где обучающая выборка представляла собой комбинацию между двумя видами данных: вырезки журналов «Курьер ЮНЕСКО» 1970-х и 2000-х гг. и вырезки из проекта *Newspaper Navigator dataset* [17] с ручными пометками о том, является ли вырезка полностью текстовой или содержит ещё и изображение.

В выборке *Newspaper Navigator dataset* находятся подходящие данные для обучения нашей нейронной сети: 603 рекламные вырезки из газет Америки 1905 г. с пометками о том, является реклама полностью текстовой, или там присутствует иллюстрация. Однако при использовании в качестве обучающей выборки только данных из *Newspaper Navigator dataset*, нейронная сеть часто ошибалась (24 % *accuracy*). Было решено добавить в обучающую выборку 714 вырезок из журналов «Курьер ЮНЕСКО» 1970-х и 2000-х гг. 70-е гг. были взяты для того, чтобы исключить пересечения обучающей выборки с исследовательской выборкой 60-х и 90-х гг. Вырезки из журналов 2000-х гг. были добавлены в обучающую выборку из-за того, что в журналах 70-х гг. не было примеров оформления текста на градиентных цветных подложках, что было популярно в журналах 90-х гг. Это позволило отделить текст на подложках от результата извлечения изображений, при этом оставить другой сложный случай: изображения и текст на цветной подложке (рис. 3).

Данные результаты говорят нам о неуниверсальности задачи извлечения картинок из исторических документов. Таким образом, для достижения хорошего результата извлечения картинок из исторического документа со своими особенностями оформления, требуется создавать собственный метод, настроенный для конкретного объекта исследования. В дальнейшем мы планируем разработать свой метод извлечения изображений из журналов «Курьер ЮНЕСКО».

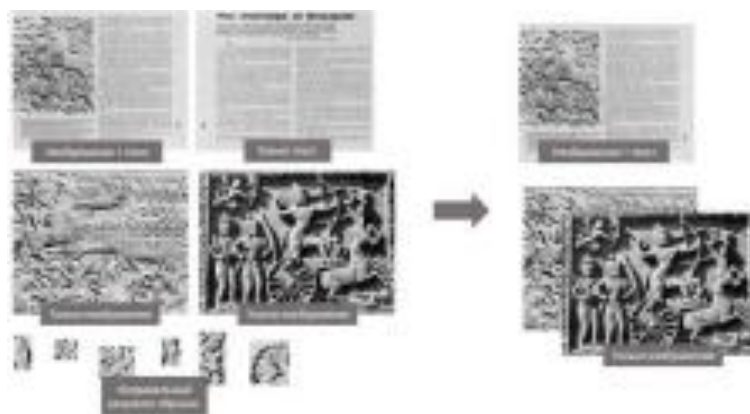


Рис. 3. Результат дополнительной обработки извлечённых изображений

Важно заметить, что большинство (88 %) картинок на цветной подложке с текстом программа вырезает именно по границе подложки, а не самой картинке (табл. 2).

Таблица 2

Процент сложных случаев в журналах 1960-х и 1990-х гг.

Журналы	Текст на цветной подложке	Картинки на цветной подложке
«Курьер ЮНЕСКО» за 60-е гг.	11 %	19,1 %
«Курьер ЮНЕСКО» за 90-е гг.	8 %	11,5 %

Исходя из того, что у нас есть как текст на подложке, так и картинка вместе с текстом на подложке, мы не можем удалить не интересующие нас тексты на цветной подложке с помощью фильтрации по обнаружению текста. Удалив картинки, где обнаружится много текста, мы потеряем также и картинки на цветной подложке. Для нас приоритетнее сохранить данные картинки в формате с подложкой и текстом, чем совсем потерять их в итоговом результате. Поэтому алгоритм удаляет только текст на цветной подложке.

В табл. 3 представлены результаты работы алгоритма. Важно отметить, что метрика *Accuracy* не отразит то, что есть разные классы проблем при извлечении картинок (см. разд. 2 «Результаты»). Например, картинки на подложке или логотипы и рекламные изображения. *F-Score* оценивает алгоритм более точно.

Таблица 3

Процент сложных случаев в журналах 1960-х и 1990-х гг.

Журналы	Было найдено картинок	<i>Accuracy</i>	<i>F-Score</i>
«Курьер ЮНЕСКО» за 60-е гг.	4 285	0,96	0,69
«Курьер ЮНЕСКО» за 90-е гг.	5 841	0,97	0,7

1.3. Извлечение текста

Поскольку все журналы «Курьер ЮНЕСКО» представлены на сайте «ЮНЕСКО» уже с распознанным текстом, нам не требовалось заниматься разметкой. Важно отметить, что текст с подложек, о которых говорилось в п. 1.2., также размечен и входит в текстовую выборку. Таким образом, нам требовалось только подготовить текстовый корпус для мультимодального анализа.

Алгоритм сохраняет весь текст с одной страницы в файл формата *txt*, а в названии файла помечает, с какой страницы он был взят. Это требуется для того, чтобы в дальнейшем можно было обратиться к определённому текстовому файлу для совместного анализа изображений и текста с одной страницы. В табл. 4 представлен результат извлечения всего текстового корпуса.

Таблица 4

Текстовый корпус журналов за 1960-е и 1990-е гг.

Журналы	Текстовый корпус (слов)
«Курьер ЮНЕСКО» за 60-е гг.	1 951 161
«Курьер ЮНЕСКО» за 90-е гг.	2 623 578

1.4. Группировка текста и изображений

Поскольку данная работа является подготовкой к кросс-модальному анализу журнала «Курьер ЮНЕСКО», нам нужно было найти способ группировки текста и изображений с одной страницы.

В созданном алгоритме процесс поиска изображений на каждой странице осуществляется с помощью цикла. Программа не перейдёт на следующую страницу до тех пор, пока не найдёт все изображения на текущей странице. Поэтому мы решили присвоить каждому найденному изображению номер страницы, который равен порядковому номеру итерации цикла. Такой подход позволяет нам учитывать то, что на одной странице может быть сразу несколько изображений. Как было сказано в п. 1.3, для текста также используется нумерация в названии, чтобы понимать, с какой страницы он был взят. Так, файл с текстом со стр. 1 будет называться «1_text». Данная часть кода созданного алгоритма в перспективе позволяет попарно анализировать текст и фото с одной и той же страницы, а далее и сравнивать.

Таким образом, созданный алгоритм позволяет извлекать отдельно изображения (*Accuracy* 96,5 %) и текст (*Accuracy* 100 %) со страниц журналов «Курьер ЮНЕСКО», при этом, в названиях файлов помечается, с какой страницы они были взяты, поэтому в будущем к этим данным будет возможно обратиться и проанализировать их попарно.

2. Результаты

Поскольку сбор текста из уже размеченного документа не представляет сложности, мы оценивали точность извлечения изображений созданного алгоритма с результатами других инструментов. Так, мы проверили бесплатную универсальную программу *PDF24 Tools* [18] и два способа платформы *Transkribus* [19] (табл. 5).

Таблица 5

Проверенные инструменты для выделения картинок в тексте

<i>PDF24 Tools</i>	Многофункциональная и простая в использовании программа для работы с <i>PDF</i> , с помощью которой можно извлекать картинки из <i>PDF</i> -файла
<i>Transkribus Printed Block Detection</i>	Автоматическая функция <i>Transkribus</i> по поиску текстовых блоков и разметки страницы
<i>Transkribus P2PaLa</i>	Метод по разметке страницы, где можно использовалось предварительное обучение на наших данных

Среди функций программы *PDF24 Tools* есть извлечение картинок из файла формата *PDF*. Нам хотелось узнать, как «универсальная» программа справится с историческими документами и сложными стилистическими особенностями дизайна журнала. В *Transkribus* были протестированы два метода разметки страницы: *Printed Block Detection* и *P2PaLa*. С помощью первого метода можно автоматически искать «печатные блоки» (картинки, абзацы) для разметки. Второй метод также делает разметку страницы, но для его обучения мы использовали 100 страниц из журналов «Курьер ЮНЕСКО» (случайные страницы из журналов 1970-х гг.), размеченных ручным способом в *Transkribus*. Количество страниц для обучения было выбрано исходя из инструкции на сайте *Transkribus*, релевантной для этого метода [20].

Все методы, включая наш, проверялись на двух случайных выборках: 10 % журналов (11 выпусков) из 1960-х и 10 % журналов (11 выпусков) из 1990-х. Для создания сравнительной таблицы всех методов были использованы метрики *Accuracy*, *Precision*, *Recall* и *F-Score* (табл. 6 и 7).

Таблица 6

Сравнение способов: 1960-е гг.

Способ	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>
Созданный метод	0,96	0,69	0,7	0,69
<i>PDF24 Toolbox</i>	0,06	0,01	0,05	0,02
<i>Transkribus Printed Block Detection</i>	0,34	0,07	0,54	0,12
<i>Transkribus P2PaLa</i> (модель, обученная на наших данных)	0,08	0,01	0,05	0,02

Сравнение способов: 1990-е гг.

Способ	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>
Созданный метод	0,97	0,7	0,69	0,7
<i>PDF24 Toolbox</i>	0,08	0,02	0,06	0,03
<i>Transkribus Printed Block Detection</i>	0,41	0,1	0,61	0,17
<i>Transkribus P2PaLa</i> (модель, обученная на наших данных)	0,08	0,02	0,06	0,03

Accuracy показывает долю правильных ответов алгоритма. Так, программа *PDF24 Tools* верно сохранит 6 % картинок из журналов 60-х гг., *Transkribus Printed Block Detection* 34 %, а наш метод 96 %. Однако *Accuracy* не отразит то, что есть разные классы проблем при извлечении картинок, и все представленные инструменты справляются с ними по-своему. Так, 34 % *Accuracy* у *Transkribus Printed Block Detection* не покажет, что данный метод будет определять текст на цветной подложке как изображение. Поэтому дополнительно была посчитана метрика *F-Score*. *F-Score* объединяет в себе информацию о точности (*Precision*) и полноте (*Recall*) нашего алгоритма. Точность (*Precision*) системы рассчитывается в пределах определённого класса (картинка с чёткими границами, логотип/реклама, текст на цветной подложке и т. д.) – так мы учтём долю картинок, которую метод неправильно включил в результат, а также потери картинок. Полнота (*Recall*) системы – это доля найденных правильных картинок относительно всех найденных картинок. Таким образом, метрика *F-Score* позволила нам оценить эффективность использованных методов точнее. Что мы получили? Табл. 6 для 1960-х гг. показывает, что лучшая эффективность работы у созданного метода (69 %). У *PDF24 Tools* и *Transkribus P2PaLa* мы видим одинаковый результат эффективности 2 %. Эти методы определяют каждую страницу журнала целиком как картинку, т. е. не идентифицируют никакой структуры внутри страницы. 2 % получилось из-за того, что в журнале «Курьер ЮНЕСКО» иногда картинка занимает всю страницу. У *Transkribus Printed Block Detection* (результат 16 %). Этот метод пропускает много изображений и ошибается при определении их границ. Вертикальные линии на картинках *Transkribus* определяет как *separator*, т. е. как разделители в таблицах и включает это в структуру страницы. Что означают эти результаты? Разрыв между методами, которые не были специально настроены (или были недостаточно настроены) для журнала «Курьер ЮНЕСКО», и нашим методом достаточно большой. Это говорит о неуниверсальности задачи извлечения картинок из исторических документов (т. е. из документов, которые не подвергались компьютерной обработке в момент создания).

Заключение

В этой работе был создан алгоритм для извлечения изображений и текстовых блоков из журналов «Курьер ЮНЕСКО». Кроме того, программа группирует изображения и тексты для мультимодального анализа. Алгоритм представляет собой комбинацию из инструментов библиотеки компьютерного зрения *OpenCV*, нейронной сети, созданной с помощью *Fast.AI* и других вспомогательных библиотек: *PyPDF2*, *PyMuPDF* и *PIL*. Такой подход позволил достичь максимальной точности извлечения изображений (97 % *Accuracy*, 70 % *F-Score*), что и представляло основную сложность. Созданный алгоритм учитывает особенности журналов «Курьер ЮНЕСКО», например, то, что текст располагается на цветных подложках, которые можно было бы принять за изображения. Разработанный подход даёт возможность успешно отделить сложные случаи оформления текстовых блоков и не включать их в результат извлечения изображений. При этом работа демонстрирует, что инструменты платформы *Transkribus*, программа *PDF24 Tools*, т. е. более универсальные способы, не справляются с извлечением изображений из исторических документов («Курьер ЮНЕСКО»), и дают не более 41 % точности (*Accuracy*).

Мы считаем, что задача извлечения картинок всё ещё является сложным и не универсальным этапом подготовки к анализу изображений, который в определённой степени препятствует изучению визуального материала отсканированных книг, журналов, газет и других исторических документов. Также этот этап необходим для проведения мультимодальных исследований, описанных в работе [11]. Так, в рамках мультимодального исследования возможно провести сравнение результатов компьютерного анализа изображений и текста с одной страницы журнала. Для этого в алгоритм встроен счётчик изображений, а в названии помечается, на какой странице они были обнаружены. Исследование изображений, сопоставление их с текстом – важный новый этап цифровых гуманитарных исследований. Однако мы должны понимать, что исследования с возможностью формулирования гипотез или выводов должны быть проведены на надёжном материале: изображения должны быть максимально точно извлечены, их количество должно быть близко к тому, сколько их в исходном объекте, мы должны учитывать особенности объекта исследования, влияющие на процесс извлечения изображений. Так мы сможем прийти к дальнейшему чтению изображений и их сопоставлению текстам. Код программы: github.com/Alina-meow/courier.

Список источников

1. Wollney E. Feminine, Competent, Submissive: a Multimodal Analysis of Depictions of Women in US Wartime Persuasive Messages during World War I and World War II / E. Wollney, M. Sternadori // *Visual Communication Quarterly*. 2019. No. 26.1. Pp. 3–21.
2. Sengupta T. Heroes and Villains: Multimodal Identity Construction in Children's Wartime Visual Narratives / T. Sengupta // *Multimodal Communication*. 2021. No. 10.3. Pp. 265–288.

3. Hiippala T. Distant Viewing and Multimodality Theory: Prospects and Challenges / T. Hiippala // *Multimodality & Society*. 2021. No. 1.2. Pp. 134–152.
4. Stern B. B. Historical and Personal Nostalgia in Advertising Text: the Fin de Siecle Effect / B. B. Stern // *Journal of Advertising*. 1992. No. 21.4. Pp. 11–22.
5. Lansdall-Welfare T. Content Analysis of 150 years of British Periodicals / T. Lansdall-Welfare, S. Sudhahar, J. Thompson et al. // *Proceedings of the National Academy of Sciences*. 2017. No. 114.4. Pp. E457–E465.
6. Drouin J. Close- and Distant-reading Modernism: Network Analysis, Text Mining, and Teaching the Little Review / J. Drouin // *The Journal of Modern Periodical Studies*. 2014. No. 5.1. Pp. 110–135.
7. Wevers M. The Visual Digital Turn: Using Neural Networks to Study Historical Images / M. Wevers, T. Smits // *Digital Scholarship in the Humanities*. 2020. No. 35.1. Pp. 194–207.
8. Arnold T. Enriching Historic Photography with Structured Data using Image Region Segmentation / T. Arnold, L. Tilton // *Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access*. 2020. Pp. 1–10.
9. Porter L. H. "Our International Journal": UN Publications and Soviet Internationalism after Stalin / L. H. Porter // *The Russian Review*. 2021. No. 80.4. Pp. 641–660.
10. Stoczkowski W. UNESCO's Doctrine of Human Diversity: a Secular Soteriology? / W. Stoczkowski // *Anthropology Today*. 2009. No. 25.3. Pp. 7–11.
11. Smits T. A Multimodal Turn in Digital Humanities: using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections / T. Smits, M. Wevers // *Digital Scholarship in the Humanities: a Journal of the Alliance of Digital Humanities Organizations*. 2023. Pp. 1–14.
12. Henning C. A. Estimating the Information Gap between Textual and Visual Representations / C. A. Henning, R. Ewerth // *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 2017.
13. Chen H. Cross-modal Image-text Retrieval with Semantic Consistency / H. Chen et al. // *Proceedings of the 27th ACM International Conference on Multimedia*. 2019.
14. Kaur P. Comparative Analysis on Cross-modal Information Retrieval: a Review / P. Kaur, H. S. Pannu, A. K. Malhi // *Computer Science Review*. 2021. No. 39. P. 100–336.
15. Lee B. C. G. The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in *Chronicling America* / B. C. G. Lee et al. // *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020.
16. Oberbichler S. Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians / S. Oberbichler et al. // *Journal of the Association for Information Science and Technology*. 2022. No. 73.2. Pp. 225–239.
17. Van Strien D. Computer Vision for the Humanities: an Introduction to Deep Learning for Image Classification. P. 1 / D. Van Strien et al. // *The Programming Historian* 2022.
18. PDF24 Tools. URL: tools.pdf24.org/ru.
19. Transkribus. URL: readcoop.eu/transkribus.
20. P2PaLa. URL: readcoop.eu/transkribus/docu/p2pala.

И. С. Гурьянов,

старший преподаватель кафедры прикладной информатики в креативных и культурных индустриях, заведующий лабораторией компьютерной графики
Сибирский федеральный университет, Красноярск, Россия

ТЕНДЕНЦИИ РАЗВИТИЯ ИСТОРИЧЕСКИХ ГИС

Аннотация. В статье рассмотрены причины возникновения и история развития исторических геоинформационных систем. Выделены четыре этапа. Каждый этап отличается источниками информации, способами и методами её обработки. Рассмотрены примеры актуальных междисциплинарных исследований. Также дан прогноз развития направления.

Ключевые слова: ГИС, историческая ГИС, базы данных, источник данных, тенденции, развитие, этапы, примеры, технологии

Развитие и внедрение в повседневную жизнь навигационного оборудования и программного обеспечения генерируют и требуют больше данных, большую точность и большой трафик. За последние 50 лет накопленная масса электронных географических данных, которая переходит из актуальной в историческую. За последние тысячелетия созданы географические карты, которые сейчас переходят в цифровой вид. Эти и многие другие факторы говорят о нарастающей актуальности геоинформационных систем параллельно с развитием цивилизаций. Сравнительно недавно на стыке информатики и истории, а также других наук возникли исторические ГИС. «Цифровой поворот» в гуманитарном знании позволил объединить исторические и географические данные, что позволило развиваться исторической геоинформатике [1].

Историю развития исторических ГИС можно условно разделить на следующие ключевые этапы.

1. Появление исторических ГИС (1980–90-е гг.). Начало разработки исторических ГИС, специально предназначенные для анализа и визуализации исторической информации. Они позволяли исследователям работать с аналоговыми географическими данными, такими как карты, атласы, фотографии и другие источники. Благодаря появлению программного обеспечения, профильных ресурсов и сообществ, а также открытым картографическим данным в Китае, Германии, России и многих других странах предпринимаются попытки создания национальных исторических ГИС-проектов [2].

2. Цифровизация исторических данных (1990–2000-е гг.). Происходит прогресс в цифровизации исторических данных. Исторические карты, документы и другие материалы были сканированы и преобразованы в электронный формат, что позволило их использование в исторических

ГИС. А. Ю. Володин отмечает, что оцифрованные данные меняют подход к исследованиям. Датификация происходит в различных гуманитарных дисциплинах и позволяет гармонизировать и переосмыслить подход к анализу данных [3].

3. Развитие методологии и анализа (2000–2010 гг.). Новые инструменты, методологии и подходы к анализу исторической информации с помощью ГИС. Исследователи начали применять пространственный анализ и моделирование для изучения исторических процессов и событий. Например, В. Н. Володин с коллегами применяет различные современные технологии в своих проектах: *web*-интерфейс, 3D-карты, игровой движок [4].

4. Интеграция с другими источниками данных (2010 г. – настоящее время). С развитием технологий и доступности больших объёмов данных исторические ГИС стали интегрироваться с другими источниками информации, такими как архивы, открытые источники, базы данных и геоинформационные системы. Это позволяет исследователям получать более полную картину прошлого и проводить более точные анализы. Исторические ГИС разнообразны в своем применении, каждая имеет свои цели, которые ставят исследователи в зависимости от области их научных интересов. На основе данных переписи населения Тобольска 1897 г. произведена реконструкция городского пространства, что актуально для урбанистов и историков [5]. На основе делопроизводственной документации, книг, примечаний к картам, переписных книг и других массовых источников изучают возникновение Белгородской черты [6]. На основе цифровых копий военных архивных документов создан банк данных мемориальных объектов, разрабатываются веб-карты для широкого использования [7]. На основе картографических и литературных источников Новосибирской области разработана методика создания исторической этнополитической ГИС [8].

Актуальные примеры использования ГИС-систем совместно с данными из различных областей знаний, таких как история, урбанистика, литература, экономика, делопроизводство, дизайн, говорят о широкой заинтересованности гуманитарных учёных во взаимодействии с методами цифровой картографии и базами данных. Учитывая развитие дигитализации и открытость данных, в будущем гуманитарная наука трансформируется, улучшая программное обеспечение исследований, расширяя область источников информации, становясь более доступной и открытой для различных пользователей. Развитие *Digital Humanities* позволит открыть множество интересных и эффективных методов исследования человеческой природы, что позволит нам узнать себя лучше и получить новые, ранее не встречавшиеся знания.

Список источников

1. Фролов А. А. «Цифровой поворот» в исторической науке и историческая геоинформатика / А. А. Фролов // Историческая информатика. 2019. № 3 (29). URL: cyberleninka.ru/article/n/tsifrovoy-povorot-v-istoricheskoy-nauke-i-istoricheskaya-geoinformatika.

2. Брюханова Е. А. Исторические ГИС online: обзор зарубежных и отечественных проектов / Е. А. Брюханова, М. В. Рыгалова // Известия АлтГУ. 2016. № 2 (90). URL: cyberleninka.ru/article/n/istoricheskie-gis-on-line-obzor-zarubezhnyh-i-otechestvennyh-proektov.

3. Володин А. Ю. Между Data и Карта: проблемы датафикации исторических исследований / А. Ю. Володин // Вестник ПГНИУ. Сер.: История. 2019. № 3 (46). URL: cyberleninka.ru/article/n/mezhdu-data-i-karta-problemy-datafikatsii-istoricheskikh-issledovaniy.

4. Латкин В. А. Технологические подходы и прикладные аспекты 3D-картографирования Транссибирской магистрали (на примере Тарманчуканского тоннеля) / В. А. Латкин, Е. П. Крупочкин, В. Н. Владимиров // Историческая информатика. 2022. № 1 (39). URL: cyberleninka.ru/article/n/tehnologicheskie-podhody-i-prikladnye-aspekty-3d-kartografirovaniya-transsibirskoy-magistrali-na-primere-tarmanchukanskogo-tonnelya.

5. Крупочкин Е. П. Реконструкция городского пространства Тобольска рубежа XIX–XX вв. средствами ГИС-технологий / Е. П. Крупочкин, Е. А. Брюханова, М. В. Рыгалова // Журнал СФУ. Гуманитарные науки. 2023. № 7. URL: cyberleninka.ru/article/n/rekonstruktsiya-gorodskogo-prostranstva-tobolska-rubezha-xix-xx-vv-sredstvami-gis-tehnologiy.

6. Жиров Н. Массовые источники XVII – первой половины XIX в. исторического региона белгородской черты: информационные возможности и проблемы обработки при создании ГИС / Н. Жиров // История: факты и символы. 2023. № 2 (35). URL: cyberleninka.ru/article/n/massovye-istochniki-xvii-pervoy-poloviny-xix-vv-istoricheskogo-regiona-belgorodskoy-cherty-informatsionnye-vozmozhnosti-i-problemy.

7. Требушкова И. Е. Использование средств геоинформационных систем в инвентаризации, исследовании и сохранении военно-исторических объектов Курской области / И. Е. Требушкова, А. И. Блинов // Географическая среда и живые системы. 2022. № 2. URL: cyberleninka.ru/article/n/ispolzovanie-sredstv-geoinformatsionnyh-sistem-v-inventarizatsii-issledovanii-i-sohranении-voenno-istoricheskikh-obektov-kurskoy.

8. Баталов Р. Н. Разработка методики исторической этнополитической ГИС / Р. Н. Баталов, Л. К. Радченко // Интерэкспо Гео-Сибирь. 2020. № 1. URL: cyberleninka.ru/article/n/razrabotka-metodiki-istoricheskoy-etnopoliticheskoy-gis.

А. А. Акашева,

кандидат исторических наук, доцент

Национальный исследовательский Нижегородский государственный университет
им. Н. И. Лобачевского, Нижний Новгород, Россия

e-mail: annakasheva@yandex.ru

ИНФОРМАЦИОННЫЕ РЕСУРСЫ АССОЦИАЦИИ «ИСТОРИЯ И КОМПЬЮТЕР»

Аннотация. В статье представлена и описана система информационных ресурсов межрегиональной профессиональной Ассоциации «История и компьютер». В неё входят официальный сайт Ассоциации, сайт её главного периодического издания журнала «Историческая информатика», *YouTube*-канал и группа в социальной сети «ВКонтакте».

Ключевые слова: общественные организации, сайты, ресурсы, Ассоциация «История и компьютер», АИК

В 1992 г. на постсоветском пространстве возникла старейшая профессиональная ассоциация в области гуманитарных наук – Ассоциация «История и компьютер» (далее – АИК). «История её создания и развития тесно связана с процессом становления и институционализации в странах Европы нового научного направления – исторической информатики» [1].

В настоящее время АИК имеет в интернете систему информационных ресурсов в виде своего официального сайта, сайта журнала «Историческая информатика», *YouTube*-канала и группы в социальной сети «ВКонтакте». Охарактеризуем их подробнее.

За всё время существования у АИК было несколько сайтов:

- 1) с 1999 г. – kleio.dcn-asu.ru;
- 2) с 2002 г. – aik.barnaul.ru;
- 3) с февраля 2004 г. – kleio.asu.ru;
- 4) с мая 2004 г. – aik.org.ru;
- 5) с 2006 г. – aikonline.ru;
- 6) с 2011 г. – aik-sng.ru.

Вся информация об URL сайтов восстановлена нами [2]. В настоящий момент официальным ресурсом является седьмая версия сайта, расположенная по адресу aik-hisc.ru. Интересными для пользователя являются разделы «Библиотека АИК», «Члены Ассоциации», «Конференции АИК». С этого сайта есть переходы на *YouTube*-канал и на группу «ВКонтакте». На сайте регулярно публикуют анонсы и новости о деятельности АИК, размещаются объявления членов Ассоциации о вышедших учебных пособиях, изданиях и прочих новостях.

Уже 11 лет Ассоциацией издаётся журнал «Историческая информатика» (список ВАК по историческим наукам, с 2023 г. в т. ч. по культурологии – 5.6.8. Документалистика, документоведение, архивоведение). Его статьи публикуются в открытом доступе на платформе издательства *Nota Bene* (URL: e-notabene.ru/istinf). В журнале отражаются достижения учёных из России и стран СНГ в области применения компьютерных технологий в истории, архивоведении, музееведении, а также в высшем историческом образовании. Статьи организованы в 18 рубрик, освещающих актуальные вопросы развития исторической информатики. С 2012 по 2016 гг. включительно публикации журнала можно посмотреть на сайте kleio.asu.ru [3].

Записи всех мероприятий АИК, проходящих в гибридном и онлайн-формате, можно посмотреть и пересмотреть на собственном *YouTube*-канале, который организован в ноябре 2020 г. и доступен по прямой ссылке (URL: youtube.com/channel/UCWDdUbeqpp4Vt5DipXMHBVQ) или через любую поисковую систему по тегу *@AIKtube* [4]. Сейчас на канале размещено почти 30 видео, объединённых в 6 плейлистов.

Наконец, в январе 2021 г. у АИК появилось сообщество «ВКонтакте» (URL: vk.com/aiksng) [5]. В сообществе есть рубрики «Кто есть кто в АИК» с портретами учёных-аиковцев и «Лучшая статья номера», где даётся обзор одной из наиболее интересных статей в свежем выпуске журнала «Историческая информатика», также приведены сведения обо всех конференциях АИК и условия вступления в Ассоциацию. Сообщество можно найти по хештэгу *#aiksng* в любом поисковике.

Список источников

1. Об АИК // Ассоциация «История и компьютер»: оф. сайт. URL: aik-hisc.ru/about_us.
2. Запрос по ключевым словам «ассоциация», «история» и «компьютер» // Internet Archive WaybackMachine. URL: web.archive.org/web/20230000000000*/Ассоциация%20история%20и%20компьютер.
3. Историческая информатика. Информационные технологии и математические методы в исторических исследованиях и образовании: оф. сайт. URL: kleio.asu.ru.
4. Ассоциация «История и компьютер» (АИК): оф. канал // YouTube: видеохостинг. URL: youtube.com/channel/UCWDdUbeqpp4Vt5DipXMHBVQ.
5. Ассоциация «История и компьютер»: оф. сообщество // ВКонтакте: соц. сеть. URL: vk.com/aiksng.

Е. В. Брылевская¹, Н. П. Федотова² (науч. руководитель),
² кандидат физико-математических наук, доцент

^{1,2} Ярославский государственный университет им. П. Г. Демидова, Ярославль, Россия

НЕКОТОРЫЙ АЛГОРИТМ ДЛЯ ПОСТРОЕНИЯ СЕМАНТИЧЕСКИХ СЕТЕЙ

Семантические сети представляют собой графовые, матричные и векторные модели, используемые для описания связей между понятиями и терминами. Одним из ключевых инструментов в их построении является матрица смежности, которая отражает степень связей между понятиями в виде числовых значений. Эти значения позволяют количественно оценить семантическую близость или удалённость между понятиями [1]. Семантические сети имеют широкий спектр приложений в различных областях, таких как информационный поиск, автоматическое реферирование и категоризация текстов, машинном переводе, рекомендательных системах, анализе социальных сетей.

Примеры алгоритмов построения семантических сетей.

1. Алгоритм *WordNet* является одним из наиболее известных ресурсов для построения семантических сетей, строит сеть синонимов, где каждый узел представляет собой синонимичную группу слов, а рёбра обозначают отношения между этими словами, такие как гиперонимия, гипонимия.

2. Алгоритм *Latent Semantic Analysis (LSA)* использует матричные вычисления и сингулярное разложение для выделения семантических паттернов в больших наборах текстовых данных.

3. Алгоритм *Distributional Semantic Models (DSM)* основан на идее, что слова, которые встречаются в похожих контекстах, имеют схожую семантику.

4. Алгоритм *Random Walk* (случайное блуждание) используется для построения семантической сети на основе вероятностей перехода между словами в текстовом корпусе.

5. Алгоритм *Deep Walk* является методом графового вложения, который строит семантическую сеть, используя идеи из области нейронных сетей.

В большинстве своём, все эти алгоритмы базируются на принципах построения матрицы смежностей. Благодаря матричной нотации возможно провести семантический анализ входного текста, но отсюда вытекает проблема – контекстуально обусловленная вариативность единиц языка. Одним из решений этой проблемы является метод взвешивания слов.

Для своего исследования я применила формулу *TF-IDF*. Формула *TF-IDF* вычисляется как произведение двух значений:

$$tf(t, d) = \frac{n_t}{\sum_k n_k}, \quad (1)$$

где n_t – число вхождений слова t в документ, а в знаменателе – общее число слов в данном документе.

$$idf(t, D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}, \quad (2)$$

где $|D|$ – число документов в корпусе; $|\{d_i \in D | t \in d_i\}|$ – число документов из коллекции D , в которых встречается t (когда $n_t \neq 0$).

Таким образом, *TF-IDF* является произведением двух сомножителей:

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D). \quad (3)$$

Формула *TF-IDF* не просто подсчитывает частоту лексических повторов. Формула *TF-IDF* вычисляется путём умножения двух значений: частоты термина (*TF*) и обратной частоты документа (*IDF*). Частота термина *TF* показывает, сколько раз термин встречается в данном документе, а обратная частота документа *IDF* показывает, насколько редко этот термин встречается во всех документах коллекции. Таким образом, если термин встречается часто в данном документе, но редко в остальных документах коллекции, его семантическая значимость в данном документе будет высокой [2]. Воспользуемся данным алгоритмом для семантического анализа произведения М. И. Цветаевой «Август» 1917 г. Для этого мне понадобится язык программирования *Python* и сторонняя библиотека *sklearn* [3]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
text = "Август – астры,\nАвгуст – звезды,\n\nАвгуст – грозды\nВинограда и рябины\n\nРжавой – август!\n\nПолновесным,\n\nблагосклонным\n\nЯблоком своим имперским,\n\nКак дитя,\n\nиграешь, август.\n..."
tfidf = TfidfVectorizer()
matrix = tfidf.fit_transform([text])
words = tfidf.get_feature_names()
values = matrix.toarray()[0]
tfidf_dict = dict(zip(words, values))
for word, value in sorted(tfidf_dict.items(),\key=lambda x:
x[1], reverse=True):
    print(f"{word}: {value:.3f}")
```

В результате выполнения кода будет выведен список слов из стихотворения и их значения *TF-IDF* в порядке убывания обратных частот (табл. 1).

Список слов и их значения *TF-IDF*

Термин	Обратная частота
Август	0,190
звездных	0,301
месяц	0,301
...	...
ладонью	0,889
гладишь	0,889
именем	0,889

С полученными результатами можно работать при помощи аппарата дистрибутивной семантики, всё зависит от цели и методов исследования.

Список источников

1. Андреев А. В. Введение в формальную семантику / А. В. Андреев, О. А. Митрофанова, К. В. Соколов. СПб.: СПбГУ, 2014.
2. Christopher D. Introduction to Information Retrieval / D. Christopher, P. P. Manning, H. Schutze // Cambridge University Press. 2014. P. 134.
3. Библиотека sklearn. URL: github.com/scikit-learn/scikit-learn.

УДК 81'33+004.912

М. А. Бутусова,

магистрант

Южный федеральный университет, Ростов-на-Дону, Россия

e-mail: mbutusova@sfedu.ru

ORCID: 0009-0000-6433-3611

СЕТЕВОЙ АНАЛИЗ РЕЧИ ГЕРОЕВ ПЬЕС А. П. ЧЕХОВА «ТРИ СЕСТРЫ» И «ВИШНЁВЫЙ САД»

Аннотация. В данном исследовании проводится сетевой анализ пьес «Три сестры» и «Вишнёвый сад» А. П. Чехова с целью семантической разметки выявленных характеристик. С помощью методов стилометрии и сетевого анализа формируется алгоритм классификации героев драматических произведений на основе семантической близости их речи. Вводятся специфические значения атрибутов *type* и *subtype* тега *castItem* (стандарт *Text Encoding Initiative*) для разметки соответствующих групп в семантическом издании *Chekhov Digital*.

Ключевые слова: стилометрический анализ, семантическая разметка, сетевой анализ, семантическое издание, *Chekhov Digital*

Исследование проводится в рамках подготовки семантического издания *Chekhov Digital*, которое разрабатывается с использованием стандарта *Text Encoding Initiative* (далее – *TEI*) на основе Полного собрания сочинений и писем А. П. Чехова в 30 томах [9], изданного в 1974–1983 гг. Такое цифровое представление текстов делает их машиночитаемыми и открывает новые возможности для исследовательской работы. На сегодняшний день семантическое издание содержит тексты произведений I–X т. Полного собрания сочинений и тексты Полного собрания писем писателя (12 томов), в которых проведена структурная разметка в соответствии с форматом печатного издания и включены элементы семантической разметки. Среди задач проекта – семантическая разметка драматического наследия А. П. Чехова и дополнение его переводами на английский язык.

Семантическая разметка – это набор тегов, с помощью которых помечаются значения слов, их связи, контексты и др., благодаря чему исследователям доступны различные средства автоматической компьютерной обработки для исследовательской работы по соотношению фактов и выявлению связей в тексте [6, с. 250]. Основой семантической разметки в издании *Chekhov Digital* служит язык *XML* (англ. *extensible markup language* – расширяемый язык разметки) и стандарт *TEI*, принятый в области цифро-

вых гуманитарных наук и определяющий набор тегов и их атрибутов для разметки структурно-семантических единиц текстов.

Включение текстов в издание требует изучения пьес, определения их специфики и разработки концепции разметки. В настоящей работе приводятся результаты исследования речи героев пьес А. П. Чехова «Три сестры» и «Вишнёвый сад» с использованием методов стилометрического и сетевого анализа (набор интегрированных техник для отражения отношений между актёрами и анализа социальных структур), который позволяет увидеть взаимоотношения между элементами сети и понять роль каждого элемента [2, с. 10 499].

На материале текстов выбранных пьес был проведён стилометрический анализ. Тексты пьес были взяты на сайте Фундаментальной электронной библиотеки и распространяемого по открытой лицензии (*CC BY*) [1], преобразованы в формат *plain text*, с помощью средств ЯП *Python* разделены на отдельные единицы, каждая из которых представляет собой текстовый файл, содержащий речь одного персонажа и названный его именем. Анализ полученного корпуса текстов проводился с помощью функции *stylo.network()*¹⁴ языка программирования *R*, которая позволяет построить частотный словарь с ограниченным количеством позиций для каждого подаваемого текста (в данном случае заданы следующие параметры *MFW Settings: minimum – 100, maximum – 1 000*) [3]. На основе полученных частотных словарей автоматически была высчитана семантическая близость текстов с мерой межтекстового расстояния Классическая Дельта (*Classic Delta*), предложенная Дж. Берроузом в 2001 г. и широко применяемая в стилометрических исследованиях [5, с. 2]. Результаты анализа были представлены в виде дендрограммы (построена с помощью метода *Bootstrap Consensus Tree*), а также матриц узлов и рёбер, где узел – это реплики одного героя (узел обозначен его именем), а рёбра – семантическая близость между речью соединяемых героев.

В данной работе для визуализации результатов, полученных в ходе стилометрического анализа, использована программа *Gephi*¹⁵. Граф, построенный на основе семантической близости реплик героев пьесы «Три сестры», показал высокую степень связности (9,95) и плотности (0,813), что означает в целом высокую семантическую близость речи героев. При этом три персонажа выполняют роль «посредников» (алгоритм *Eigenvector Centrality*), т. к. их речь больше других связана с речью персонажей с высокой степенью связности: это реплики Анфисы, Родэ и Федотика. Такой результат объясняется техническими причинами: объём речи этих героев невелик, и состоит она в основном из междометий, местоимений и прочих высокочастотных слов (Анфиса: «Пойду и я поищу. Машенька, ау! А-у, а-у!»), в то время как речь главных героев связана с тема-

¹⁴ Eder M. *Stylometry with R: a Package for Computational Text Analysis* / M. Eder, J. Rybicki, M. Kestemont // *R Journal*. 2016. No. 8 (1). Pp. 107–121. URL: journal.r-project.org/archive/2016/RJ-2016-007/index.html.

¹⁵ Bastian M. *Gephi: an Open Source Software for Exploring and Manipulating Networks* / M. Bastian, S. Heymann, M. Jacomy // *International AAAI Conference on Weblogs and Social Media*. 2009.

ми, которые могут не встречаться у других персонажей. Поэтому речь героев с малым объёмом реплик не учитывалась.

С помощью методов стилометрии были выделены две группы семантически близких текстов: в первую группу вошли реплики Маши, Ольги, Ирины, Андрея, Вершинина и Тузенбаха; во вторую – реплики прочих персонажей. Полученные результаты схожи с делением на главных и второстепенных персонажей: в данном случае предполагается, что речь главных героев схожа, т. к. они раскрывают одни и те же ключевые для произведения темы и используют схожую лексику, а речь второстепенных героев носит служебный характер. Однако после исключения трёх узлов (реплики Анфисы, Родэ и Федотика) стилометрический анализ показал наличие третьей группы, в которую входят реплики двух персонажей: Ирина и Ольга, речь которых несколько отличается от речи других представителей первой группы. Таким образом, сетевой анализ может быть использован для автоматической кластеризации персонажей с семантически близкой речью и выведением классификации на её основе.

Такое же исследование было проведено для текста пьесы «Вишнёвый сад». Стиллометрический анализ показал, что речь героев довольно сильно связана, но в меньшей степени, чем речь героев пьесы «Три сестры»: средняя степень – 8,571; плотность графа – 0,659. При исследовании реплик персонажей «Вишнёвого сада» были выделены две группы: в первую вошли реплики Любви Андреевны, Лопухина, Гаева и Трофимова, во вторую – речь остальных героев. После исключения персонажей с малым объёмом реплик (Пищик, Прохожий и Начальник станции) результаты кластеризации не изменились.

Таким образом, сетевой анализ, проведённый на основе стиллометрических данных, позволяет разделить речь героев на группы. Подобный подход может быть использован для автоматической классификации героев пьес (на основе кластеризации их речи) и обозначения их класса в разметке семантического издания. Отметка о принадлежности героя к той или иной группе позволит автоматизировать поиск по речи героев. Полученная классификация не является строгим разделением героев на главных и второстепенных, поэтому они будут помечены как *group_1* и *group_2* в качестве атрибутов *type* тега *castItem*, что потребует введения нового значения атрибута *type* (в настоящее время в документации *TEI* описано два возможных значения атрибута *type* тега *castItem*: *role* (если тегом размечен один герой) и *list* (если тегом размечен список героев)) [4]; наличие третьей группы (как в примере с пьесой «Три сестры») может быть отмечено в атрибуте *subtype*. В таком случае разметка имени героя Маша будет выглядеть следующим образом: `<castItem type='group_1' subtype='subgroup_1'>Маша</castItem>`, а героя Ольга – так: `<castItem type='group_1' subtype='subgroup_2'>Ольга</castItem>`.

Для второй группы также будет полезным деление на подгруппы для пометки героев с малым объёмом реплик. Стиллометрический анализ показал, что включение их речи в исследование может исказить общий ре-

зультат, поэтому для них введены подтипы второй группы: *major* (для основной части героев второй группы) и *minor* (для персонажей с малым объёмом реплик). Разделение героев второй группы на подтипы осуществляется на основе объёма речи: если количество слов героя превышает определённый порог, то герой относится к подгруппе *major* и размечается соответственно (`<castItem type='group_2' subtype='major'>Аня</castItem>`); если герой произносит менее 120 слов, он отмечается подгруппой *minor* (например, `<castItem type='group_2' subtype='minor'>Прохожий</castItem>`). Порог в 120 слов определён для пьесы «Три сестры», в которой высокие показатели по мерке *Eigenvector Centrality* были выявлены в речи трёх персонажей: Анфисы, Федотика и Родэ, поэтому все они были исключены из исследования на следующем этапе как герои с недостаточным для анализа объёмом реплик.

Некоторая стандартизация всё же необходима, чтобы определение группы как первой или как второй было последовательным. Типом *group_1* будет отмечена меньшая по размеру группа, т. к. центральных персонажей, как правило, меньше, чем второстепенных, а типом *group_2* – остальные персонажи: в пьесе «Три сестры» к первой группе относится речь 42,86 % героев, а в пьесе «Вишнёвый сад» – 28,57 %.

В результате данного исследования предложен метод классификации героев пьес на основе семантической близости их речи, измеренной с помощью методов стилометрии и сетевого анализа. Полученная классификация может быть использована при разметке пьес в семантическом издании *Chekhov Digital*.

Исследование выполнено в рамках соглашения о научном сотрудничестве между Южным федеральным университетом (ЮФУ) и Национальным исследовательским университетом «Высшая школа экономики» (НИУ ВШЭ) («Зеркальные лаборатории НИУ ВШЭ»), проект № 6.13.1-02/250821-1 «Конвергенция языковых пластов русского языка в зеркале цифровых решений».

Список источников

1. About CC Licenses. Creative Commons. URL: creativecommons.org/about/cclicenses.
2. Chiesi A. M. Network Analysis / A. M. Chiesi; ed.: N. J. Smelser, P. B. Baltes // International Encyclopedia of the Social & Behavioral Sciences. Pergamon, 2001. Pp. 10 499–10 502. DOI: 10.1016/B0-08-043076-7/04211-X.
3. stylo.network: Bootstrap Consensus Networks, with D3 Visualization // R Package Documentation. URL: rdrr.io/cran/stylo/man/stylo.network.html.
4. TEI Consortium, eds. "<castItem>" Guidelines for Electronic Text Encoding and Interchange. 2023. URL: tei-c.org/release/doc/tei-p5-doc/en/html/ref-castItem.html.
5. Мамаев Н. Метод дельты Бёрроуза для определения авторства анонимных и псевдонимных литературных произведений на русском языке / Н. Мамаев // The 2nd International Conference "R. Piotrowski's Readings LE & AL'2017 St. Petersburg". 2018. URL: pure.spbu.ru/ws/portalfiles/portal/36984115/_._._.pdf.
6. Орехов Б. В. Что такое семантическое издание и почему в будущем все издания станут семантическими? / Б. В. Орехов, М. Гронас // A/Z: Essays in Honor of Alexander Zholkovsky. Boston MA: Academic Studies Press, 2018. С. 246–268.

7. Северина Е. М. Новые филологические практики: семантическое издание текстов А. П. Чехова / Е. М. Северина, М. Ч. Ларионова // Филология: научные исследования. 2020. № 10. DOI: 10.7256/2454-0749.2020.10.33970. URL: nbpublish.com/library_read_article.php?id=33970.

8. Сушко В. А. История возникновения и методология сетевого анализа / В. А. Сушко // ЖССА. 2018. № 1. URL: cyberleninka.ru/article/n/istoriya-vozniknoveniya-i-metodologiya-setevogo-analiza.

9. Чехов А. П. Полное собрание сочинений и писем: в 30 т. / АН СССР. ИМЛИ им. А. М. Горького. М.: Наука, 1974–1983. URL: febweb.ru/feb/chekhov/default.asp?feb/chekhov/texts/che-te02.html.

В. Д. Ведерникова¹, Е. М. Северина² (науч. руководитель),
¹ магистрант;

² доктор философских наук, профессор

^{1,2} Южный федеральный университет, Ростов-на-Дону, Россия

e-mail: ¹ vedernikova@sfedu.ru, ² emkovalenko@sfedu.ru

ORCID: ¹ 0009-0003-6251-0866, ² 0000-0001-6518-2771

ЦИФРОВОЙ АНАЛИЗ ИДИОСТИЛЯ АВТОРА И ПЕРЕВОДЧИКА (НА МАТЕРИАЛЕ ТЕКСТОВ ПРОИЗВЕДЕНИЙ ДЖ. К. РОУЛИНГ)

Аннотация. Работа посвящена изучению идиостиля автора и переводчика с помощью цифровых методов на материале оригинальных текстов серии Дж. К. Роулинг «Гарри Поттер» и их переводов на русский язык, опубликованных издательствами «Махаон» и «Росмэн». Используются методы стилометрии: функции *stylo()* (кластерный анализ), *oppose()* (контрастивный анализ) и *rolling.classify()* пакета *Stylo* для языка *R*. Выявлены различия в использовании переводных эквивалентов в их коннотативном значении, показана возможность исследования изменений идиостиля автора и переводчика под воздействием внешних факторов с помощью цифровых методов.

Ключевые слова: *стилометрия, идиостиль переводчика, идиостиль автора, кластерный анализ*

Сегодня одним из актуальных вопросов переводоведения является изучение идиостиля автора и переводчика с помощью цифровых методов атрибуции текста [1]. Трудность такого исследования связана с тем, что установление уникальных стилистических характеристик переводчика возможно только в сравнении с такими же характеристиками автора оригинального текста. Кроме того, при создании автором одной или нескольких серий книг возможны значительные изменения в его идиостиле под влиянием внешних факторов – времени и литературного жанра. Изменение идиостиля автора может повлиять на появление новых стилистических характеристик идиостиля переводчика при последующем переводе этих серий книг. Так, научный интерес данного исследования заключается в том, чтобы выяснить, как изменения в авторском стиле под влиянием внешних факторов отражаются на идиостиле переводчика и возможно ли это определить с помощью цифровых методов.

Для исследования данного вопроса были взяты оригинальные тексты серии «Гарри Поттер» Дж. К. Роулинг и их переводы на русский язык, опубликованные издательствами «Махаон» и «Росмэн». Переводы издательства «Махаон» осуществлены одним переводчиком, М. Спивак [2], в то время как переводы издательства «Росмэн» выполнены группой переводчиков под руководством главного редактора М. Д. Литвиновой [3].

Идиостиль автора и переводчика изучались методами стилометрического анализа. «Стилометрия – это статистический анализ отклонений

между литературными стилями разных авторов и жанров» [4]. СтилOMETрические методы во всём их разнообразии имеют две общие черты: текстовые элементы должны быть каким-то образом преобразованы в числа, а числа уже исследованы статистическими методами [5]. Классическим стилOMETрическим методом исследования идиостиля автора является кластерный анализ. Материал исследования был преобразован в формат *plain text*, в кодировке *UTF-8*.

Кластерный стилOMETрический анализ позволяет провести разделение текстов на сравнительно однородные группы (кластеры). Для исходных (непредобработанных) оригинальных и переводных текстов кластеризация проводилась с помощью функции *stylo()* пакета *Stylo* для языка *R*. Для изучения авторского и переводческого идиостиля дополнительно были взяты следующие тексты: для анализа оригинальных текстов – тексты серии «*Cormoran Strike*» *R. Galbraith (J. K. Rowling)* («*The Cuckoo's Calling*», «*The Silkworm*», «*Career of Evil*», «*Lethal White*») и для анализа текстов перевода на русский язык – авторские и переводные тексты М. Спивак («Твари, подобные Богу», «Чёрная магия с полным её разоблачением» и перевод К. Эдварса «Дочь хранителя тайны»).

Кластерный стилOMETрический анализ оригинальных текстов показал, что оригинальные тексты подразделяются на два кластера: первый кластер состоит из всех текстов серии «*Harry Potter*» и первых двух текстов серии «*Cormoran Strike*» («*The Cuckoo's Calling*», «*The Silkworm*»), второй кластер – из третьей и четвёртой текстов серии «*Cormoran Strike*» («*Career of Evil*», «*Lethal White*»). Внутри кластера текстов серии «*Harry Potter*» также есть разделение на кластеры: в первый вошли тексты серии с 1 по 4, во второй – с 5 по 7.

Такая кластеризация позволяет высказать предположение о схожести стилей написания серии «*Harry Potter*» и ранних работ *R. Galbraith (J. K. Rowling)*. Выделение более поздних работ серии «*Cormoran Strike*» в отдельный кластер позволяет предположить, что происходит изменение идиостиля автора, возможно, под воздействием требований жанра криминального романа. Кроме того, выделение в отдельные кластеры первого и седьмого текстов серии «*Harry Potter*» может также свидетельствовать о стилевых изменениях под воздействием внешних факторов: как известно, изначально первая книга «Гарри Поттера» была опубликована как самодостаточный роман, но коммерческий успех заставил автора писать продолжение [6].

Кластерный стилOMETрический анализ собранного корпуса текстов на русском языке разделил тексты на два кластера: в первый кластер вошли все тексты переводов на русский язык серии «Гарри Поттер» (независимо от переводчика), во второй – авторские и переводные тексты М. Спивак. Таким образом, авторский сигнал оказывается сильнее, чем сигнал переводчика. Кроме того, переводы издательства «Росмэн» и переводы издательства «Махаон» на русский язык были выделены в отдельные кластеры. При этом в переводе обоих издательств первая книга серии «Гарри Пот-

тер» была выделена в отдельный кластер так же, как и в оригинальных текстах, что подтверждает сохранение переводчиками авторского сигнала в переводах. Разделение остальных текстов по кластерам, отличных от кластеров оригинальных текстов, может рассматриваться как проявление идиостиля переводчика и требует дополнительного изучения.

Контрастивный стилметрический анализ текстов перевода с помощью функции *oppose()* пакета *Stylo* для языка *R* позволил определить характерные для каждого перевода переводные эквиваленты. Например, лексические формы *wizard* (*wizardy*) и *witch* (*witchcraft*) в текстах перевода имеют различные переводные эквиваленты: в переводе, выполненном переводчиком издательства «Махаон», – это «колдун» и «ведьма», а в переводе, выполненном переводчиками издательства «Росмэн», – «волшебник»/«чародей»/«колдун» и «ведьма»/«чародейка». Исследование семантики данных переводных эквивалентов на основе Русского семантического словаря, тонального словаря *PolSentiLex* (2016 г., платформа *Linis Crowd*) и их коллокаций в Национальном корпусе русского языка (НКРЯ) показало, что данные лексические формы передают различные коннотативные значения: слово «волшебник» чаще всего употребляется с лексическими формами «добрый» и «злой», причём почти в равных пропорциях и, соответственно, может передавать как положительную, так и негативную коннотацию, а слово «колдун» используется в основном в коллокации «злой колдун» и, соответственно, передаёт негативную коннотацию [7; 8]. В переводе издательства «Росмэн» различие в коннотативных значениях используется на композиционном уровне: с лексической формой «волшебник» отождествляются положительные персонажи, ученики Хогвартса, а с лексическими формами «чародей» и «колдун» – отрицательные персонажи. В переводе издательства «Махаон» данного различия нет: для обозначения и положительных, и отрицательных героев используется лексическая форма «колдун». Возможно, данное различие в употреблении переводных эквивалентов было одной из причин негативного отношения к переводу издательства «Махаон» поклонников истории «мальчика-который-выжил» [9].

С помощью функции *rolling.classify()* пакета *Stylo* для языка *R* было проведено исследование текстов перевода издательства «Росмэн» на предмет выявления вклада каждого переводчика в перевод текстов серии Дж. К. Роулинг «Гарри Поттер». Перевод пятого текста серии (переводчики: В. Голышев, В. Бабков, Л. Мотылев) представляет собой «равномерное смешение идиостилей» переводчиков, что может указывать на распределение текста перевода на отдельные равные части и их соответствующий перевод каждым переводчиком. Перевод четвертого текста серии (переводчики: М. Д. Литвинова, Е. Саломатина, А. Лях, М. Межуев) по большей части осуществлён М. Д. Литвиновой, хотя авторство перевода заявлено за несколькими переводчиками. При исследовании второго текста серии в переводах издательства «Махаон» (переводчик М. Спивак, 2014 г.) и издательства «Росмэн» (переводчик М. Д. Литвинова, 2001 г.) оказалось,

что идиостиль обоих переводчиков проявляется в каждом переводе, что свидетельствует, по-видимому, о силе авторского сигнала в сравнении с сигналом переводчика.

Таким образом, стилометрический анализ позволяет выявить специфику идиостиля не только автора, но и переводчика. Такие методы, как кластерный стилометрический анализ, контрастивный стилометрический анализ и анализ с помощью *rolling.classify* позволяют установить не только специфические различия в переводных эквивалентах, но и общую тенденцию изменений идиостиля и автора, и переводчика под воздействием внешних факторов.

Список источников

1. Скоринкин Д. Силометрия: как в разное время люди искали авторов текстов / Д. Скоринкин, А. Затонская // Системный Блокъ. 2021. URL: sysblok.ru/knowhow/stilometrija-kak-v-raznoe-vremja-ljudi-iskali-avtorov-tekstov.

2. Издательства реального мира. Махаон // Гарри Поттер вики: сообщество фэндома. URL: harrypotter.fandom.com/ru/wiki/%D0%9C%D0%B0%D1%85%D0%B0%D0%BE%D0%BD.

3. Издательства реального мира. Росмэн // Гарри Поттер вики: сообщество фэндома. URL: harrypotter.fandom.com/ru/wiki/%D0%A0%D0%9E%D0%A1%D0%9C%D0%AD%D0%9D.

4. Гагарина Д. А. Культурный код: искусство и культура в цифровой среде / Д. А. Гагарина. Калининград, 2018.

5. Eder M. Stylometry with R: a Package for Computational Text Analysis / M. Eder, J. Rybicki, M. Kestemont // The R Journal. 2016. Vol. 8. Pp. 107–121.

6. История создания Гарри Поттера // Сообщество PotterBase: мир волшебства. 2019. URL: potter-base.ru/117-istoriya-sozdaniya-garri-pottera.html.

7. Русский семантический словарь: толковый словарь, систематизированный по классам слов и значений / под общ. ред. Н. Ю. Шведовой. М.: Азбуковник, 1998.

8. Национальный корпус русского языка. URL: ruscorpora.ru.

9. Спивак М. Особенности перевода «Поттерианы»: издания / М. Спивак // Гарри Поттер вики: сообщество фэндома. URL: harrypotter.fandom.com/ru/wiki/%D0%9C%D0%B0%D1%80%D0%B8%D1%8F_%D0%A1%D0%BF%D0%B8%D0%B2%D0%B0%D0%BA.

П. И. Максименко,
студент, стажёр-исследователь
Национальный исследовательский университет «Высшая школа экономики»,
Санкт-Петербург, Россия
e-mail: pimaksimenko@edu.hse.ru
ORCID: 0000-0003-2621-6144

РУССКОЯЗЫЧНАЯ ЭЛЕКТРОННАЯ БАЗА ФАНФИКШН-ТЕКСТОВ: ПРИНЦИПЫ СОЗДАНИЯ И АНАЛИЗ МЕТАДААННЫХ

Аннотация. В работе рассматривается популярный феномен массовой любительской сетевой литературы – фанфикшн. Описывается создание электронной базы, включающей в себя более 135 тыс. текстов, размещённых на крупнейшем русскоязычном фанфикшн-ресурсе «Книга фанфиков» (URL: ficbook.net). Квантитативный анализ метаданных, входящих в электронную базу, позволяет исследовать явление фанфикшн в русскоязычной среде и выявить его значимые особенности.

Ключевые слова: фанфикшн, электронная база, метаданные, веб-скрейпинг

Введение

На сегодняшний день фанфикшн (англ. *fan fiction* – «фанатская литература») – один из наиболее активно развивающихся типов сетевой литературы. Каждый день на онлайн-платформах появляются десятки тысяч текстов, созданных фанатами по мотивам любимых произведений. Фанфикшн определяют как феномен любительской массовой литературы, основанный на использовании, расширении или переосмыслении художественного мира, созданного в каком-либо популярном произведении (как правило, литературном, кинематографическом, драматическом) [1].

В англоязычной среде фанфикшн становится объектом научного исследования уже в конце XX в. и интересует учёных прежде всего с социокультурной точки зрения [2–4]. Большое внимание в западных работах уделяется понятию фандома (англ. *fandom*), понимаемому как сообщество поклонников чего-либо или кого-либо, вовлечённых в различные дискурсивные и социальные практики, одними из которых и являются написание и чтение фанфикшн-текстов [5]. В то же время русскоязычный фанфикшн остаётся малоисследованным, предположительно ввиду того, что фан-тексты изначально оценивались как маргинальное явление, «низкое проявление паралитературы», «не заслуживающее серьёзного внимания» [6]. Позже исследователи начали рассматривать фанфикшн как порождение современной массовой (в частности фанатской) культуры и литературы, отражающее их черты [7]. Помимо вторичности текста по отношению к оригиналу (канону), к важным особенностям фанфикшн как литературного явления можно отнести его создание в рамках фанатских сообществ

и соответствие их стандартам, безвозмездный характер писательского творчества и большую сфокусированность на персонаже, чем на вымышленном мире [8].

Большинство русскоязычных исследований фанфикшн основаны на последовательном литературоведческом анализе и т. н. методе *close reading* [9; 10], в то время как иной подход, предполагающий создание корпусов и электронных баз, включающих выборки большого объёма, показывает свою эффективность [11–13]. Он позволяет наиболее полно проанализировать феномен фанфикшн (в рамках созданной выборки) и предоставляет возможность применения многочисленных компьютерных методов.

1. Источник данных

В качестве источника данных для создания электронной базы был выбран наиболее популярный в России специализированный ресурс для публикации фанфикшн – «Книга фанфиков» (URL: ficbook.net), насчитывающий более 4 млн фан-текстов (также называемых фанфиками) по более чем 25 тыс. фандомам [15]. Внутренняя организация портала предполагает разделение фандомов по девяти категориям (в зависимости от типа оригинального произведения): «Аниме и манга», «Книги», «Мультфильмы», «Игры», «Фильмы и сериалы», «Известные люди», «Комиксы», «Мюзиклы», «Ориджиналы» и «Другое». Самой многочисленной является категория «Известные люди», к которой относятся около 1 млн работ и более 10 тыс. фандомов. За ней следуют категории «Аниме и манга» и «Фильмы и сериалы». Наиболее популярными фандомами являются южнокорейская поп-группа *Bangtan Boys (BTS)*, книги и фильмы о Гарри Поттере, аниме *Naruto*, *Fairy Tale*, компьютерная игра *Genshin Impact*. Отметим, что «Книга фанфиков» позволяет публиковать ориджиналы – авторские тексты, не опирающиеся на какое-либо произведение и не заимствующие чужих героев, элементы вымышленного мира и событий. Ориджиналы часто рассматриваются как ответвление фанфикшн, т. к. имеют схожую структуру, жанры и персонажей [16]. На портале также можно публиковать переводы иноязычных фан-текстов.

Ресурс предлагает пользователям развёрнутую систему тегов, характеризующих размер, статус написания, возрастной рейтинг, направленность, описываемых персонажей и тематику работы. Каждый фан-текст предваряет т. н. «шапка», содержащая название работы, никнейм автора и ссылку на его профиль, информацию по всем перечисленным параметрам и краткое описание сюжета фанфика.

2. Создание электронной базы

Электронная база была создана с помощью веб-скрейпинга – технологии автоматического извлечения данных из интернета, как правило, для последующего сохранения в файловой системе, извлечения и анализа [19]. Основное преимущество этой технологии заключается в том, что она отличается скоростью и высокой производительностью и подходит для сбора и обработки больших объёмов данных [19]. Для применения

данного метода на материале веб-сайта «Книга фанфиков» был создан веб-скрейпер – автоматизированная программа, которая совершает запросы на веб-сервер ресурса и производит извлечение информации путём парсинга данных. Веб-скрейпер был реализован в виде программы на языке *Python* с использованием библиотеки *Selenium*, представляющей собой *API WebDriver*, который функционирует как веб-браузер. С его помощью производится загрузка веб-страниц, извлечение с них данных, а также поиск элементов и взаимодействие с ними. Тексты работ были сохранены в виде файлов в формате *.txt*, а извлечённые метаданные представлены в табличном виде (*pandas DataFrame*).

3. Описание электронной базы

Полученная электронная база насчитывает 136 257 текстов, созданных 27 792 авторами в рамках 3 411 фандомов. Фан-тексты опубликованы в период с июня 2009 г. по январь 2013 г.¹⁶ 134 615 произведений написаны на русском языке, что составляет более 98 % от общего числа текстов, в то время как на переводы приходится всего 1,2 % электронной базы (1 642 текста).

Кроме того, в электронной базе содержатся метаданные обо всех текстах, а именно: название фанфика, ссылка на веб-страницу, где размещена работа, никнейм автора и ссылка на его профиль на платформе «Книга фанфиков», описание, предшествующее тексту, информация о фандоме, жанре и рейтинге, к которым отнесена работа, о персонажах (пейринге) и присвоенных работе метках. Важно также отметить, что все фан-тексты, вошедшие в электронную базу, имеют статус «Завершён» (по состоянию на май 2023 г.).

4. Анализ метаданных и текстов

Рассмотрим распределение текстов электронной базы по фандомам. Приблизительно 28,5 % выборки составляют ориджиналы. Ориджиналы можно оценивать и как использование начинающими авторами ресурсов фанфикшн для продвижения, и как выход за рамки заимствования героев и элементов сюжета из других историй [16]. Тем не менее, учитывая сравнительно большое количество оригинальных работ на фанфикшн-платформе, можно говорить о востребованности фан-сообществ в т. ч. как площадки для публикации авторских текстов, не являющихся порождением фан-культуры в полной мере.

В список из 25 наиболее востребованных фандомов входят 11 фандомов, относящихся к категории «Аниме и манга» (*Katekyo Hitman Reborn!*, *Naruto*, *Fairy Tail*, *Bleach* и др.), а также три фандома корейских музыкальных групп (*Super Junior*, *SHINee*, *EXO – K/M*), что говорит о ярко выраженном интересе фикрайтеров к азиатскому искусству и культуре. В категории «Фильмы и сериалы» самыми многочисленными оказались фандомы «Шерлок» (*BBC*), «Сверхъестественное», «Закрытая школа», «Дневники вампира». Любопытно, что фандом «Гарри Поттер» в рассмат-

¹⁶ Объём электронной базы ограничен из-за лимита ресурса на частоту поступающих запросов при веб-скрейпинге.

риваемой выборке не так популярен в сравнении с текущей информацией, размещённой на «Книге фанфиков»: среди наиболее востребованных фандомов электронной базы он занимает только восьмое место, насчитывая 2 596 фан-текстов (всего около 2 % выборки).

Отметим, что количество фандомов на ресурсе существенно возросло за последние десять лет. Несмотря на то, что многие новые фан-сообщества насчитывают большее количество текстов, некоторые фандомы, появившиеся на рубеже 2000–10-х гг., остаются востребованными и пополняются публикациями по сей день. Среди них можно выделить такие фандомы, как «Гарри Поттер», *Naruto*, *Fairy Tail*, «Мстители», «Сверхъестественное», «Шерлок» (*BBC*) и др.

Направленность в фанфикшн характеризует наличие и особенности любовной линии, описанной в тексте. Так, «джен» предполагает отсутствие или малозначительность романтических отношений в работе. «Гет» маркирует описание гетеросексуальных отношений, «слэш (фемслэш)» – гомосексуальных. Смешанная направленность указывает на наличие нескольких равнозначных романтических линий разных направленностей. «Книга фанфиков» также вводит направленность «статья» – «публицистический текст о фандоме или писательском искусстве» [15].

Рассмотрим распределение текстов выборки по направленности. Несмотря на то, что более 42 % работ относятся к направленности джен, большая часть произведений сосредоточена на описании романтических (и/или сексуальных) отношений – они составляют ~57 %. Статистические данные подтверждают и суждение о том, что фан-тексты чаще описывают гомосексуальные отношения, чем гетеросексуальные. В то же время отношения между мужчинами (слэш) привлекают фикрайтеров в десятки раз больше, чем отношения между женщинами (фемслэш). Полученные данные свидетельствуют о том, что слэш действительно превалирует среди фан-текстов, содержащих любовную линию, что говорит в пользу исследовательских гипотез о привлекательности слэша как альтернативы общепринятым представлениям о гендерных ролях [3; 20; 21]. Тем не менее выборку можно назвать довольно сбалансированной по соотношению трёх основных направленностей – джена, слэша и гета. Следовательно, можно предположить, что распределение текстов по направленности варьируется внутри различных фан-сообществ.

Так, в фандоме «Шерлок» (*BBC*) работы направленности «слэш» составляют ~49,5 % от всех текстов, «джен» – ~35,7 %, «гет» – ~10,5 %, что объясняется фанатским восприятием и интерпретацией дружеских отношений главных героев оригинала Шерлока Холмса и Джона Ватсона в качестве гомосексуальных (т. н. «синдром Старски и Хатча» [3]). В рамках фандома «Гарри Поттер» можно увидеть иное соотношение направленностей: около 40 % текстов относятся к гету, ~35 % – к джену и только ~26 % – к слэшу. В качестве причин такого распределения можно назвать наличие как мужских, так и женских канонических персонажей первого и второго плана, а также описание гетеросексуальных отношений в ориги-

нальных произведениях, играющее значимую роль в развитии сюжета и героев.

Вопреки стереотипному представлению о фанфикшн как об эротических произведениях [22], более чем $\frac{3}{4}$ текстов не содержат какого-либо описания сексуальных взаимодействий, а самым многочисленным рейтингом является *G* (~45 % работ), зачастую не предполагающий наличие любовной линии в истории. Кроме того, детальное графическое описание эротики или насилия содержат лишь около 12 % фанфиков, где на особо жестокие работы (*NC-21*) приходится менее 1 %. Такое распределение может говорить о том, что интерес фан-авторов более сфокусирован на интерпретации событий и личности героев канона, а также на эмоциональном аспекте взаимоотношений персонажей, чем на эротических взаимодействиях между ними.

Обратимся к анализу корреляции рейтинга и направленности фанфикшн-произведений. Так, более чем $\frac{2}{3}$ текстов рейтинга *G* относятся к джену, в то время как слэш преобладает для высокорейтинговых фанфиков (*R*, *NC-17*, *NC-21*). Работы о взаимоотношениях разнополых персонажей занимают от 16 до 27 % для каждого из рейтингов. Более того, статистические данные подтверждают гипотезу, выдвинутую М. МакКардл. Именно пейринг задаёт направленность работы и во многом влияет на возрастные ограничения и рейтинг: джен и гет противопоставлены слэшу (фемслэшу), т. к. гетеросексуальные отношения (в отличие от гомосексуальных) считаются приемлемыми в обществе и, следовательно, допустимыми для ознакомления читателями любого возраста [23]. Действительно, для общедоступных текстов рейтингов *G* и *PG-13* количество текстов направленностей «гет» и «джен» значительно превышает количество текстов направленностей «слэш» и «фемслэш» (~84,8 против ~12,4 % и ~56 против ~39,7 % соответственно). Обратная тенденция, также предсказанная МакКардл, прослеживается для рейтингов *NC-17*, *NC-21*: число слэш- и фемслэш-произведений превосходит число джен- и гет-произведений на ~48,6 и ~21,2 % соответственно. Для рейтинга *R* распределение сравнительно равномерное: на гет и джен приходится около 45 %, на слэш и фемслэш – около 50 %.

Перейдём к анализу категории меток, присвоенных работам авторами на основе жанровых особенностей, содержания, настроения и модели построения текста. Важно упомянуть, что система «жанров» фанфикшн была сформирована самими фикрайтерами и «ориентирована на эмоциональную, а не литературную составляющую текстов», поэтому одна работа может сочетать в себе несколько жанров [17]. Уникальные жанровые обозначения для фан-текстов были созданы во многом из-за того, что общепринятые литературные формы (рассказ, повесть, роман, поэма) не способны отразить присущие фанфикам особенности [18]. В то же время некоторые жанры фанфикшн опираются на традиционные литературные жанры и переосмысливают их – например, романтика (романс), драма, юмор (комедия) [17].

Около одной трети текстов имеют жанровую метку «романтика», что соотносится с данными о том, что более половины произведений описывают романтические отношения между персонажами. Кроме того, востребованность романтической линии свидетельствует о «стремлении фикрайтеров создать для любимых героев идеальные взаимоотношения» [10]. Любопытно, что «ангст» встречается почти так же часто, при этом отражая тревожное, депрессивное настроение историй, которое можно назвать противоположным позитивно окрашенному жанру «романтика». С другой точки зрения, «ангст» может трактоваться как эмоциональная составляющая описания любовной линии и сочетаться с тегом «романтика» – в электронной базе обе метки имеют 4 146 работ. Ещё одна интерпретация популярности жанра «ангст» подразумевает «сильное влияние канона» на фан-тексты, если в оригинальном произведении эмоциональным переживаниям героев отводилось значимое место [10]. Приведённые объяснения релевантны и для обоснования частотности жанра «драма», во многом схожего с жанром «ангст».

Наличие меток «смерть основных персонажей», «дарк», «насилие» может указывать на то, что авторы ориентируются на канон, включающий описание жестоких и трагичных моментов. Отмеченный интерес фикрайтеров к описанию «тяжёлых» событий характерен для фанфикшн: читатели фан-текстов парадоксально стремятся проживать страдания вместе с персонажами [21]. Дело в том, что одной из функций чтения фанфикшн является шоковый эффект, зачастую связанный с высоким уровнем «телесного реализма» в повествовании, который обозначают перечисленные метки, а также теги «нецензурная лексика», *PWP*, «изнасилование» [21].

Популярность тегов *OOC* и *AU*, подразумевающих значимое изменение образов и характера канонических персонажей и перенос героев в отличные от оригинала обстоятельства, говорит о стремлении авторов фанфикшн создавать новые интерпретации личности любимых персонажей и событий их жизни, соответствующие индивидуальному фанатскому представлению о них. Более того, по частотности тега «повествование от первого лица» можно судить об отождествлении фикрайтера с главным героем фан-текста, которым может быть как персонаж канона, так и введённый автором (ОЖП, ОМП). Нередко таким образом осуществляется желание фикрайтера представить себя участником повествования.

В отношении организации текста самая используемая форма – стихотворная. Причём почти 60 % (10 815 из 18 284) текстов с этой меткой относятся к категории ориджиналов, что подтверждает функциональность фанфикшн-ресурсов как инструмента для продвижения своего творчества, в т. ч. не имеющего отношения к фан-культуре. Другой популярный «жанр», связанный с форматом построения текста, – «сонгфик», который привлекает фикрайтеров, вероятно, возможностью «создать особую атмосферу или подчеркнуть то, что находится между строк» в тексте с помощью строк из музыкального произведения [18].

Все 136 257 фанфикшн-текстов, включённых в электронную базу, были токенизированы и лемматизированы с применением библиотеки *NLTK* и морфологического анализатора *rumorphy2*, реализованных на языке *Python*.

Суммарный объём выборки составил 505 391 801 токенов. Средний объём текста – 3 709 токенов. Более 57 % электронной базы занимают тексты длиной менее 1 тыс. токенов, около $\frac{3}{4}$ текстов короче 2 тыс. токенов. В то время как на тексты объёмом более 10 тыс. токенов приходится всего около 8 % выборки. К тому же 55 080 работ (~40 %) отмечены тегом «драбл», характеризующим короткий отрывок, часто детально описывающий одну идею, сцену или персонажа.

Из приведённых данных можно сделать вывод о том, что большинство фан-текстов сосредоточены на описании одной или нескольких сцен и близки к малым литературным формам – рассказу и повести.

Важно отметить, что произведённая обработка позволит использовать электронную базу для лингвистического анализа и выявления языковых особенностей текстов, например, в рамках одного или нескольких фандомов или для сравнения фан-работ в диахронической перспективе.

Заключение

На данный момент фанфикшн занимает важное место в ряду современных сетевых литературных явлений и привлекает всё больший интерес не только авторов и читателей, но и исследователей разных научных областей. Фан-тексты оказываются успешными по причине «многофункциональности, интенсивности и свободы» писательского и читательского опыта [21]. Кроме того, фанфикшн был и остается как одним из способов взаимодействия потребителей культуры (фанатов) с медиаконтентом, так и порождением культуры конвергенции [24].

Результатом настоящей работы стало создание русскоязычной электронной базы фанфикшн, включающей в себя более 135 тыс. текстов и данные о них. На основе квантитативного анализа полученной информации были сделаны выводы о масштабном развитии фанфикшн в русскоязычной среде за последние десять лет за счёт появления новых фандомов и расширения существующих. Статистический анализ подтверждает значимость романтической составляющей фан-текстов и преобладающую ориентированность на гомосексуальные отношения, что в то же время во многом зависит от специфики оригинального произведения. Для многих фанфиков характерна повышенная степень «телесного реализма», заставляющая читателей проживать эффект шока, который является неотъемлемой частью опыта прочтения. Тем не менее большая часть работ не содержит описания эротических сцен, тем самым опровергается стереотипное представление о фанфикшн как о преимущественно порнографических текстах. Значимым аспектом взаимодействия с фан-текстом для фикрайтера и читателя оказывается соотнесение себя с любимым персонажем или воплощение собственного образа в рамках индивидуальной интерпретации канона.

Перспективами настоящего исследования являются расширение электронной базы, применение её материалов для создания моделей машинного обучения (для решения задачи классификации текстов по тематическим тегам), сопоставления фан-текстов с другими видами (жанрами) литературы, изучения отдельных явлений в рамках феномена фанфикшн.

В 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 г.

Список источников

1. Leavenworth M. L. A Truth Universally Acknowledged? Pride and Prejudice and Mind-reading Fans / M. L. Leavenworth // *Storyworlds: a journal of narrative studies*. 2015. Vol. 7. No. 2. Pp. 93–110.
2. Jenkins H. *Textual Poachers: Television Fans and Participatory Culture* / H. Jenkins. New York: Routledge, 1992.
3. Bacon-Smith C. *Enterprising Women: Television Fandom and the Creation of Popular Myth* / C. Bacon-Smith. Philadelphia, 1992.
4. Hellekson K. Work in Progress / K. Hellekson, K. Busse // *Fan Fiction and Fan Communities in the Age of the Internet*. Jefferson: McFarland Press, 2006. Pp. 5–32.
5. Coppa F. A Brief History of Media Fandom / F. Coppa // *Fan Fiction and Fan Communities in the Age of the Internet*. Jefferson: McFarland Press, 2006. Pp. 41–60.
6. Прасолова К. А. Фанфикшн: литературный феномен конца XX – начала XXI в. (творчество поклонников Дж. К. Роулинг): дисс. канд. филол. наук / К. А. Прасолова. Калининград, 2009.
7. Четина Е. М. Фандомы и фанфики: креативные практики на виртуальных платформах / Е. М. Четина, Е. А. Ключикова // *Вестник ПГНИУ*. 2015. № 3 (31). С. 95–104.
8. Coppa F. *The Fanfiction Reader: Folk Tales for the Digital Age* / F. Coppa. University of Michigan Press, 2017.
9. Попова С. Н. Лингвостилистика фанфикшн (на материале англоязычных сайтов, посвящённых творчеству Дж. Р. Р. Толкина): дисс. канд. филол. наук / С. Н. Попова. М., 2009.
10. Коробко М. А. Жанр в фанфикшн: закономерности использования (на материалах фандомов «Шерлок», «Мерлин», «Сверхъестественное») / М. А. Коробко // *Учёные записки ОГУ им. И. С. Тургенева*. 2015. № 6 (69). С. 154–157.
11. Sindoni M. G. I Really Have no Idea what Non-fandom People Do with their Lives. A multimodal and Corpus-based Analysis of Fanfiction / M. G. Sindoni // *Linguae e Linguaggi*. 2015. Vol. 13. Pp. 277–300.
12. Milli S. Beyond Canonical Texts: a Computational Analysis of Fanfiction / S. Milli, D. Bammann // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Pp. 2048–2053.
13. Evans S. A. Where no one Has Gone Before: a Meta-Dataset of the World's Largest Fanfiction Repository / S. A. Evans et al. // *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Pp. 6106–6110.
14. Колпакова Т. А. Эволюция фанфикшн с 2005 по 2020 гг. на примере фандома «Гарри Поттер» / Т. А. Колпакова // *Социальные и гуманитарные науки: теория и практика*. 2020. № 1 (4). С. 85–101.
15. Книга фанфиков. URL: ficbook.net.

16. Жилыева А. И. Ориджинал: между фанфикшн и литературой / А. И. Жилыева, А. С. Лазарева, Н. В. Кургузова // Перспективы науки – 59: матер. VI Междунар. конкурса НИР (Казань, 2017). Казань: Рокета Союз, 2017. С. 186–191.
17. Антипина Ю. В. Жанровые особенности фанатской прозы (на примере фанфикшн по творчеству братьев Стругацких) / Ю. В. Антипина // Вестник ЧелГУ. 2011. № 13. С. 21–25.
18. Шатрова Е. Д. К проблеме определения жанров произведений фанфикшн / Е. Д. Шатрова, Л. А. Ласица // Вестник ОГУ им. И. С. Тургенева. 2017. № 1 (201). С. 41–48.
19. Mitchell R. Web Scraping with Python / R. Mitchell. Sebastopol: O'Reilly Media, Inc., 2018.
20. Russ J. Pornography by Women, for Women, with Love / J. Russ // Magic Mommas, Trembling Sisters, Puritans and Perverts: Feminist Essays. Trumansburg, 1985. Pp. 79–99.
21. Самутина Н. Великие читательницы: фанфикшн как форма литературного опыта / Н. Самутина // Социологическое обозрение. 2013. Т. 12. № 3. С. 137–191.
22. Caputo K. The New Pornography: the Rise of Fanfic / K. Caputo. 2016.
23. McCardle M. Fan Fiction, Fandom, and Fanfare: What's All the Fuss / M. McCardle. Boston, 2003.
24. Jenkins H. Convergence Culture: Where Old and New Media Collide / H. Jenkins. New York: New York University Press, 2006.

М. С. Панкова,

магистрант

Южный федеральный университет, Ростов-на-Дону, Россия

e-mail: mapankova@sfedu.ru

КОРПУС ДОНСКИХ ГОВОРОВ КАК OPPORTUNISTIC CORPUS

Аннотация. В статье рассматривается диалектный Корпус донских говоров, относящийся к типу оппортунистических корпусов. Оппортунистические корпусы – это специализированные лингвистические корпусы, которые не обязаны быть сбалансированными и/или репрезентативными. При помощи построения корреляционной матрицы, анализа тональности текстов и автоматического морфологического анализа диалектных слов исследуются характеристики Корпуса донских говоров, подтверждающие его принадлежность к оппортунистическим корпусам.

Ключевые слова: корпусная лингвистика, *opportunistic corpus*, *оппортунистический корпус*, диалектный корпус, корреляционная матрица

В научной литературе представлено несколько определений понятия «лингвистический корпус», которые объединяются общими характеристиками. Лингвистический корпус представляет собой собрание языковых данных в электронном виде, снабжённых лингвистической разметкой и собранных в соответствии с определёнными критериями [3; 4; 8]. Корпус предназначен для решения различных задач, связанных с изучением языка. Существует два подхода к лингвистическому анализу на основе языковых корпусов: *corpus-based* и *corpus-driven* [4; 11].

Подход *corpus-based* – это анализ, основанный на корпусе, т. е. у исследователя есть заранее сформулированные гипотезы и корпус служит источником данных для этих гипотез и источником примеров.

В подходе *corpus-driven*, т. е. в исследованиях, направляемых корпусом, происходит наоборот: корпусные данные позволяют сформулировать лингвистические гипотезы. Исследователь заранее не может предсказать, что дадут ему данные, которые представлены в языковом корпусе.

В подходе *corpus-driven* становится важным принцип сбора данных для того, чтобы делать лингвистические обобщения на их основе. Согласно этому принципу, можно выделить мониторные, или мониторинговые, корпусы: они сбалансированные и динамические, т. е. пополняющиеся со временем. Также есть сбалансированные корпусы на основе репрезентативной выборки (*sample, snapshot*). Если же корпус нельзя отнести ни к мониторинговым, ни к собранным на основе репрезентативной выборки, его называют «оппортунистическим» [10].

«Оппортунистический», или, как он назван в статье Е. А. Нахимовой, «приспособленный» [5] корпус содержит все данные, которые удалось со-

брать для решения конкретной задачи. Такой корпус несбалансированный и/или нерепрезентативный, но при этом служит определённой цели. Например, в [5] был собран корпус для изучения прецедентных имён на базе интернет-архивов российских СМИ. Мониторинговые корпуса также можно считать оппортунистическими, потому что они чаще всего не являются репрезентативными [11]. Некоторые исследователи полагают, что любой корпус можно назвать оппортунистическим, поскольку все корпуса в той или иной степени несбалансированны [9]. Диалектные корпуса в силу своей специфики не являются репрезентативными [2].

Корпус донских говоров создан в рамках проекта «Конвергенция языковых пластов русского языка в зеркале цифровых решений», соглашение о научном сотрудничестве («Зеркальные лаборатории») между НИУ ВШЭ и ЮФУ. Задачей корпуса является отражение диалектных особенностей нижнедонских говоров, входящих в группу южнорусских говоров. Данные для корпуса были представлены в виде аудиозаписей и текстовых расшифровок к ним [6]. Расшифровка осуществлялась с использованием программы *ELAN* [8].

Текстовые данные были подготовлены таким образом, что для анализа использовались только реплики информантов, речь интервьюеров была исключена. Тематика текстов отражена в ответах на вопросы интервьюеров: т. к. интервью были собраны в процессе диалектологических экспедиций, интервьюеров интересовали события из жизни информантов, а также особенности обычаев и обрядов. Соответственно, в текстах представлена информация о детстве, играх, школьном периоде, служении в армии, болезнях и народных способах их лечения, сватовстве и свадебных обрядах, церковных обрядах и праздниках. Кроме того, информанты могли рассказывать о том, как они жили во время Великой Отечественной войны.

В исследовании текстов Корпуса донских говоров были взяты 11 записей из трёх населённых пунктов, среди которых есть несколько записей от одного и того же информанта – они представляют собой расшифрованные фрагменты целостного интервью. Это тексты 07, 09 и 11.

Была построена корреляционная матрица (рис. 1), показывающая формальную связанность текстов по содержащейся в них лексике, и было выявлено, что в целом тексты сильно коррелируют друг с другом. Коэффициент корреляции составляет от 0,5 до 0,8. Исключение составляют три текста под номерами 07, 09 и 11 – это фрагменты большой беседы, а не отдельные тексты: их коэффициенты корреляции меняются от 0,1 до 0,3.

Построенная корреляционная матрица показывает достаточно сильную лексическую схожесть текстов, т. е. корпус является несбалансированным по использованной лексике и жанру.

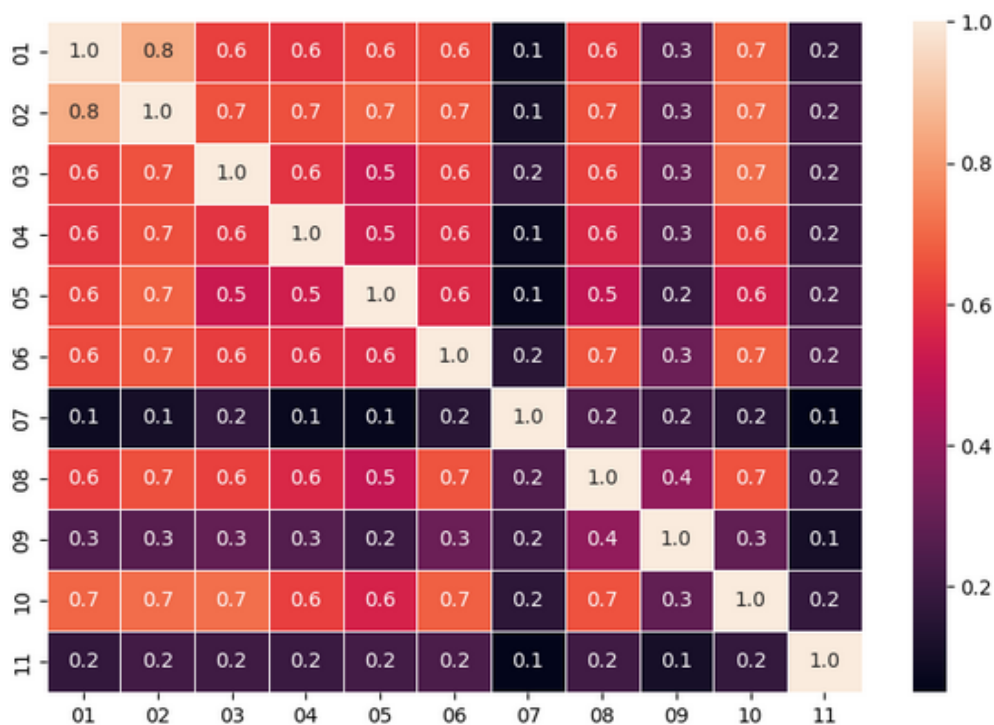


Рис. 1. Корреляционная матрица

Нами также был проведён анализ тональности (сентимент-анализ) текстов корпуса. Исследование проводилось на основе машинного обучения – для этого использовалась модель, обученная на текстах из русскоязычной социальной сети. Модель на основе наивного Байесовского классификатора (*Naïve Bayes Classifier, NBC*) базируется на концепции условной вероятности того, принадлежит ли документ d классу c . Для определения наиболее вероятного класса для документа d необходимо вычислить условные вероятности принадлежности документа d к каждому из представленных классов по отдельности и выбрать класс с наибольшей вероятностью [7].

В результате все тексты корпуса были разделены на два класса: негативно и позитивно окрашенные. При этом для негативно окрашенных текстов характерны такие слова, как «болеть», «пенсия», «ругаться», «убивать», «погибать», «похоронить», «больница», «плакать», «зараза», «мамка» и др. В «позитивном» классе были использованы слова «жениться», «молитва», «икона», «добро», «молодец», «сватать», «замуж», «добрый», «церковный», «святой», «немец» и др. Можно заметить, что слово «немец» попало в список лексики «положительных» текстов. У информантов немцы связаны с военным временем, что скорее сигнализирует о сложном периоде их жизни, в то время как модель была построена на основе современных текстов из социальной сети. По-видимому, для современных носителей русского языка у слова «немец» уже исчезла отрицательная коннотация, поэтому таким образом обученная модель отнесла диалектные тексты, в которых присутствует слово «немец», к позитивно окрашенным.

Кроме того, было проведено исследование качества обработки диалектной лексики, которая встречается в речи донского казачества, ав-

томатическими морфологическими анализаторами *pymorphy2* и *MyStem*. Оказалось, в частности, что существительное «дружко», обозначающее распорядителя на свадьбе со стороны жениха [1, с. 141], этими программами приводится к наречию в качестве начальной формы. А существительное мужского рода «котух», которое означает «хлев для мелкого рогатого скота и свиней» [1, с. 235], во фразе «Он и в сараях, и в котухах, и всё», согласно морфологическим анализаторам, неверно определяется как одушевлённое существительное женского рода.

В результате проведённого исследования лексической близости текстов на основе корреляционной матрицы, анализа тональности текстов и анализа автоматической морфологической обработки диалектной лексики было выявлено, что Корпус донских говоров является несбалансированным. Корпус однороден по своему лексическому и жанровому составу: это набор интервью с типовыми вопросами, а не спонтанная речь. В корпусе представлен ограниченный набор тематик. Корпус донских говоров не является репрезентативным, т. к. в нём представлены тексты из ограниченного числа донских районов. И, наконец, он специфичен с точки зрения своего типа – он звуковой, с точки зрения своего устройства – это интервью, а также с точки зрения встречающейся в нём диалектной лексики.

Таким образом, корпус донских говоров может быть определён как «оппортунистический корпус».

Список источников

1. Большой толковый словарь донского казачества: около 18 000 слов и устойчивых словосочетаний / РГУ. М.: Русские словари: Астрель: АСТ, 2013. 608 с.
2. Гольдин В. Е. Корпус русской диалектной речи: концепция и параметры оценки / В. Е. Гольдин, О. Ю. Крючкова // Компьютерная лингвистика и интеллектуальные технологии: Диалог-2011: матер. Междунар. конф. (М., 2011). С. 359–367.
3. Захаров В. П. Корпусная лингвистика / В. П. Захаров, С. Ю. Богданова. СПб.: СПбГУ. РИО. Филологический факультет, 2013. 148 с.
4. Копотев М. Введение в корпусную лингвистику / М. Копотев. Прага: Animedia Company, 2014.
5. Нахимова Е. А. Использование корпусной методологии при сопоставительном изучении прецедентных имён / Е. А. Нахимова // Политическая лингвистика. 2013. № 3 (45). С. 48–56.
6. Флягина М. В. Корпус донских диалектов / М. В. Флягина, Н. В. Калиничева, Е. М. Северина. М.: Международная лаборатория языковой конвергенции, НИУ ВШЭ, 2022–23. URL: lingconlab.ru/don_rnd.
7. Domingos P. On the Optimality of the Simple Bayesian Classifier under Zero-one Loss / P. Domingos, M. Pazzani // Machine Learning. 1997. No. 29. Pp. 103–137.
8. ELAN: Computer Software. Ver. 6.3. Nijmegen: Max Planck Institute for Psycholinguistics. The Language Archive, 2022. URL: archive.mpi.nl/tla/elan.
9. Halliday M. A. K. Lexicology and Corpus Linguistics / M. A. K. Halliday, A. Cermakova, W. Teubert et al. A&C Black, 2004. 256 p.
10. McEnery T. Corpus Linguistics. Method, Theory, and Practice / T. McEnery, A. Hardie. Cambridge: Cambridge University Press, 2011. 294 p.
11. Varadi T. The Linguistic Relevance of Corpus Linguistics / T. Varadi // Proceedings of the Corpus Linguistics Conference (2001). Pp. 587–593.

В. А. Белковский, И. П. Максаков
Сибирский федеральный университет, Красноярск, Россия

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ КОМПЬЮТЕРИЗИРОВАННОГО КОНТЕНТ-АНАЛИЗА КАК МЕТОДА ГУМАНИТАРНОГО ИССЛЕДОВАНИЯ В СОВРЕМЕННОЙ ПРАКТИКЕ

Аннотация. Контент-анализ является признанным и эффективным методом исследования в социальных науках. Огромную пользу ему принесли ИКТ и рост вычислительных мощностей, а с появлением интернета, особенно электронной почты и веб-сайтов, контент-анализ стал повсеместным. Поиск по ключевым словам, являющийся основным элементом контент-анализа, наиболее широко распространен во многих интернет-приложениях – от поисковых систем до программ для взлома паролей. В этой статье рассматриваются некоторые из наиболее важных применений контент-анализа и обсуждается его всё более важное место в гуманитарных областях.

Ключевые слова: *цифровые инструменты, дистанционное преподавание, контент-анализ*

Образовательное взаимодействие между учащимися и учителями в виртуальной среде требует большей гибкости и адаптируемости, чем в традиционном классе [4]. В учебный процесс были внесены уникальные изменения: синхронное обучение с помощью инструментов видеоконференций, асинхронное обучение с помощью заранее записанных учебных пособий и создание хорошо организованного цифрового класса. Учителя могут персонализировать платформы онлайн-обучения и системы управления, такие как *Google Classroom* или *Class Notebook*, с учётом определённых уровней обучения, предмета, стандартов учебной программы и предпочтений преподавания [7]. Целью данной статьи является рассмотрение повсеместного распространения и разветвлений контент-анализа, многочисленных технологий, а также многочисленных функций контент-анализа в гуманитарной среде.

С развитием цифрового программного обеспечения в целом и интернета в частности огромные и постоянно растущие объёмы текстовой информации стали доступны во всём мире. Поисковые системы интернета и встроенные поисковые функции браузеров и других типов программного обеспечения стали широко распространёнными, легкодоступными и удобными для пользователя. Такие инструменты для контент-анализа постоянно развиваются и становятся всё более совершенными, оставаясь при этом очень удобными для пользователя. Более того, анализ цифрового контента не исключает другие методологии, такие как анализ дискурса, нарратив-

ный анализ или анализ аргументации. Подобные качественные методы, на наш взгляд, полностью совместимы с контент-анализом.

Уникальность контент-анализа заключается в том, что он стал легкодоступным, удобным для пользователя и почти неизбежным методом в цифровом мире. Мы утверждаем, что – за исключением более специализированных технологий и явного применения научного контент-анализа – студенты и сотрудники учебных учреждений уделяют мало внимания вопросам использования таких инструментов и последствий их использования.

Контент-анализ возможно использовать для различных целей в научных работах: для отражения культурных моделей групп и институтов, компаний или для выявления направленности определённых субъектов и т. п. Так, полезным разграничением единиц анализа является выборка, контекстные единицы [6; 8]. В первом случае мы фокусируемся на всём сообщении – например, на газетной статье или веб-странице. Единицы записи – это различные отдельные компоненты, на которые можно разделить сообщение, – например, абзацы, предложения или отдельные слова. В последнем случае мы рассматриваем весь контекст, в котором находится сообщение – например, всю газету или веб-сайт. В качестве общего примера общий «трёхуровневый» подход к контент-анализу может включать частоту категорий, ключевые слова в контексте и список частотности слов.

Первый уровень (частота категорий) анализирует, сколько раз в каждой речи встречаются категории лексики, созданной по основным фреймам. Фреймы – это общие коды, посредством которых акторы интерпретируют сложные проблемы. Эти схемы интерпретации упрощают внешнюю реальность посредством избирательного процесса [2]. Различные фреймы представляют собой альтернативные способы решения темы исследования, которая может быть практически чем угодно, от «неравенства доходов» до «прав человека» и «кибертерроризма» или «жизни во Вселенной».

Второй уровень анализа (*KWIC*, *keyword-in-context* или ключевые слова в контексте) иллюстрирует извлечённый фрагмент текста, в который термин вставлен длиной в три строки, что позволяет лучше понять его значение. Фактически, слово, изолированное от контекста, может вызвать некоторые ошибки. Например, одинаковые слова могут иметь разное значение («оставить кого-нибудь в покое» и «покойся с миром») или некоторые термины могут быть просто отрицаниями. Таким образом, интерпретация и отбор являются важными задачами для обеспечения эффективности инструментов контент-анализа.

Третий уровень анализа – словарь: частота слов, входящих в предварительный словарь, используемых в тексте.

В последние годы социологи существенно усовершенствовали инструменты контент-анализа, особенно благодаря появлению персональных компьютеров и соответствующего специализированного программного обеспечения. Программное обеспечение, созданное специально для контент-анализа, несомненно, облегчило задачу исследователя в чрезвычайно

трудоёмких действиях, таких как измерение частот и поиск битов информации, скрытых за значительным количеством «белого шума». Частая критика этих методологических инструментов указывает на отсутствие или недооценку «общего контекста» (контекстных единиц), в котором могут происходить вербальные взаимодействия.

Программы для контент-анализа, такие как *Hamlet II* или *Words in Context*, позволяют исследователю быстро оценить частотность терминов и их расположение в тексте. *Hamlet II* способна продемонстрировать такие черты, как простота, непосредственность и наличие необходимых требований для разных уровней анализа, что хорошо соответствует целям большинства исследований в области гуманитарных, социальных наук.

Программы, подобные упомянутым здесь и другим, ищут в текстовых файлах слова или категории в заданном словарном списке, подсчитывают их совместные частоты в любой заданной единице контекста, в предложениях или в виде словосочетаний внутри заданного диапазона слов, а затем предоставляют соответствующую информацию. Этот процесс «сокращения» и классификации называется «маркировкой» [9]. Для операционализации абстрактных понятий, таких как «национальные интересы», «социальное неравенство» или «неприятие риска», необходимо построить категорию, которая методологически сложна, поскольку требует уловить суть социального явления [3].

Поскольку большая часть информации, доступной в интернете, всё ещё основана на текстовых материалах, можно с уверенностью сказать, что алгоритмы поисковых систем представляют собой высокоразвитый тип анализа контента. Такие алгоритмы могут просматривать огромные массивы текста, создавать таблицы частот и извлекать информацию, что делает их незаменимым помощником в процессе принятия решений и операционных исследований [5].

Однако контент-анализ имеет множество других функций, которые также могут быть использованы в различных сферах. Список приведённых ниже функций не является исчерпывающим, но он демонстрирует широкий спектр возможностей, которые открывает компьютерный анализ контента: кибератаки (рассылка спама, кражи личных данных, взломы компьютеров); разведка и наблюдение; расследование преступлений; исследования рынка; академические исследования; журналистские расследования; личные цели (хобби, планирование путешествий и т. д.).

Важно отметить, что контент-анализ может быть полезен в различных областях и использован для достижения разных целей. Например, в области маркетинга он может помочь в изучении предпочтений и поведения потребителей, а в журналистике – в раскрытии скрытых фактов и коррупции. Таким образом, контент-анализ является мощным инструментом, который может быть использован для достижения разных целей и решения различных задач.

Контент-анализ, который когда-то был особым количественным методом текстового анализа среди многих других, стал доминирующим

и даже повсеместным способом получения информации в эпоху цифровых технологий, поскольку поиск по ключевым словам, ключевой элемент анализа контента, стал наиболее распространённой функцией многих интернет-приложений, от поисковых систем до программ для взлома паролей. Чтобы увидеть это и понять последствия, понятие контент-анализа необходимо расширить за пределы ограниченной и требовательной концепции научного контент-анализа. Сегодня существует множество удобных инструментов для анализа контента, некоторые простые поисковые системы (например, инструмент поиска в *MS Word* и других программах), а также более продвинутый поиск движки, такие как *Google*, и более специализированные инструменты, такие как программное обеспечение *Hamlet*. Контент-анализ также широко применим и используется во многих различных контекстах и для самых разных целей.

Мы считаем, что компьютеризированный онлайн-контент-анализ обладает двумя главными преимуществами. Во-первых, он легкодоступен, удобен для пользователя, даёт немедленные результаты, может применяться к огромным объёмам материала и имеет глобальный охват. Например, при использовании сканера сайтов компьютеризированный контент-анализ может автоматически сканировать веб-страницы и получать информацию о большом количестве сайтов в короткие сроки. Во-вторых, компьютеризированный контент-анализ позволяет избежать человеческих ошибок при сборе и кодировании данных. Он может быть использован для анализа большого количества материала и обработки данных с большой скоростью.

Однако следует учитывать, что результаты компьютеризированного контент-анализа зависят от выбора ключевых слов и интерпретации результатов. Это может привести к некоторым ограничениям в использовании методов контент-анализа. В отличие от человеческого анализа, где исследователь может установить свои собственные критерии для анализа данных, компьютеризированный контент-анализ полностью зависит от того, какие ключевые слова используются и как интерпретируются результаты. Выбор ключевых слов и интерпретация результатов – это не то, в чём сам контент-анализ может помочь.

Список источников

1. Arroyo M. Content Analysis of Digital Tools Use during 2020–21 Remote Teaching / M. Arroyo, L. Quinn, L. Paretti et al. // *Educational Research: Theory and Practice*. 2022. No. 33 (1). Pp. 55–60.
2. Benford R. D. Framing Process and Social Movement: an Overview and Assessment / R. D. Benford, D. A. Snow // *Annual Review of Sociology*. 1992. No. 26. Pp. 611–639.
3. Druckman D. *Doing Research: Methods of Inquiry for Conflict Analysis* / D. Druckman. London: Sage, 2005. 412 p.
4. Henriksen D. Folk Pedagogies for Teacher Transitions: Approaches to Synchronous Online Learning in the Wake of COVID-19 / D. Henriksen, E. Creely, M. Henderson // *Journal of Technology and Teacher Education*. 2020. No. 28 (2). Pp. 201–209.

5. Janasik N. Text Mining in Qualitative Research Application of an Unsupervised Learning Method / N. Janasik, T. Honkela, H. Brunn // *Organizational Research Methods*. 2009. No. 12 (3). Pp. 436–460.
6. Krippendorff K. *Content Analysis: an Introduction to its Methodology* / K. Krippendorff. London: Sage, 1980. 200 p.
7. Vermette L. Freedom to Personalize my Digital Classroom: Understanding Teachers' Practices and Motivations / L. Vermette, J. McGrenere, C. Birge et al. // *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2019. No. 11 (2). Pp. 1–14.
8. Weare C. Content Analysis of the World Wide Web: Opportunities and Challenges / C. Weare, W. Y. Lin // *Social Science Computer Review*. 2000. No. 18 (3). Pp. 272–292.
9. Weber R. P. *Basic Content Analysis* / R. P. Weber. London: Sage, 1990. 102 p.

А. И. Артемьева,
преподаватель
Сибирский федеральный университет, Красноярск, Россия
e-mail: artanastasiia@mail.ru
ORCID: 0009-0008-2277-1953

ТЕХНОЛОГИИ ОЦЕНКИ ДОСТУПНОСТИ ТЕКСТОВОЙ ИНФОРМАЦИИ В ПРАКТИКЕ СОЗДАНИЯ БЕЗБАРЬЕРНОЙ КОММУНИКАЦИОННОЙ СРЕДЫ

Аннотация. В последние десятилетия активно развиваются цифровые методы в гуманитарных исследованиях, используются цифровые ресурсы, а также компьютерные технологии с целью решения как локальных, так и глобальных проблем. Однако явно прогрессирующее развитие компьютерных и иных систем для обработки данных не гарантирует эффективного и продуктивного использования технологий в рамках инклюзивной практики. В настоящем исследовании дана оценка технологий, анализирующих текстовый материал на предмет доступности для определённых групп населения в практике создания безбарьерной коммуникационной среды.

Ключевые слова: *простой язык, ясный язык, восприятие текста, читабельность текста, количественная оценка текста*

Ежегодно растёт количество людей, диагностируемых когнитивными и иным нарушениями, влияющими на качество жизни. Согласно данным Всемирной организации здравоохранения, в современном мире насчитывается около 55 млн людей с нейродегенеративными заболеваниями. К 2050 г. ожидается, что количество людей с данным диагнозом может увеличиться в три раза [13]. Помимо когнитивных расстройств, приводящих к сложностям в повседневной жизни, население планеты страдает и от других психических, неврологических, а также психоневрологических заболеваний, которые отягощают бытовое функционирование индивида.

Безбарьерная коммуникационная среда создаётся и поддерживается, однако не только для вышеперечисленных групп населения, но и для тех граждан, которые плохо владеют языком, не получили должного образования, позволяющего без каких-либо трудностей порождать и воспринимать тексты различных стилей, а также тех людей, которые в силу других причин имеют трудности с визуальным и смысловым восприятием письменных текстов.

В последние годы государством был принят ряд нормативных документов, призванный обеспечивать условия для людей с ограниченными возможностями здоровья. Одним из шагов для реализации документов являлось утверждение программы № 363 от 29.03.2019 «Доступная среда» [14].

В национальном стандарте РФ о «безбарьерной среде жизнедеятельности инвалидов» безбарьерная среда определяется следующим образом: «среда жизнедеятельности, в которой отсутствуют или сведены к минимуму физические, средовые, информационные и социально-психологические барьеры для инвалидов» [15].

Т. В. Жеребило даёт определение коммуникативной среде, важному научному концепту в рамках настоящего исследования. Под коммуникативной средой исследователь понимает «этносоциоязыковую общность» с регулярными «внутренними коммуникативными связями и определённой территориальной локализованностью» [4, с. 116].

Лица с когнитивными расстройствами, с особенностями интеллектуального развития, а также с миграционным прошлым и т. п. испытывают определённые трудности восприятия письменных текстов.

Одну из основных проблем составляют тексты институционального взаимодействия, поскольку они написаны крайне сложным канцелярским языком, содержат терминологическую лексику, оформлены в сложном для восприятия стиле. Несмотря на то, что такие тексты представляют собой объявления, инструкции и рекомендации и размещены в практически любом институте (медицинские организации, отделения услуг почтовой связи, пенсионный фонд и т. п.), уровень их доступности крайне мал.

Тексты таких жанров (инструкции, объявления, расписания, рекомендации и т. п.) должны быть понятны любому взрослому гражданину, поскольку «каждый человек должен иметь возможность доступа к информации, которая определяет сферу его социального взаимодействия, профессиональную или трудовую занятость, удовлетворение его потребностей» [2, с. 6].

В последние десятилетия возросла потребность в адаптации и переводе текстового материала для любой группы реципиентов. Адаптированные тексты, которые были трансформированы согласно правилам ясного и простого языка, призваны быть в разы легче для восприятия.

Существует ряд работ, в которых предлагается применять простой язык для адаптации текстов для определённых групп реципиентов. В англоязычной лингвистической практике для обозначения доступного, адаптированного текста используются дефиниции *Easy-to-Read*, *Easy Language*, в немецкоязычных источниках употребляются номинации *leichte Sprache*, *einfache Sprache* [7–9; 11; 12]. В публикациях отечественных авторов можно встретить обозначения «простой язык», «ясный язык» [3].

Под простым языком (*leichte Sprache*) в немецкой научной мысли принято понимать такой вариант языка, при котором построение предложения, а также его лексическая наполненность систематически редуцируются [7].

Важно понимать, что перед адаптацией текста на более доступный язык необходимо провести оценку на предмет его сложности или «читабельности». Данная работа практически невозможна без компьютерных и вычислительных систем, поскольку процесс оценки уровня читабельно-

сти текста включает в себя определённое количество этапов и задействует количественные параметры, среди которых: среднее количество слов в предложении, средняя длина слова, количество слов в тексте. На основании параметров вычисляются сложность текста, предполагаемый реципиент (при этом сервис на основании параметров указывает возможный возраст адресата, а также его степень образования).

Наиболее распространёнными формулами, в которых задействованы количественные параметры, являются формулы Флеша – Кинкейда, Дэйла – Чалл, *SMOG*. «Результатами применения формул являются показатели класса или возраста, на который рассчитан текст. На сегодняшний день известно более 200 формул для определения читабельности текста на многих языках мира» [5, с. 373].

К сервисам автоматической оценки читабельности текста относятся такие зарубежные веб-сайты, как *readable.com*, *readabilityformulas.com*, *datayze.com* и др. В русскоязычном пространстве данную услугу предоставляют сервисы *copywritely.com*, *plainrussian.ru*. Оценка читабельности на всех вышеперечисленных сайтах производится на основе нескольких общих подходов к автоматическому анализу текста.

Наиболее универсальную оценку текста в русскоязычном сегменте предоставляет сайт *plainrussian.ru*. Результаты проверки предстают в виде совокупности пяти подходов к автоматическому разбору текста на предмет его понятности и доступности. Среди подходов сервиса выделяется индекс Флеша – Кинкейда, рассматривающий текст по двум параметрам: средняя длина предложения (СДП) и среднее число слогов в слове (СЧС). По показателям СДП и СЧС даётся оценка по 100-балльной шкале: чем выше значение, тем более доступным является материал [6, с. 236–238]. Автоматический индекс удобочитаемости (*ARI*) высчитывается исходя из суммирования частного количества знаков на количество слов с пробелами и частного количества слов на число предложений. После вычисления тексту присваивается один из 14 уровней, первый из которых сопоставим с уровнем чтения детей детского сада – первого класса, а последний – с преподавателем высшего учебного заведения.

Индекс Колман – Лиану анализирует текст, высчитывая среднее количество знаков и предложений на 100 слов [10, с. 283]. Индекс *SMOG* вычленяет слова в тексте, в которых присутствуют три и более слога. В рамках создания и внедрения индекса Дэйла – Чалл был разработан словарь, включающий в себя 3 тыс. простых для понимания слов, которые должен знать ребёнок, обучающийся в четвёртом классе. Программа автоматически считывает данные слова, и чем больше их присутствует в тексте, тем доступнее является материал [8, с. 404–405].

Материал и методы исследования

Материалом в рамках настоящего исследования послужили информационные и предписывающие тексты, размещённые на информационных стендах в социальных и медицинских учреждениях г. Красноярска в 2021–2023 гг. Корпус из 41 ед. был сформирован методом сплошной выборки

и обработан с помощью общенаучных методов: наблюдение, анализ, систематизация, контент-анализ, лингвистический анализ текста.

С целью демонстрации работы русскоязычного сервиса по оценке читабельности представим изображение, послужившее демонстративным материалом в рамках настоящего исследования.

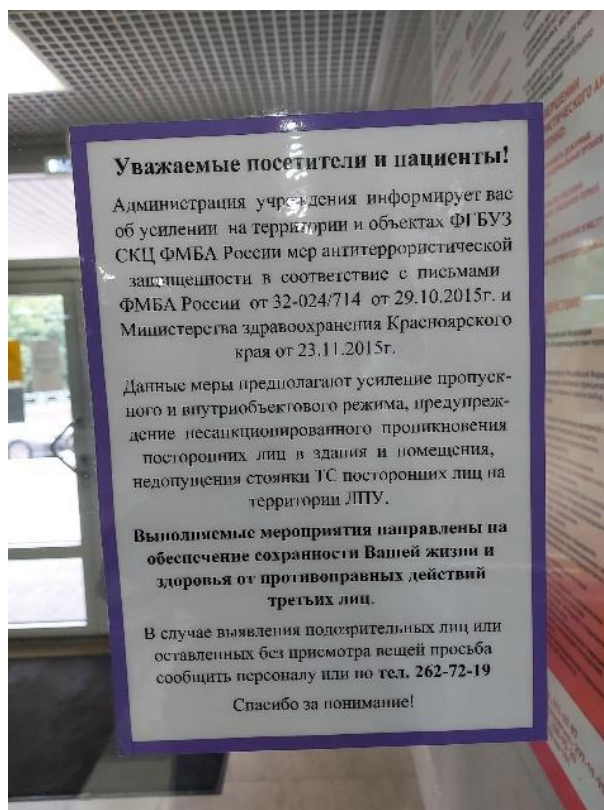


Рис. 1. Демонстративный материал

Данный текст является информирующим текстом, размещённым на входе в государственное медицинское учреждение. Визуально данный текст достаточно легко воспринимать – деление на абзацы, межстрочные отступы, а также выделение полужирным начертанием наиболее важной для реципиента информации способствуют плавному чтению текста. Углубляясь в более детальный анализ и интерпретацию текста, можно выделить следующее: несмотря на наличие вышеперечисленных элементов текста, положительно влияющих на его восприятие, он является сложным для понимания за счёт обилия длинных дву- и более составных слов, аббревиатур, нерелевантных числовых данных, сложного синтаксиса, включающего в себя сложные предложения с причастиями, страдательным залогом, однородными членами предложения, притяжательного падежа и т. п.

Русскоязычная платформа автоматической оценки читабельности текста *plainrussian.ru* выдаёт следующий результат (рис. 2).



Рис. 2. Результат автоматической оценки читабельности

Основываясь на методе количественного анализа текста, сервис подсчитал уровень сложности текстовой информации («достаточно сложно читать»), представленной на рис. 1. Кроме того, системой представлена информация о реципиентах данного текста – студенты 1–3 курсов высшего учебного заведения, а также даётся уровень понятности – 13,56. Ниже размещена информация об использованных при автоматизированном анализе индексах вычисления и расчётные показатели текста.

Этот же текст был вбит в другую русскоязычную платформу, которая вычислила следующее (рис. 3).



Рис. 3. Результат автоматической оценки читаемости

Сервис по оценке читабельности *copywritely.com* при вычислении использует два индекса: индекс туманности Ганнинга (рассчитывается исходя из длины и сложности предложения) и индекс удобочитаемости Флеша. Проведённый системой анализ демонстрирует, что полученные с помощью индексов данные не суммируются, в связи с этим не даётся общая оценка читабельности текста, поскольку индекс туманности Ганнинга определяет текст как доступный, в то время как индекс удобочитаемости Флеша даёт противоположную оценку – текст сложно читать. Таким образом, сервис *copywritely* уступает сервису *plainrussian*.

Необходимо подчеркнуть, что обе системы анализируют исключительно количественные данные. Несомненно, в оценке текста от *copywritely* указано, что предложения грамматически сложные и лексически наполненные, однако никаких данных, подтверждающих это, система не выдаёт и в перспективе это не поможет при адаптации данного текста в более доступный вариант.

О сложности вопроса оценки уровня читабельности говорит и Ю. Ф. Шпаковский, отмечающий, что до сих пор:

«а) не найден достоверный и общепринятый метод экспериментального определения трудности восприятия текстов;

б) многие признаки текстов, заведомо существенные в определении трудности, не удалось количественно измерить» [1, с. 72].

Количественный анализ текста является важным при трансформации текста в рамках практики безбарьерной коммуникационной среды. Однако не менее важным остаются и количественные данные – например, есть ли в тексте сложные для понимания конструкции (обратимые конструкции, страдательный залог, не прямой порядок слов, сложноподчинённые связи, причастный и деепричастный обороты и т. д.).

Таким образом, технологии оценки текстового материала являются неотъемлемой частью при создании адаптированных текстов для широкой группы населения. Однако на данный момент представленные системы не позволяют в полной мере реализовать эффективную адаптацию текстовой информации ввиду отсутствия общепризнанных формул по подсчёту количественных данных. Следующая проблема состоит в том, что количественные показатели не предоставляют исчерпывающих и релевантных данных о способах адаптации текстового материала в более доступный вариант. При этом при оценке уровня читабельности не учитываются и качественные показатели – например, шрифт, использование специальных знаков и символов, разные способы выделения информации, выравнивание текста по ширине или центру, включение в текст неиконичных и сложных графических предложений, введение числовых данных и т. п. Создание сервиса, применяющего не только метод количественного анализа, но и качественного, позволило бы адаптировать письменные тексты более эффективно, особенно с применением компьютерных технологий и искусственного интеллекта.

Благодарю профессорско-преподавательский состав Института филологии и языковой коммуникации за вдохновение на написание данной работы.

Список источников

1. Шпаковский Ю. Ф. Оценка трудности восприятия текста / Ю. Ф. Шпаковский // Труды БГТУ. Сер.: Издательское дело и полиграфия. 2012. № 9. С. 72–74.
2. Хитрюк В. В. «Ясный язык»: как сделать информацию доступной для чтения и понимания / В. В. Хитрюк, Е. Н. Сороко, Т. В. Гришан и др. Минск: БелАПДДиМИ, 2018. 42 с.
3. Нечаева Н. В. Ясный и простой языки как средство обеспечения доступности сайтов организаций / Н. В. Нечаева, Э. М. Каирова, И. С. Борщевский // Матер. III Междунаур. НПК. Екатеринбург: Ажур, 2021. С. 105–114.
4. Жеребило Т. В. Термины и понятия лингвистики. Общее языкознание. Социоллингвистика: словарь-справ. / Т. В. Жеребило. Назрань: Пилигрим, 2011. 280 с.
5. Газизулина Л. Р. Сложность и читабельность как критерии оценки учебного текста при обучении иностранному языку в неязыковом вузе / Л. Р. Газизулина // Мир науки, культуры, образования. 2019. № 1 (74). С. 372–374.
6. Tavernier J. Flesch and Dale-Chall Readability Measures for INEX 2011 Question-Answering Track / J. Tavernier, P. Bellot // Lecture Notes in Computer Science. 2012. Pp. 235–246.
7. Maaß C. Leichte Sprache. Barrierefreie Kommunikation / C. Maaß // Das Regelbuch. Berlin: Universität Hildesheim. LIT Verlag, 2015. 191 s.
8. Kaur S. The Influence of Text Statistics and Readability Indices on Measuring University Websites / S. Kaur, K. Kaur, P. Kaur // International Journal of Advanced Research in Computer Science. 2018. Pp. 403–414.
9. Jekat S. Barrieren Abbauen, Sprache Gestalten / S. Jekat, M. Kappus, K. Schubert. Winterthur: ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2018. 66 s.
10. Coleman M. A Computer Readability Formula Designed for Machine Scoring / M. Coleman, T. L. Liau // Journal of Applied Psychology. 1975. No. 60 (2). Pp. 283–284.
11. Canay D. E. Leicht, Leichter, Leichte Sprache – Eine Untersuchung zu Sprachlichen Kodierungen und Mentalen Modellen / D. E. Canay. Kiel: Christian-Albrechts-Universität zu Kiel, 2019. 102 s.
12. Bredel B. Duden: Leichte Sprache. Theoretische Grundlagen. Orientierung für die Praxis / B. Bredel, C. Maaß. Berlin: Dudenverlag, 2016. 356 s.
13. Пресс-релизы // Всемирная организация здравоохранения: оф. сайт. URL: who.int/ru/news.
14. Доступная среда: инф.-аналит. портал гос. программы РФ. URL: zhit-vmeste.ru.
15. Консорциум Кодекс: эл. фонд прав. и норм.-тех. документации. 2022. URL: docs.cntd.ru.

Е. А. Сальников¹, А. А. Бонч-Осмоловская²,

² кандидат филологических наук, доцент

^{1,2} Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

e-mail: ¹ egorsalnikov1@gmail.com, ² abonch@gmail.com

ORCID: ¹ 0000-0003-2572-4112, ² 0000-0001-5826-8286

ПРИМЕНЕНИЕ СТИЛОМЕТРИИ ДЛЯ ОПРЕДЕЛЕНИЯ СГЕНЕРИРОВАННЫХ ТЕКСТОВ

Аннотация. В рамках данного доклада будет проанализировано использование стилометрической метрики дельта Бёрроуза в качестве метода для определения искусственного (т. е. сгенерированного языковой моделью) текста. Данными для эксперимента послужили дневники – как дневниковые записи случайно выбранных авторов, так и дневниковые записи М. М. Пришвина. В качестве данных языковых моделей послужили дневниковые записи, сгенерированные при помощи языковых моделей *ChatGPT* и *Vicuna_12b*. Путём кластеризации подкорпусов на основе дельты Бёрроуза была установлено, что стилометрический подход позволяет достаточно уверенно отличать тексты, написанные людьми, от текстов, сгенерированных языковыми моделями.

Ключевые слова: *LLM, стилометрия, дельта Бёрроуза, ChatGPT*

Введение

С развитием генеративных моделей вопрос автоматического определения сгенерированного контента встаёт всё более остро. Современные большие языковые модели (*LLM*) – такие, как *ChatGPT*, – способны порождать тексты, на первый взгляд, не отличимые от текстов, написанных людьми. При ближайшем рассмотрении в них, конечно же, встречаются логические ошибки, подмена фактов, синтаксическая структура, несвойственная текстам, написанным людьми, а также весьма необычный выбор лексических средств. Однако автоматическое определение «искусственности» текста или хотя бы чёткая формализация задачи всё ещё вызывает существенные трудности. Актуальность данной задачи, с одной стороны, заключается в необходимости определять сгенерированный контент для борьбы с неправомерным использованием крупных языковых моделей – созданием фейк-ньюс или разведением ботов, – а с другой, для получения устойчивого критерия, предназначенного для оценки качества генерации и последующего улучшения языковых моделей.

На данный момент существует достаточно большое количество систем оценки генеративных моделей (или бенчмарков), однако все они обладают рядом существенных недостатков. Большинство подобных систем рассматривают генерацию текста на естественном языке как набор конкретных, чётко формализованных задач – суммаризация [1; 2], ответы на вопросы [3], симплификация [4; 5]. Следствием этого становится чрез-

мерная ориентация на задачу, из-за чего при оценке учитываются лишь определённые аспекты сгенерированного текста. Это в свою очередь также находит отражение и в применяемых внутри них метриках (см. *ROUGE* [6], широко используемый при оценке диалоговых агентов). В то же время некоторые системы оценки генеративных моделей всё же стремятся отойти от излишней ориентированности на конкретную практическую задачу и вместо этого предлагают сконцентрироваться на анализе семантической составляющей текста [7]. Однако вследствие чрезвычайно сложной формализации области семантики такие подходы зачастую ведут к чрезмерному упрощению.

Переходя к детекции сгенерированного контента, стоит отметить, что за последнее время в данной области произошёл существенный прогресс. Чтобы убедиться в этом, достаточно ознакомиться с результатами соревнования *RuATD-2022* [8], целью которого стал поиск наиболее эффективного метода определения сгенерированного текста на русском языке. Так, для бинарной постановки задачи, в рамках которой требовалось определить, является ли текст сгенерированным или написанным человеком, лучшие решения демонстрировали точность около 86 %. Однако большая часть решений, показавших наивысшую эффективность, оказалась построена на дообучении трансформерных моделей, что существенно ограничивает возможности интерпретации результатов. Кроме того, с момента проведения соревнования прошло уже больше года, за это время в области автоматической генерации текста произошли заметные изменения, связанные в первую очередь с появлением *ChatGPT* и других крупных языковых моделей, поэтому применимость методов, продемонстрировавших свою эффективность для моделей прошлого поколения, к более современным моделям вызывает сомнения.

Таким образом, на сегодняшний день всё ещё существует острая необходимость в создании достаточно простого, основанного на формальных показателях текста подхода, способного уверенно определять искусственную или естественную природу текста. В данном докладе будет рассмотрен один из возможных кандидатов на роль данного метода – дельта Бёрроуза, широко используемая в стилометрических исследованиях.

Стилометрия

Стилометрия – метод стилистического анализа текста, предназначенный для атрибуции текстов спорного авторства, а также датировки спорных текстов. За время своего развития стилометрия претерпела ряд существенных изменений. На ранних этапах исследователи отдавали предпочтение количественному анализу и экспериментировали с такими критериями текста, как: средняя длина слова, средняя длина предложения, количество слов той или иной части речи и пр. Эти критерии являлись дискриминирующими в лучшем случае лишь для конкретных текстов и, следовательно, не были универсальными. Переломным моментом для стилометрии стало появление дельты Бёрроуза – метрики стилистической бли-

зости текстов, предложенной австралийским компьютерным лингвистом Джоном Бёрроузом в 2002 г. [9].

Суть метода заключается в следующем. Для корпуса исследуемых текстов формируется выборка наиболее частотных слов. Далее для каждого слова подсчитывается его частота по анализируемому тексту, затем – средняя частота по всему корпусу, а далее разность показателей делится на стандартное отклонение. Таким образом данная величина, названная *z-score*, рассчитывается для каждого слова первоначального набора в каждом из рассматриваемых текстов. Следовательно, обладая набором *z*-оценок одного и того же набора слов для двух разных текстов, становится возможным измерить расстояние между ними, что и является дельтой Бёрроуза (1). Формула дельты имеет следующий вид:

$$\Delta = \sum_{i=1}^n \frac{|z(x_i) - z(y_i)|}{n}. \quad (1)$$

С момента появления эффективность и надёжность дельты Бёрроуза многократно подтверждалась эмпирически [10], вследствие чего на данный момент метод широко применяется для стилометрических исследований – в т. ч. и для атрибуции текстов спорного авторства.

В то же время задачу определения сгенерированного текста вполне можно представить как атрибуцию текста спорного авторства с той лишь разницей, что в данном случае стилистическая близость спорного текста должна устанавливаться по отношению не только к потенциальному «автору-человеку», но и к потенциальному «автору-машине» – в данном случае, языковой модели. Недавнее исследование [11] продемонстрировало, что *ChatGPT* всё ещё не в состоянии скопировать авторский стиль (в случае рассматриваемой работы – стиль десяти английских авторов, в т. ч. Диккенса) и тем самым «обмануть» дельту Бёрроуза. Но что касается текстов на других языках – например, на русском? А текстов другого тематического домена? Наконец, свойственна ли эта проблема другим языковым моделям или же является недостатком, присущим исключительно *ChatGPT*? Это ещё предстоит выяснить.

Данные

В качестве данных для исследования были использованы дневниковые записи, полученные из корпуса «Прожито» [12]. Дневники как жанр обладают определённой спецификой и тем самым представляют особый интерес для исследования. Дневники обладают ярко выраженным «авторским отпечатком», что делает применение к ним стилометрического анализа ещё более актуальным. Кроме того, дневники как жанр не обладают чётко формализованными критериями, поэтому случайный набор текстов, относящихся к дневникам, будет обладать достаточно высокой гетерогенностью. Всё это делает дневники идеальным кандидатом для тестирования подхода, основанного на применении дельты Бёрроуза.

В качестве основного набора текстов были выбраны дневниковые записи Михаила Михайловича Пришвина. Такой выбор был в первую очередь обусловлен обилием дневников Пришвина. Среди них было несложно выделить дневниковые записи, удовлетворяющие критерию минимального объёма в 500 символов. Кроме того, дневники Пришвина известны достаточно хорошо, а значит, теоретически могли так или иначе быть представлены в тренировочных данных для *ChatGPT*, что в свою очередь имело шансы отразиться на качестве генерации текстов «в стиле Пришвина».

Помимо этого, был также составлен другой датасет оригинальных дневников, составленный из дневниковых записей авторов, относящихся к различным временным периодам. Такой смешанный датасет представлял интерес для сравнения, выступая в качестве текста, написанного усреднённым «автором-человеком».

Подготовка датасетов

В общей сложности в эксперименте было использовано четыре датасета, объём каждого из которых составлял 50 текстов. Длина каждого текста, вошедшего в тот или иной датасет, составляла от 500 до 1 500 символов.

В качестве базового набора текстов было решено выбрать 50 дневниковых записей из дневника М. М. Пришвина (*PRISHVIN_ORIG*). Эти тексты использовались как для дальнейшего сравнения с помощью дельты Бёрроуза, так и для предшествующей ей генерации дневниковых записей с помощью генеративных моделей.

Второй датасет (*RANDOM*) был составлен из 50 оригинальных дневниковых записей различных авторов, относящихся к XIX, XX и XXI вв. Теоретически индивидуальный авторский сигнал в данном подкорпусе выражен менее ярко и вместо него выступает некий усреднённый коллективный сигнал, свойственный «автору-человеку».

Два оставшихся корпуса представляют собой два набора по 50 текстов, сгенерированных двумя различными языковыми моделями – *ChatGPT* (*PRISHVIN_GPT*) и *Vicuna_12b* (*PRISHVIN_VICUNA*). Генерация в обоих случаях проходила по методу *few-shot learning*, когда модели на вход подаётся несколько примеров текстов исходного домена и предлагается продолжить цепочку. С примером такой «затравки» (*prompt*) можно ознакомиться в табл. 1.

В качестве текстов для затравки были использованы дневниковые записи Пришвина, часть из которых также попала в корпус *PRISHVIN_ORIG*. После генерации около 100 записей тексты объединялись между собой до достижения минимальной длины в 500 символов. Таким образом, одним сгенерированным текстом считались несколько объединённых записей общим объёмом более 500 и менее 1 500 символов.

Суммарное количество токенов для каждого датасета представлено в табл. 2.

Таблица 1

Prompt	<p>Напиши дневниковую запись за 10.01.1921 от лица М. М. Пришвина: Святочный сон: в лесу на сосне привесил я портрет своей Козочки, волки стали собираться к портрету, а я их стрелять. Разрушитель не дикий человек, а недоучка, тот, кто, выкрав, напр., из физического кабинета спираль Румкорфа, знает, что её можно приспособить для закуривания папирос. –1 Р. Снег мелкий, но частый и сильный ветер. Ночью на 29-е – буря и +5 Р.</p> <p>Напиши дневниковую запись за 16.01.1921 от лица М. М. Пришвина: Искусство есть способность человека изображать предмет своей веры и любви (Христ.). (Возрождение – предмет наслаждений.) Вера без дел мертва, а вера без любви – зла и есть, кажется (надо подумать), основа величайших злодейств. Зло существует на кредит любви. Сатира, пародия – должники искусства.</p> <p>Напиши дневниковую запись за 26.01.1921 от лица М. М. Пришвина:</p>
Результат генерации ChatGPT	<p>Сегодня я прочитал долгое и принципиальное письмо от одного из моих адвокатов. Мне предстоит сразу же начать подготовку к защите своей позиции, поэтому я встал рано и занялся прочтением литературы по юридической теории. Я размышляю над тем, как использовать закон в мою пользу, чтобы достичь моей цели. Я верю, что правая юридическая аргументация поможет мне достичь моей цели. Да будет мне помощью мудрость и да будет мне поддержкой судьба!</p>

Таблица 2

Название датасета	Размер (токены)
RANDOM	9 471
PRISHVIN_ORIG	9 843
PRISHVIN_GPT	6 222
PRISHVIN_VIGUNA	7 312

Результаты экспериментов

Расчёт дельты Бёрроуза был произведён при помощи библиотеки *stylo* для языка *R* [13]. С результатами кластеризации на основе дельты можно ознакомиться на дендрограмме (рис. 1).

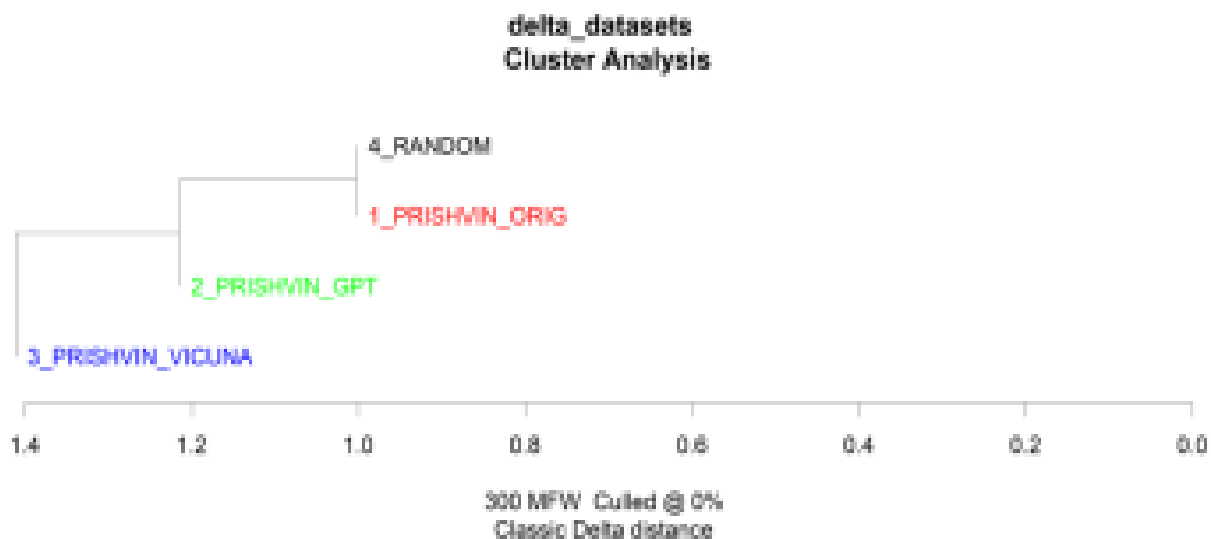


Рис. 1. Дендрограмма

Как видно на дендрограмме, оригинальные тексты Пришвина оказались наиболее стилистически близки к случайному набору оригинальных дневниковых записей. Расстояние между ними составляет всего 0,9130. Далее на большем отдалении от оригинальных записей Пришвина находятся тексты, сгенерированные при помощи модели *ChatGPT*. Расстояние между ними составляет 1,2979. И, наконец, наименее близкими по стилю к оригинальным записям Пришвина оказались тексты, сгенерированные моделью *Vicuna_12b*, – расстояние для данной пары составляет 1,3483.

Из вышеприведённых результатов можно сделать следующие выводы. Во-первых, несмотря на сравнительно высокое качество генерации, современные крупные языковые модели всё ещё не способны уловить стилистические особенности исходного текста. Достаточно простой метод дельты Бёрроуза уверенно определяет, кем именно был написан текст – человеком или машиной. Примечательно и то, что тексты, собранные из записей нескольких «людей-авторов», чей авторский сигнал теоретически должен быть зашумлён, всё же оказываются стилистически ближе к оригинальным текстам Пришвина, чем к текстам, сгенерированным языковой моделью. Эта особенность позволяет сделать предположение, что крупные языковые модели всё ещё не способны улавливать определённые, пока ещё сложноформализуемые аспекты языка, свойственные исключительно текстам, написанным людьми.

Во-вторых, эксперимент доказывает, что неспособность улавливать стилистические особенности исходного текста свойственна не только модели *ChatGPT*, но и, вероятно, всем крупным языковым моделям. В случае текущего эксперимента стилистическая разница между оригинальными текстами и текстами, сгенерированными *Vicuna_12b*, оказалась даже выше стилистической разницы между оригинальными текстами и текстами, сгенерированными *ChatGPT*. Для более уверенных выводов по данному вопросу необходимо в дальнейшем повторить эксперимент, используя вместо *ChatGPT* и *Vicuna* крупные языковые модели, специально обученные под русский язык (например, *GigaChat* от *Sber*).

Заключение

В заключение следует отметить, что метод стилометрического анализа текста, основанный на дельте Бёрроуза, на данный момент является перспективным кандидатом на роль универсального способа автоматического определения сгенерированного текста. Тем не менее для подтверждения или опровержения полученных результатов в дальнейшем необходимо провести эксперименты с применением различных вариаций стилометрической дельты, с использованием более разнообразных текстов различных тематических доменов, а также используя для генерации другие крупные языковые модели, лучше предназначенные для работы с текстами на русском языке.

Авторы доклада выражают благодарность Михаилу Мельниченко за предоставление корпуса дневниковых записей «Прожито» и Борису Орехову за консультации по стилометрии.

Список литературы

1. Narayan S. Don't Give me the Details, Just the Summary! Topic-aware Convolutional Neural Networks for Extreme Summarization / S. Narayan, S. B. Cohen, M. Lapata // arXiv Preprint. 2018. arXiv: 1808.08745.
2. Scialom T. MLSUM: the Multilingual Summarization Corpus / T. Scialom et al. // arXiv Preprint. 2020. arXiv: 2004.14900.
3. Rastogi A. Towards Scalable Multi-domain Conversational Agents: the Schema-guided Dialogue Dataset / A. Rastogi et al. // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. Vol. 34. No. 5. Pp. 8 689–8 696.
4. Xu W. Optimizing Statistical Machine Translation for Text Simplification / W. Xu et al. // Transactions of the Association for Computational Linguistics. 2016. Vol. 4. Pp. 401–415.
5. Alva-Manchego F. ASSET: a Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations / F. Alva-Manchego et al. // arXiv Preprint. 2020. arXiv: 2005.00481.
6. Lin C. Y. Rouge: a Package for Automatic Evaluation of Summaries / C. Y. Lin // Text Summarization Branches Out. 2004. Pp. 74–81.
7. Lin B. Y. CommonGen: a Constrained Text Generation Challenge for Generative Commonsense Reasoning / B. Y. Lin et al. // arXiv Preprint. 2019. arXiv: 1911.03705.
8. Shamardina T. Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian / T. Shamardina et al. // arXiv Preprint. 2022. arXiv: 2206.01583.
9. Burrows J. "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship / J. Burrows // Literary and Linguistic Computing. 2002. Vol. 17. No. 3. Pp. 267–287.
10. Hoover D. L. Testing Burrows's Delta / D. L. Hoover // Literary and Linguistic Computing. 2004. Vol. 19. No. 4. Pp. 453–475.
11. Reborá S. GPT-3 vs. Delta. Applying Stylometry to Large Language Models / S. Reborá // La Memoria Digitale XII Convegno Annuale AIUCD. 2023. Pp. 292–297.
12. Центр «Прожито»: оф. сайт. URL: prozhito.org.
13. Eder M. Stylometry with R: a Suite of Tools / M. Eder, M. Kestemont, J. Rybicki // DH. 2013. Pp. 487–489.

Е. А. Долуденко¹, Л. А. Горохова², А. В. Горохова³,

^{1,2} кандидат филологических наук, доцент;

³ старший преподаватель

¹ Адыгейский государственный университет, Майкоп, Россия

² Пятигорский государственный университет, Пятигорск, Россия

³ Кубанский государственный университет, Краснодар, Россия

e-mail: ¹ ellen-313@mail.ru, ² gorohova@pgu.ru, ³ giacinta-veneziana@yandex.ru

ORCID: ¹ 0000-0002-7861-3742, ² 0000-0001-6675-3257, ³ 0000-0002-9233-3958

ОПЫТ КОРПУСНОГО АНАЛИЗА ДНЕВНИКОВЫХ ЗАПИСЕЙ СТУДЕНТОВ-ПЕРЕВОДЧИКОВ

Аннотация. Целью настоящего исследования является изучение корпуса текстов дневниковых записей, созданных студентами в процессе освоения английского языка как иностранного, с точки зрения морфологического состава и уровня используемой лексики, а также гендерных особенностей речи. Материал исследования представлен корпусом дневниковых записей на английском языке, создававшихся студентами-переводчиками в рамках учебного задания с 2010 по 2023 гг. общим объёмом более 1 млн словоупотреблений. Был осуществлён частотный анализ используемой лексики посредством *AntConc*, а также кластерный анализ с помощью библиотеки *Stylo*. Помимо этого, было проведено сопоставление текстов, созданных девушками и юношами, а также современных текстов и записей, созданных более 10 лет назад.

Ключевые слова: *дневники, корпусные исследования, гендерные особенности речи, английский как иностранный, цифровые гуманитарные исследования*

Введение

Ведение дневника на изучаемом языке с целью совершенствования навыков письма и расширения речевой компетенции обучаемых – не новая идея. Ей посвящено множество исследований и соответствующих публикаций в зарубежной и отечественной методической литературе: [2; 3; 7; 14; 21; 24]. Дневник способствует развитию аналитических навыков и активизирует творческое начало – неперенный атрибут эффективной коммуникации. Работа над дневником позволяет обучаемым абстрагироваться от строгих канонов академического и конвенционального письма, даёт возможность свободно выражать свои мысли и чувства, рефлексировать по поводу собственного опыта, как положительного, так и отрицательного, приводить в порядок мысли, избавляться от негативных впечатлений и т. п. Кроме того, ведение дневника выполняет чисто дидактическую функцию: расширяет репертуар лингвистических средств обучаемого, совершенствует лексические и грамматические навыки, формирует функциональную грамотность, т. к. речь в дневниковых записях чаще всего идёт о повседневных делах и проблемах.

Критический подход к обучению рассматривает язык как средство достижения конечного результата, который позволит обучаемому обрести контроль над своей жизнью, анализируя свою деятельность и её социальный контекст [19]. Замечено, что студенты, обучающиеся на программах подготовки специалистов в области перевода, достаточно успешно осваивают ту или иную отраслевую специфику, могут эффективно осуществлять коммуникацию в условиях профессиональной деятельности, но при этом теряются и впадают в ступор, когда речь идёт об использовании изучаемого языка для решения элементарных бытовых проблем. Ведение регулярных дневниковых записей, отражающих повседневную жизнь обучаемого, призвано восполнить указанный пробел.

По мнению многих исследователей, работа над дневником способствует росту мотивации и вовлечённости обучаемых [16; 26]. Конечно, такая деятельность предусматривает некий диалог обучаемого с преподавателем, наличие обратной связи, комментирование типичных ошибок, однако основное назначение дневника – стимулировать аутентичную языковую практику за пределами учебной аудитории [20]. Личный дневник как продукт рефлексивного письма способствует развитию критического мышления, формирует автономность действий обучаемого, снижает уровень тревожности и укрепляет самооценку [11].

Тексты, создаваемые студентами в ходе изучения английского языка как иностранного, в последние годы неоднократно становились объектом изучения корпусной лингвистики [1; 8; 12; 13; 15]. Методы корпусного исследования позволяют сравнить тексты, созданные обучающимися с различной языковой компетенцией [10], проанализировать используемые ими синтаксические структуры [17], проследить употребление определённых конструкций в диахронии [22] и т. д. Использование дневниковых записей в качестве материала для корпусных исследований также является распространённой практикой [4–6; 23; 25]. Тем не менее потенциал применения методов корпусной лингвистики для анализа дневниковых записей, созданных студентами, изучающими английский язык как иностранный, ещё не раскрыт в полной мере. Как бы то ни было, исследования такого рода могут представлять интерес в силу того, что предполагают большую самостоятельность учащихся по сравнению с другими учебными заданиями (например, сочинениями и эссе на заданную тему), а также решение более разнообразных коммуникативных задач.

Таким образом, *цель* данного исследования состоит в диахроническом изучении корпуса текстов дневниковых записей, созданных студентами в процессе освоения английского языка как иностранного, с точки зрения морфологического состава и уровня используемой лексики, а также гендерных особенностей речи.

Материалы

Материалом исследования послужил корпус дневниковых записей на английском языке, которые создавались студентами-переводчиками в рамках учебного задания по дисциплине «Практический курс перевода»

на протяжении 13 лет (2010–2023 гг.). Объём корпуса составляет более 1 млн словоупотреблений (387 файлов дневников, созданных 283 авторами).

Общий объём текстов за вычетом стоп-слов – 480 432 словоупотребления. При этом число лемм в корпусе за вычетом стоп-слов составляет 16 394. Таким образом, средняя повторяемость слов в дневниковых записях – 62,1.

Объём подкорпуса текстов, созданных авторами-женщинами, составляет 840 тыс. словоупотреблений; объём подкорпуса текстов, созданных авторами-мужчинами, – 168 тыс. словоупотреблений.

Процедура исследования

При разметке данных учитывались такие параметры, как гендер автора, уровень языковой компетенции автора (среднее от оценок в семестре), а также год создания дневника.

Все леммы исследуемого корпуса были ранжированы по частоте употребления с помощью программы *AntConc*. Было проведено сопоставление частотности используемых лексических единиц в подкорпусах текстов, созданных женщинами и мужчинами, а также в подкорпусах «старых» (2010–2013 гг.) и «новых» (2022–2023 гг.) текстов.

Кроме того, полученные данные были проанализированы с помощью инструментов стилометрии, содержащихся в библиотеке *stylo* [18]. Был осуществлён кластерный анализ всего корпуса текстов с использованием классической дельты при $MFW = 100$. Помимо этого, был также проведён анализ подкорпуса созданных женщинами текстов с помощью формулы Крейга [9] с использованием подкорпуса созданных мужчинами текстов в качестве референтного, что позволило выделить «предпочитаемые» и «избегаемые» лексические единицы для рассмотренного подкорпуса. Такой же анализ был проведён для подкорпуса текстов, созданных за последние годы (2022–2023 гг.), с использованием подкорпуса «старых» текстов (2010–2013 гг.) в качестве референтного.

Результаты

Исследование сравнительной частоты употребления лексических единиц в рамках всего корпуса текстов позволило установить, что лексическое ядро данного корпуса целиком составляют слова, характерные для начальных уровней владения языком (A1–A2). Среди 100 наиболее часто встречающихся слов преобладают глаголы (45 %), второе место занимают существительные (24 %), на долю прилагательных приходится 12 %, наречий – 9 %, местоимений – 4 %, числительных – 3 %, служебных частей речи – 3 %.

Все единицы, вошедшие в первую сотню по частоте употребления, связаны с повседневной деятельностью. Наиболее широко представлены семантические группы «Время» (например, *time, day, today, year, hour, week* и т. д.) и «Учёба» (например, *university, class, homework, study, lesson* и т. д.).

Кластерный анализ корпуса дневниковых записей позволил сделать ряд выводов. В подавляющем большинстве случаев была выявлена общность стилометрических характеристик текстов, принадлежащих одному и тому же автору (созданных в течение разных семестров). При этом такие параметры, как гендер и уровень языковой компетенции авторов, не оказывали значимого влияния на попадание текстов в один и тот же кластер.

Анализ предпочитаемой лексики показал, что в дневниковых записях, созданных авторами-женщинами, слова, входящие в семантические группы «Семья» (например, *sister, mom, family, wedding, kids* и т. д.) и «Еда» (например, *cook, coffee, cake, tasty, pizza* и т. д.) встречаются значительно чаще, чем в референтном корпусе «мужских» текстов. Также можно отметить, что для текстов, созданных женщинами, более характерно использование эмоционально окрашенной лексики, преимущественно с положительной коннотацией (например, *beautiful, amazing, nice, delicious* и т. д.).

Сопоставление текстов, созданных в последние годы, с написанными в 2010–2013 гг. не позволило выявить каких-либо закономерностей в отношении состава предпочитаемой лексики, помимо слабо выраженного тренда на повышение частотности употребления слов, связанных с восточными языками и культурами (например, *Chinese, Arabic, Korean* и т. д.) в современных записях.

Обсуждение

Преобладание низкоуровневой (A1–A2) лексики в дневниковых записях студентов с большой долей вероятности объясняется тем, что, как правило, в процессе выполнения задания обучающиеся задействуют свой активный словарный запас, не прибегая к справочной литературе. Имеет значение и то, что содержание дневников вращается в основном вокруг таких тем, как семья, учёба, работа, досуг, лексика для описания которых закладывается на самых ранних этапах изучения языка. Тем не менее возможно модифицировать задание, чтобы мотивировать обучающихся использовать более высокоуровневую лексику.

Важной особенностью дневниковых записей студентов оказалось то, что тексты, принадлежащие одному автору, стабильно демонстрируют высокую степень сходства стилометрических характеристик. Это позволяет строить предположения о существовании «индивидуального стиля» применительно к данным текстам, несмотря на то, что авторы пишут на неродном для себя языке. Одной из причин этого явления, возможно, является то, что студенты используют одни и те же выражения для описания повторяющихся ситуаций своей повседневной жизни, что обуславливает появление характерных особенностей текстов одного автора. Отсутствие значимого сходства стилометрических параметров между текстами авторов с одинаковым уровнем языковой компетенции с большой степенью вероятности объясняется тем, что, чем выше навык владения языком, тем больше разнообразие используемых лексических единиц и конструкций, в резуль-

тате чего становится затруднительным выделить особенности, характерные для всех авторов с данным уровнем языковой компетенции.

Безусловно, заслуживают внимания выявленные гендерные различия, которые могут объясняться как действительной разницей в повседневной жизни юношей и девушек, так и особенностями самопрезентации студентов, которые неосознанно стремятся соответствовать принятым в их культуре нормам и стереотипам касательно поведения, характерного для определённого гендера.

Результаты представленного исследования могут быть использованы для повышения эффективности обучения путём моделирования образовательного процесса с учетом полученных данных, а также для расширения базы корпусных исследований текстов, создаваемых обучающимися в ходе освоения английского языка как иностранного.

Дальнейшие направления исследования могут включать в себя сопоставительный анализ текстов дневниковых записей, созданных носителями разных языков, а также более детальное изучение характерных особенностей данных текстов в зависимости от гендерной принадлежности авторов и их языковой компетенции.

Список источников

1. Гвишиани Н. Б. Пишем по-английски: к выпуску веб-версии Международного корпуса «английский как иностранный» / Н. Б. Гвишиани // Вестник МГУ. Сер. 9: Филология. 2021. № 5. С. 45–57.

2. Евсева Л. Н. Использование нарративного метода заполнения рефлексивного учебного дневника с целью развития умения критической рефлексии в ходе изучения дисциплины «Иностранный язык» студентами нелингвистических направлений подготовки / Л. Н. Евсева, С. А. Жемчугова, Е. Л. Сафроненкова // Мир науки. Педагогика и психология. 2019. № 5. URL: cyberleninka.ru/article/n/ispolzovanie-narrativnogo-metoda-zapolneniya-refleksivnogo-uchebnogo-dnevnik-a-s-tselyu-razvitiya-umeniya-kriticheskoy-refleksii-v-hode.

3. Игнатенко И. И. Формирование автономии студентов магистратуры в изучении иностранного языка / И. И. Игнатенко // Преподаватель XXI в. 2017. № 2-1. URL: cyberleninka.ru/article/n/formirovanie-avtonomii-studentov-magistratury-vizucheni-inostrannogo-yazyka.

4. Мазуркина Е. С. Эмоциональность гендерно маркированных воспоминаний в англоязычных онлайн-дневниках / Е. С. Мазуркина // Актуальные проблемы межкультурной коммуникации: матер. IV Всеросс. студ. НПК (М., 2020). М.: РУДН, 2020. С. 171–177.

5. Мельниченко М. А. «Прожито» от рукописи до корпуса: сбор, разметка, анализ дневниковых текстов / М. А. Мельниченко, Н. Б. Тышкевич // Цифровая гуманитаристика: ресурсы, методы, исследования: матер. Междунар. НК в 2 ч. (Пермь, 2017). Ч. 1. Пермь: ПГНИУ, 2017. С. 134–137.

6. Тивьяева И. В. Мемориализация пандемии коронавируса в сетевых дневниках русскоязычных и англоязычных участников интернет-коммуникации / И. В. Тивьяева // Русистика и компаративистика: сб. тр. / гл. ред. С. А. Васильев. Т. 15. М.: Книгодел, 2021. С. 236–261. DOI: 10.25688/2619-0656.2021.15.14.

7. Юсупова Л. Н. Ведение дневника на английском языке в рамках курса по академическому письму / Л. Н. Юсупова // Концепт. 2016. № S19. URL: cyberleninka.ru/article/n/vedenie-dnevnik-a-na-angliyskom-yazyke-v-ramkah-kursa-po-akademicheskomu-pismu.

8. McEnery T. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use / T. McEnery, V. Brezina, D. Gablasova et al. // *Annual review of Applied Linguistics*. 2019. No. 39. Pp. 74–92. DOI: 10.1017/S0267190519000096.
9. Craig H. Shakespeare, Computers, and the Mystery of Authorship / H. Craig, A. F. Kinney. Cambridge: Cambridge University Press, 2012. 256 p.
10. Espada-Gustilo L. Linguistic Features that Impact Essay Scores: a Corpus Linguistic Analysis of ESL Writing in Three Proficiency Levels / L. Espada-Gustilo // *3L Language, Linguistics and Literature: the Southeast Asian Journal of English Language Studies*. 2011. Vol. 17. No. 1. Pp. 55–64.
11. Hussein H. Students' Reflective Journals and Creative Writing in EFL / H. Hussein, D. Jamal, I. Sadi // *Universal Journal of Educational Research*. 2020. No. 8. Pp. 3484–3495.
12. Ishikawa S. The ICNALE Guide: an Introduction to a Learner Corpus Study on Asian Learners' L2 English / S. Ishikawa. 1st ed. Abingdon: Routledge, 2023. 230 p.
13. Kwon H. English Learner Corpora and Research in Korea / H. Kwon // *Corpora*. 2022. Vol. 17. No. 1. Pp. 5–22.
14. Larrotta C. Journaling in an Adult ESL Literacy Program / C. Larrotta // *New Directions for Adults and Continuing Education*. 2009. No. 121. Pp. 35–44.
15. *Learner Corpus Research Meets Second Language Acquisition* / ed.: B. Le Bruyn, M. Paquot. Cambridge: Cambridge University Press, 2021. 300 p.
16. Lo J. Enhancing Students' Engagement and Motivation in Writing: the Case of Primary Students in Hong Kong / J. Lo, F. Hyland // *Journal of Second Language Writing*. 2007. Vol. 4. No. 16. Pp. 219–237.
17. Lu X. A Corpus-based Evaluation of Syntactic Complexity Measures as Indices of College-level ESL Writers' Language Development / X. Lu // *TESOL Quarterly*. 2011. Vol. 45. No. 1. Pp. 36–62.
18. Package 'stylo': оф. сайт. URL: cran.rproject.org/web/packages/stylo/stylo.pdf.
19. Pennycook A. Introduction: Critical Approaches to TESOL / A. Pennycook // *TESOL Quarterly*. 1999. No. 33. Pp. 329–348.
20. *Promoting Journal Writing in Adult Education: New Directions for Adult and Continuing Education (J-B ACE Single Issue ... Adult & Continuing Education)* / ed.: L. M. English, M. A. Gillen. 1st ed. Hoboken: John Wiley & Sons, 2001. 122 p.
21. Rana L. B. The Use of Dialogue Journals in an ESL Writing Class from Vygotskian Perspective / L. B. Rana // *Journal of NELTA Surkhet*. 2018. No. 5. Pp. 1–14. DOI: 10.3126/jns.v5i0.19481.
22. Siyanova-Chanturia A. Multi-word Expressions in Second Language Writing: a Large-scale Longitudinal Learner Corpus Study / A. Siyanova-Chanturia, S. Spina // *Language Learning*. 2019. Vol. 70. No. 2. Pp. 420–463.
23. Stolberg D. Canadian Heritage German Across Three Generations: a Diary Based Study of Language Shift in Action / D. Stolberg // *Journal of Historical Sociolinguistics*. URL: degruyter.com/document/doi/10.1515/jhsl-2019-0005/html.
24. Tanner M. The Effects of Dialogue Journals in Enhancing ESL Student's Writing / M. Tanner, L. Clement // *Proceedings of the Deseret Language and Linguistic Society Symposium*. 1997. Vol. 23. No. 1. Pp. 113–121.
25. Hall-Lew L. The Lothian Diary Project: Investigating the Impact of the COVID-19 Pandemic on Edinburgh and Lothian Residents / L. Hall-Lew, C. Cowie, S. McNulty et al. // *Journal of Open Humanities Data*. 2021. Vol. 7. No. 4. Pp. 1–5. DOI: 10.7488/ds/3009.
26. Rosário P. Writing Week-journals to Improve the Writing Quality of Fourth Graders' Compositions / P. Rosário, J. Högemann, J. C. Núñez et al. // *Reading and Writing*. 2017. No. 30. Pp. 1009–1032. DOI: 10.1007/s11145-016-9710-4.

Emmanuel Ngue Um
University of Yaounde I, Cameroon

THE ERASURE OF AFRICA IN THE ERA OF CHATGPT

Former French president Nicolas Sarkozy claimed that African societies haven't made enough history in a speech he gave in Dakar on July 27, 2007. His comment drew harsh condemnation for being rooted in racism, contempt, and a lack of knowledge of African history [4–6]. After the political correctness outcry over President Sarkozy's remark fades, what remains is the same undeniable fact that Africa continues to occupy a marginal position in canonical human history and more significantly, the history of knowledge. A clear illustration of this is how Sub-Saharan African institutional structures, particularly in the fields of the economy, political and judicial systems, and education, are supported by values and standards that uphold the colonial legacy rather than they are appropriately informed by systematic African frameworks of knowledge. Having received their education in colonially born institutions, the majority of the continent's so-called educated elites of today, including myself, adopt the prejudiced notion of universal science, where "universalism" more precisely refers to the structure of global power relationships than it does to the inclusion of the knowledge economy on a global scale. In this situation, being knowledgeable as an African frequently entails being able to replicate and internalize epistemic values and contents elaborated by Westerners. The mastery of colonial languages is also necessary in order to be knowledgeable as an African. In Sub-Saharan Africa, providing higher education in an African language like Swahili or Amharic is still a rare exception. Particularly notable for their disrespect for schooling in African languages are the former French colonies [9].

Speaking from personal experience, I happen to be currently the head of a department of a state-run college in Cameroon which specializes in teacher training for indigenous languages and cultures. This department is infamously underappreciated when it comes to the allocation of positions in competitive entrance exams and the hiring of teaching staff, not to mention the scorn and contempt that students or this department receive from their families, fellow students, professors in other departments, non-teaching staff, and society at large. Yet, colonial and imported languages including French, English, Spanish, German, Italian, Chinese, and more recently, Russian, have made significant inroads in Cameroon's educational system.

Possibly the only continent in the world where the reclaiming of colonial identities can result in ferocious contestations and political turmoil is Africa, which includes French colonialist-conquered satellite territories scattered throughout the Indian Ocean. This is demonstrated by the "Anglophone crisis"

in Cameroon, where citizens from the areas that were colonized by the British before the country gained its independence are fighting for fair acknowledgment of their British heritage against their numerically superior French-speaking counterparts. Paradoxically, the administrative border between the English- and French-speaking regions crosses over fluid language and ethnic spaces, which is evidence of the arbitrary nature of such divisions. However, pre-colonial histories and trajectories of the cultures involved will continue to be muddled as such distinctions increasingly take on social significance and turn into instruments of political bargaining for the post-colonial elites. The colonialists' narratives will fill in the gaps left by society's ignorance of their own past in the interim.

Unsurprisingly, the traditional eras that characterize the intellectual development of humanity, from Antiquity through the Middle Ages to the Pre-modern period, glaringly omit Africa. There is nothing like African Antiquity or Middle Age Africa that systematically accounts for the history of societies living in this region of the world in those periods, save from those related to encounters with Easterners and Westerners.

The claim that "[T]he four basic wisdom traditions were founded around 800 to 300 BCE in four geographic regions: Greece, Palestine, India, and China" made by Vervaeke and Ferraro [13, p. 23] in the book "The Scientific Study of Personal Wisdom" chapter "Relevance, meaning, and the cognitive science of wisdom" shocked me. I grew up in an African environment where attaining wisdom or "cognitive improvement that affords a good life" [13, p. 21] is a fundamental goal of social life. In large part, and as President Sarkozy put it, the apparent vacuum in pre-colonial history in Africa can be attributed to the fact that post-colonial Africa has not yet adjusted to the structures that will allow its societies to produce a commodified history similar to that of antiquity and the middle ages in the Western and Eastern worlds. As an example, 70 years of post-colonial academics have resulted in an insignificant number of text resources that could be used to create a corpus of literature that really counts as African literature. The phrase "African literature" is typically used as a catch-all for works written by Africans in non-African languages. Such proxy terminologies are just unthinkable in the context of Russian, English, German, or Japanese literature, in all respects.

The concept of hermeneutical injustice suggests that power relations can limit an individual's ability to understand their own experience [3]. Within a century and a half of their encounter with European colonialists, African societies have seen substantial modifications that have transformed their sense of identity [8]. As ChatGPT and its aliases are set to become fully furnished knowledge providers in the future, it can be feared that Africa's share in the global economy of knowledge will decrease even further, possibly to the point of quasi-invisibility regarding vast swathes of knowledge systems, values, beliefs, technologies, philosophy, literature, and the arts. Currently, investments in AI are focused more on improving the performance of Large Language Models (LLM) than on diversifying AI. African societies are grappling

with the epistemic inequities inherited from colonialism at the same time as they don't appear to be prepared to handle the added challenge posed by AI. These societies' knowledge systems remain largely rooted in oral tradition and other non-written and codified "archives".

African epistemologies and perspectives on the knowledge economy, which contributed to social capital and delivered well-being to communities for millennia, resulting in stable empires and kingdoms, must not be discounted in the face of the paradigm brought about by Western and Eastern knowledge systems, values, and beliefs. Research on mathematics education in Africa has shown, for instance, that games and numeral systems widely used in traditional societies of Africa exhibit the logical foundations and creative processes of canonical mathematics [7; 11].

This essay does not aim to provide anti-colonial justifications for the reclamation of African knowledge systems. Up until this point, there has been a lot of research done and literature produced on this. My argument centers on the consequences of the fundamental shift from information to agency in knowledge production processes. Until the event of the Fourth Industrial Revolution (4IR) [12], knowledge has been a resource created by socioeconomic agents. However, in the age of Artificial Intelligence, knowledge has evolved into an auto-reproductive agent. As a result, knowledge that is unavailable in formats that permit such self-agency will inevitably disappear into silence. More importantly, the artifacts that mediate self-generated knowledge are likely to define new modes of literacy, much like how written knowledge came to have authority over oral knowledge in Antiquity.

The analysis I am proposing in this essay is a daring but realistic assessment of why I think Africa is destined to serve as a placeholder for artificially generated or reconstructed knowledge. I will lead this reflection in three sections.

In the first part of my paper, I discuss how the epistemic injustices that underpin humanities scholarship in Africa have given truth value to stories and viewpoints on knowledge that conform to experimentation more than experiences, probabilistic likelihood more than factuality, and generality more than details. Here, I will focus on the case of Africa's entangled linguistic "archive", whose tuple-based epistemologies have examined the African linguistic reality in arithmetic terms rather than as an intricate system revealed through fluid cultural spaces. I will demonstrate how ethnolinguistic representations of groups' attitudes toward one another and themselves receive little consideration in African linguistics, and how linguistic nomenclatures and classifications generally run opposed to the expansive scopes of socio-historical networks [9].

Given the psychological subservience of the African elite to the intellectual patronage of the West, in the second part of my discussion, I'll adhere to a plainly pessimistic assessment about Africa's ability to establish new knowledge paths in the age of AI, let alone catch up with dominant paradigms. I will cite the overall dismal state of the continent's educational vision and the tethered

relationship between political and social institutions and knowledge in support of this line of reasoning. A radical step toward creating shared infrastructures for the production, dissemination, and sharing of knowledge, such as choosing to promote one or two vehicular languages throughout the continent, seems to be necessary for a significant transformation of Africa's knowledge economy to take place, despite the ongoing political turmoil in various West African countries against the backdrop of anti-colonialist demands.

The critique of Africa's knowledge infrastructures and structures will be expanded to include perspectives on education in the third part of my discussion. I'm willing to wager that many Africans, like myself, feel oppressed by what Bourdieu [1] refers to as the "conformant family" that is, a sphere in which researchers are constrained by habit and where codes and values are transmitted from the center to the periphery [8]. I am convinced many educated African people share my dissatisfaction with the kind of education they receive, and if they had the option, they would give their children a different education. However, the conformant family's dominant only way of thinking is such a potent machine that it inhibits alternate thinking. The cohort majority should accept as valued whatever the power mechanisms that dominate the conformant family deem valuable according to the culture of such intellectual enclosures.

In spite of the glaring dead-end of hyper disciplinarianism of neo-liberal education which Digital Humanities (DH) scholarship and Artificial Intelligence have challenged in recent years, curricula in African schools and universities have vastly remained unchanged. I believe that the term "Digital Humanities" encompasses more than just the skill sets necessary for enabling humanities research. DH is a new epistemology, as Professor Elisabeth Burr stated in a 2019 speech at the University of Leipzig. I believe that DH is also a new hermeneutics that exposes the interdependence of all things, as the Dalai Lama's interconnectedness perspective on education is described in his biography [2]. In this sense, a vision that seeks to change how African societies relate to knowledge is one that transforms people into subjects rather than objects of education. I am certain that if DH scholarship is implemented in Africa in a way that fosters a deeper awareness by teachers and students about everything that knowing entails, education in Africa may become a setting where intelligence and compassion are fostered simultaneously. If my projection is correct, then perhaps AI is a chance for Africa to escape erasure.

References

1. Bourdieu P. *La Distinction* / P. Bourdieu. Paris: Editions de Minuit, 1979.
2. Dalai Lama. *The Wisdom of Compassion* / Dalai Lama, V. Chan. London: Transworld Publishers, 2012.
3. Fricker M. *Epistemic Injustice: Power and the Ethics of Knowing* / M. Fricker. Oxford: Oxford University Press, 2007.
4. Gassama M. *L'Afrique Répond à Sarkozy: Contre le Discours de Dakar* / M. Gassama. Paris: P. Rey, 2012.

5. Kounkou C. L'ontologie Négative de L'Afrique. Cahiers D'études Africaines 198–199–200 / C. Kounkou. 2010. URL: journals.openedition.org/etudesafricaines/16349.
6. Lecolle M. Le Discours de Dakar. Représentations et Stéréotypes Dans un Ngué Um Emmanuel (2022). Numérique, Éducation Culturelle et Développement en Afrique Francophone. Rencontre Scientifique sur le Numérique Dans L'espace Francophone – 2IF, Lyon, France. DOI: 10.5281/zenodo.6719415discours. En Afrique sur L'Afrique. Le Discours et la Langue, Journal of French Linguistics and Discourse Analysis / M. Lecolle // Ethnotypes and Sociotypes: Norms, Discourse, Cultures. 2009. No. 1 (1). Pp. 39–57.
7. Mizony M. Les Jeux Stratégiques Camerounais et Leurs Principes Mathématiques / M. Mizony // Annales de la Faculté des Sciences. 1971. No. 6. Pp. 19–38.
8. Ngué Um E. Africa's Multilingualism and Epistemic (In)justice / E. Um Ngué // Multilingualism, Identity, and Language Endangerment in Africa / ed.: E. Gutova, K. Korangy. Springer Reference Series.
9. Ngué Um E. Numérique, Éducation Culturelle et Développement en Afrique Francophone / E. Um Ngué // Rencontre Scientifique sur le Numérique Dans L'espace Francophone – 2IF, Lyon, France. 2022. DOI: 10.5281/zenodo.6719415.
10. Ngué Um E. Had Ferdinand de Saussure Spoken Wolof or Basaa... , the Discipline of Linguistics Would Have Fared Differently / E. Um Ngué // Language, Culture and Society. 2020. Vol. 2. No. 1. Amsterdam: John Benjamins Publishing. Pp. 107–116.
11. Njock G. E. Mathématiques et Environnement Socioculturel en Afrique Noire / G. E. Njock // Présence Africaine. 1985. No. 135. Pp. 3–21.
12. Schwab K. The Fourth Industrial Revolution / K. Schwab // Geneva: World Economic Forum. 2016.
13. Vervaeke J. Relevance, Meaning and the Cognitive Science of Wisdom / J. Vervaeke, L. Ferraro // The Scientific Study of Personal Wisdom. From Contemplative Traditions to Neuroscience / ed.: M. Ferrari, N. M. Weststrate. New York: Springer, 2013. Pp. 21–51.

А. В. Полоян¹, К. Д. Приходько²,

¹ кандидат филологических наук

^{1,2} Южный федеральный университет, Ростов-на-Дону, Россия

e-mail: ¹ avpoloyan@sfedu.ru, ² kpri@sfedu.ru

ORCID: ¹ 0000-0002-5620-1624, ² 0009-0005-8556-8301

СОСТАВЛЕНИЕ ПРОМПТА В НЕЙРОСЕТИ MIDJOURNEY ДЛЯ ГЕНЕРАЦИИ ИЛЛЮСТРАЦИЙ К ТВОРЧЕСТВУ А. П. ЧЕХОВА

Аннотация. Промпт-инженер, владеющий описанным в справочных материалах инструментарием определённой нейросети, для формирования оптимальных запросов должен изучать не только инструкции, но и лучшие практики, опыт коллег. В статье рассмотрены различные варианты построения запросов к нейросети *Midjourney* для генерации иллюстраций к произведениям А. П. Чехова, а также отмечены наиболее распространённые ошибки и артефакты нейросети в прорисовке объектов.

Ключевые слова: промпт-инжиниринг, искусственный интеллект, нейросеть *Midjourney*, генерация изображений, А. П. Чехов

Вполне отрегулированная законодательно система авторского права заставляет разработчиков, желающих дополнить свой проект яркими иллюстрациями, искать подходящие по стилю и сюжету изображения на отведённых для этого площадках стоковых изображений. Настоящим технологическим прорывом в этой области стали нейросети, генерирующие изображения по текстовому запросу или по картинке-образцу. Теперь во многих подобных нейросетях грамотно составленным промптом (сущ., образованное конверсией от английского глагола *to prompt* – «побуждать, породить») [5] стало возможным задавать и сюжетную линию, и стиль изображения. Из ряда уже зарекомендовавших себя нейросетей для создания изображений (*Kandinsky* от Сбер, «Шедеврум» от «Яндекс», *Imagen* от Google, *Stable Diffusion* от *StabilityAI*, *Dall-E* и *Midjourney* от *OpenAI*, *Craiyon*, *Playgroundai*, *StarryAI*, *Writesonic* и др.) нейросеть *Midjourney* часто выделяют как наиболее функциональный и поддающийся тонким настройкам инструмент [2]. Именно *Midjourney* была выбрана нами для создания иллюстраций по мотивам произведений А. П. Чехова к проекту *Chekhov Digital* [4, с. 153].

Составление промпта, ведущего к получению от моделей ИИ желаемого результата, – процесс трудоёмкий, требующий от оператора вполне определённых знаний и умений, что породило спрос на относительно новую профессию – промпт-инженер. Работа промпт-инженера сосредоточена в основном на двух основных направлениях: на разработке запросов, которые генерируют ответы на основе языковых моделей, а также на улучшении самих языковых моделей для обеспечения более точных и релевантных выходных данных [3].

Безусловно, вполне качественный результат нейросеть может выдать и на основе простого запроса из 2–3 слов на русском языке (большая часть нейронных сетей «понимает» русский язык за счёт машинного перевода). Однако использование английского языка в промпте с учётом некоторых основных правил формализации даёт преимущество в затратах по времени на поиск требуемого результата.

Само название «нейросеть» и её структура вдохновлены принципами работы головного мозга: нейросеть состоит из «нейронов» – более простых алгоритмов, обрабатывающих объекты и множество их признаков. Когда нейросеть обрабатывает запрос, она обращается к объекту и принадлежащим ему атрибутам. В 2014 г. метод анализа формальных понятий был модифицирован введением в формальный контекст отрицания в качестве значения для атрибутов. Т. е. характерное отсутствие какого-либо свойства у объекта – это такой же атрибут, как и наличие у объекта какого-либо свойства [1, с. 19]. Для метода анализа формальных понятий стала возможной классическая логика предикатов, что впоследствии позволило уточнять запросы к нейросетям исключениями.

Основу промпта составляет объект (сюжетная часть). Далее через запятую или предлоги объект дополняется характеристиками (художественная часть) и необходимыми свойствами (техническая часть). Если на изображении предполагается несколько однородных объектов, их количество следует указать конкретным числом, но следует помнить, что более семи объектов на данном этапе нейросеть прорисовывает менее чётко и с большим количеством ошибок.

Важно заранее продумать, как будут расположены объекты (Чехов в театре смотрит пьесу), как они выглядят (грустная девушка в длинном платье), какие действия они будут выполнять (дама сидит/стоит/читает), какие аксессуары будут рядом или на объекте (девушка с собакой и книгой). К технической части можно отнести запросы по цветовой гамме (мрачно / яркое освещение / неон / сепия / закат), расположение камеры (селфи / GoPro / фото в движении), детализацию (3d, AAA, 4k), стиль (1890s / реализм). Стили и другие характеристики подробно описаны на сайте docs.midjourney.com.

Примеры запросов, составленных на английском языке.

1. *A lady is sitting on a chair and a man, who is in a cherry blossom garden, Russia 1980s style, 3d, AAA, high quality* (см. прил., рис. 1).

2. *Three sisters, Anton Chekhov style, full shot, 3d* (см. прил., рис. 2).

В рамках данного исследования для произведений А. П. Чехова было создано и проанализировано около 40 изображений. Из них 12 полностью удовлетворяли запросу (30 %), 16 имели заметные погрешности (40 %), остальные 30 % не подходили ввиду явных несоответствий запросу (например, вместо одной из трёх сестер по центру был прорисован сам Чехов, либо чиновники были одеты довольно современно, как депутаты Государственной думы).

Наибольший интерес вызвали изображения, которые имели едва заметные ошибки и артефакты. Артефактами в *IT*-сфере называют различные искажения – например, предмет, назначение которого трудно определить из-за его бесформенных очертаний, или неуместная тень, больше напоминающая третью ногу у преклонившего колени чиновника (см. прил., рис. 3).

Несмотря на заявленные улучшения в пятой версии *Midjourney*, анатомических ошибок нейросеть делает всё ещё достаточно много: лишние или недостающие пальцы, указательный палец часто длиннее остальных, глаза разного размера или формы, нефокусированный взгляд, нарушение пропорций тела. Например, на рис. 4 у дамы с собачкой не соблюдена пропорция для ног (колени расположены заметно далеко от туловища). На рис. 5 у девушки из-за неправильного расположения большого пальца правая рука выглядит как левая. Если для сюжета, который требуется проиллюстрировать, кисти рук не так важны, то на данном этапе развития нейросетей было бы более правильным их скрывать, указав в промпте параметр *--no x*, где *x* – слово (например, *hand*). Другие виды отрицаний («без», «не» и т. п.) нейросетью игнорируются. К слову, избегать отрицаний помогают синонимы: для рассказа «Дама с собачкой» при первом запросе возле девушки нейросетью был прорисован огромный пёс, поэтому уточнение *not big* (небольшой) было заменено на *small* (маленький), что позволило получить приемлемый результат.

Также в *Midjourney* реализована функция регулировки значимости объекта, её можно как повысить, так и понизить (установить отрицательный вес) командой *--iw*.

1. *Hand --iw -0.8*.

Для понижения значимости также имеется возможность разделить запрос на фрагменты с помощью команды *::* (удвоенное двоеточие). В одном из таких фрагментов достаточно перечислить все нежелательные для прорисовки объекты и назначить данному фрагменту отрицательный вес.

2. *Lady is reading a book :: umbrella, hand :: -0.8*.

В данном случае с понижением веса следует учитывать, что сумма весов всех фрагментов должна быть положительным числом, т. к. чем ближе сумма всех фрагментов к нулю, тем менее предсказуемым будет результат.

Анатомия «страдает» не только у людей, но и у птиц – на рис. 5 крылья птицы ассиметричны, к тому же не являются крыльями чайки. Таким образом, если для иллюстрации важен конкретный вид птицы или животного, в промпт следует добавлять больше описаний либо точное название вида (для чайки со светлым оперением на крыльях – *Larus smithsonianus* – серебристая чайка). Левое крыло птицы имеет артефакт в виде просвета. Положение птицы в пространстве визуально не определено – она находится и на переднем плане, и за девушкой одновременно (нарушение перспективы).

В детализации предметов на исследуемой подборке также были отмечены следующие неточности: поперечная страница в развёрнутой книге (рис. 4), лишние детали у мебели (артефакты) или, напротив, отсутствуют ножки у стула, ручка у чайника, передвижная сила у гужевого транспорта (на рис. 6 все повозки прорисованы нейросетью без лошадей), одна из повозок прорисована лишь наполовину. На рис. 1 ножка стула стоит на подоле платья, хотя при такой посадке она была бы скрыта тканью. Проблема наблюдалась также с парными предметами: один аксессуар из пары может отсутствовать (например, серьга только на одном ухе) или выглядеть иначе (одна из туфель с закруглённым носом, а вторая – с острым).

На сгенерированных иллюстрациях попадают неуместные объекты: например, двойной купол на соборе в форме елочной игрушки (рис. 6), нетипично и расположение величественного собора между жилыми домами.

Встречаются аксессуары, не соответствующие ситуации или эпохе: шариковые ручки в обеих руках у сидящего за письмом Чехова (были изобретены спустя полвека); книга в руках у девушки отпечатана в современной типографии; калькулятор на столе с хаотичным расположением кнопок; несуществующие знаки отличия на мундире генерала; мужчины в шляпах в помещении; обручальное кольцо на левой руке у замужней по сюжету дамы; вишнёвый сад с обильно цветущей сакурой (японская вишня завезена в Россию в начале XX в.).

Вывод. Несомненно, стоит отметить высокое качество генерируемых изображений в нейросети *Midjourney 5* по сравнению с предыдущими её версиями или некоторыми другими нейросетями, существующими на данный момент; а также удобство работы и возможность донастройки полученных результатов. Однако при использовании нейросети для генерации иллюстраций к художественным произведениям промпт-инженер должен быть знаком хотя бы в общих чертах с сюжетной линией произведений, иметь представление об эпохе описываемых событий, культурных традициях. Человеку легче «добрать» необходимые фоновые знания для анализа полученных изображений, чем машине. Также оператору необходимо владеть программным инструментарием и справочным материалом для исправления ошибок и корректировки обнаруженных неточностей в работе нейросети.

Авторы выражают благодарность Анне Барбовой (г. София, Болгария) за помощь в генерации изображений в Midjourney 5.

Список источников

1. Витяев Е. Е. Вероятностные формальные понятия на контекстах с отрицаниями и их непротиворечивость / Е. Е. Витяев, В. В. Мартынович // Информационные технологии в гуманитарных исследованиях: сб. тр. № 19. Новосибирск: НГУ, 2014. С. 5–21.
2. Кузьмин Е. Как пользоваться Midjourney для генерации рисунков и правильно составлять запросы / Е. Кузьмин // Тинькофф журнал. URL: journal.tinkoff.ru/midjourney-tips.

3. Промпт-инженер: какие навыки освоить, чтобы зарабатывать на ChatGPT // Хабр. URL: habr.com/ru/articles/728984.

4. Северина Е. М. Цифровые филологические практики: проект «Chekhov Digital» / Е. М. Северина, А. А. Бонч-Осмоловская, А. М. Кудин // Актуальные проблемы филологии и педагогической лингвистики. 2022. № 2. С. 153–165. DOI: 10.29025/2079-6021-2022-2-153-165.

5. Англо-русский словарь // Cambridge Dictionary. URL: dictionary.cambridge.org.

Приложение



Рис. 1. Результат промпта из примера (1) для пьесы А. П. Чехова «Вишнёвый сад»



Рис. 2. Результат промпта из примера (2) для пьесы А. П. Чехова «Три сестры»

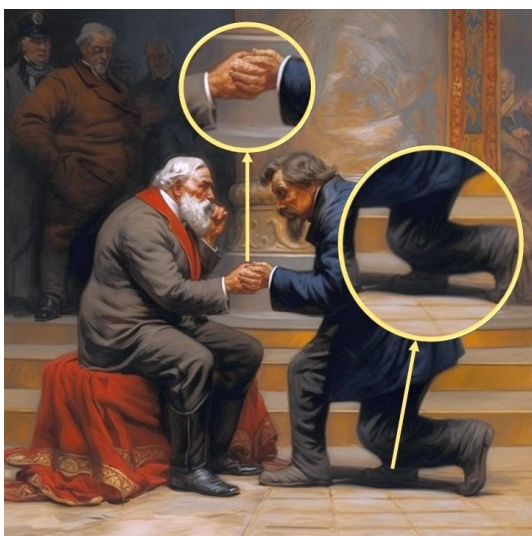


Рис. 3. Иллюстрация к рассказу А. П. Чехова «Смерть чиновника»



Рис. 4. Иллюстрация к рассказу А. П. Чехова «Дама с собачкой»



Рис. 5. Иллюстрация к произведению
А. П. Чехова «Чайка»



Рис. 6. Иллюстрация к произведению
А. П. Чехова «История одного
торгового предприятия»

Т. Ю. Шерстинова¹, М. А. Кирина², Я. К. Хлусова³,

¹ кандидат филологических наук, доцент

^{1,2,3} Национальный исследовательский университет «Высшая школа экономики»,
Санкт-Петербург, Россия

e-mail: ¹ tsherstinova@hse.ru, ² mkirina@hse.ru, ³ yakklusova@edu.hse.ru

ORCID: ¹ 0000-0002-9085-3378, ² 0000-0002-7381-676X, ³ 0009-0004-3619-8191

КОРПУС РУССКОГО РАССКАЗА КАК БАЗА ДЛЯ ПРОВЕДЕНИЯ СОЦИОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РУССКОЙ ЛИТЕРАТУРЫ

Аннотация. Корпус русского рассказа XX в. – новый ресурс, разрабатываемый на базе Корпуса русского рассказа 1900–1930 гг. В статье рассматривается концепция трёх его модулей, которые предназначены для проведения социолингвистических исследований русской литературы: библиографической информации об источнике публикации, биографической информации об авторе текста и информации о социальном бэкграунде литературных персонажей. Обсуждаются перспективы междисциплинарного подхода к изучению литературной системы путём сочетания внешне- и внутри-текстовых социолингвистических факторов, находящих отражение в художественном тексте. Кроме того, описываются механизмы литературного производства через особенности функционирования института периодических изданий советского русского рассказа (региональные, жанровые и т. д.). Создание базы данных публикаций, писателей и персонажей русского рассказа XX в. и их использование для анализа текстов рассматриваемого периода позволят комплексно взглянуть на проблематику создания цифровых литературных ресурсов, требующих исторической контекстуализации, с одной стороны, и на спектр задач, которые можно решать с помощью применения к художественному тексту методов компьютерного моделирования, – с другой.

Ключевые слова: русский рассказ, цифровое сохранение культурного наследия, литературные ресурсы, digital humanities, биографии писателей, литературные персонажи

Введение

В статье рассматриваются возможности использования разрабатываемого литературного цифрового ресурса – Корпуса русского рассказа XX в. – для проведения социолингвистических исследований русской литературы. Этот цифровой ресурс базируется на Корпусе русского рассказа 1900–1930 гг.¹⁷, который содержит информацию о русских писателях рассматриваемой эпохи и их текстах [1; 2]. Расширение исследовательского материала должно позволить названному цифровому литературному ресурсу стать перспективной площадкой для проведения стилиметрических, лингвистических и литературоведческих исследований русской прозы в социологическом контексте.

© Шерстинова Т. Ю., Кирина М. А., Хлусова Я. К., 2023

¹⁷ URL: russian-short-stories.ru.

Идея и ключевые принципы создания Корпуса русского рассказа основываются на представлении о литературной системе, предложенной около 90 лет назад выдающимся представителем русской формальной школы Юрием Николаевичем Тыняновым [3]. По Ю. Н. Тынянову, любой текст, независимо от его литературных достоинств и известности его автора, является «литературным фактом» соответствующей эпохи и потому должен быть принят во внимание [3]. Согласно Тынянову, объективное литературоведение должно анализировать все произведения рассматриваемой исторической эпохи максимального числа писателей, работавших в то время, причём не только в столице и крупных городах страны, но и в провинции.

Цель создания подобного ресурса видится нам в первую очередь в том, чтобы на его основе смоделировать национальный литературный процесс как сложную и динамично развивающуюся систему. Метод, предлагаемый Тыняновым, позволяет говорить об объективности проводимых исследований, т. к. помимо популярных писателей в наш фокус попадают также и «периферийные» авторы. Помимо этого, названный подход помогает добиться более развёрнутого литературного представления различных сторон общественной и культурной жизни, а также языкового и стилистического разнообразия [4]. Следует отметить, что идеи Тынянова, предложенные много лет назад, имеют много общего с современным подходом в литературоведении, известным как «дальнее чтение» [5; 6].

В настоящее время подготовлена часть корпуса для текстов 1900–30-х гг. Его аннотированная выборка, Корпус-300, содержит 310 текстов, написанных 300 авторами – как известными, так и фактически забытыми в наши дни, общим размером около 1 млн словоупотреблений [1]. Художественные произведения данного периода создавались во время драматических изменений в жизненном укладе российского общества: Русско-японская война, Первая русская революция 1905 г., Первая мировая война, Октябрьская и Февральская революции, последовавшая за ними Гражданская война, формирование нового советского государства. В соответствии с этим создателями Корпуса используются три периода для разметки и сопоставления текстов: 1) довоенный период (1900–1913 гг.), 2) период острых социальных катаклизмов – войн и революций (1914–1922 гг.), 3) раннесоветский период (1923–1930 гг.).

В настоящий момент осуществляется расширение Корпуса русского рассказа на 1930–1999 гг. и нормализация собираемых данных в нескольких направлениях. В последующих разделах мы затронем те аспекты, которые, на наш взгляд, представляют интерес в контексте социолингвистических исследований, проведение которых становится возможным на базе рассматриваемого материала: 1) составление библиографической базы публикаций (на основе советских литературных журналов); 2) составление биографической базы русских писателей и нормализация данных; 3) составление базы персонажей русского рассказа.

1. Библиографическая база публикаций

Первым шагом для формирования библиографической базы публикаций писателей, творчество которых приходилось на XX в., стала компиляция списка советских литературных журналов этого периода. Основным источником информации стали следующие ресурсы:

- 1) каталог Государственной публичной исторической библиотеки России¹⁸;
- 2) раздел «Литературные журналы СССР» Википедии¹⁹;
- 3) сайт «Журналы СССР»²⁰;
- 4) статьи и пособия по отечественной журналистике советского периода [9; 11; 13].

Важно отметить, что, как и в [1], было решено отказаться от включения в таблицу детских изданий и эмигрантских журналов. Кроме того, на текущем этапе учитывались только те журналы, в которых есть художественные произведения. Если издание носило чисто публицистический характер, оно не принималось во внимание (даже когда темы статей были посвящены литературе).

Краткое описание и пометки о содержании базировались на материалах таких библиотек, как Областная универсальная научная библиотека²¹, Государственная театральная библиотека²², Президентская библиотека²³, *ImWerden*²⁴, *Internet Archive*²⁵. Метаданные о журналах также дополнялись, помимо уже упомянутых источников, на основе Лексикона русской литературы XX в. [12], ресурсов проекта СовЛит²⁶, указателя «Советская сатирическая печать 1917–1963 гг.» [10]. На сайте «Лаборатория Фантастики»²⁷ удалось найти библиографическую информацию по некоторым журналам.

Для систематизации данных базы советских литературных журналов пилотно была реализована 16-размерная структура. Далее кратко охарактеризуем принципы заполнения базы по выбранным категориями с учётом специфики рассматриваемого материала и параллельной работы с несколькими источниками одновременно.

1. *Id.* Порядковый номер журнала.

2. Название (журнала). Если несколько одноимённых журналов издавалось в одном и том же городе, то рядом с названием указывалась цифра (порядок по году издания, например, московские «Знамя (1)» и «Знамя (2)», первый из которых выпускался в 1919–1922 гг., а второй – с 1931 г. и по сей день). Если такие журналы имели разное место издания,

¹⁸ URL: shpl.ru/readers/helpful_links/internetresursy_po_periodicheskoj_pechati/zhurnaly1/zhurnaly_sovetskogo_perioda_s_1917_po_1991_gg.

¹⁹ URL: ru.wikipedia.org/wiki/Категория:Литературные_журналы_СССР.

²⁰ URL: sites.google.com/site/zurnalyssr.

²¹ URL: booksite.ru.

²² URL: sptl.spb.ru.

²³ URL: prlib.ru.

²⁴ URL: imwerden.de.

²⁵ URL: archive.org.

²⁶ URL: ruthenia.ru/sovlit/index.htm.

²⁷ URL: fantlab.ru.

то в скобках указывалась первая буква города (например, ленинградский «Бегемот (Л)» и симбирский «Бегемот (С)»).

3. Годы издания. Если журнал до сих пор издаётся (например, «Вокруг света»), то ячейка заполнена по формату «1xxx–».

4. Место издания (город). Если имя города менялось, указывается то, которое было актуальным на годы выпуска журнала. Если изменение происходило непосредственно в этот временной период, то название записывалось через дефис (Петроград-Ленинград).

5. Издательство/издатель. Имя частного лица или специального учреждения.

6. Редакторы, секретари, редколлегия, редсовет.

7. Авторы (художественных произведений).

8. Количество выпусков. Подсчитано на основе [9; 10].

9. Частота выпусков в году. Используются следующие пометки: ежемес(ячно)/двухмес(ячно)/еженед(ельно)/двухнед(ельно)/ежегодн(о)/декад(но); н/г – нерегулярно; единств – вышел только один номер, а задумывавшаяся периодичность неизвестна; прочерк – не найдено никакой информации вообще. В большинстве случаев указана задуманная частота, редко – фактическая (журналы закрывались быстрее). Данные получены на основе [9] и дополнительных источников (где необходимо).

10. Количество произведений в номере (в среднем).

11. Описание (журнала). Информация о журнале по [9] и по материалам основным источников: характеристика журнала с точки зрения статуса и содержания (например, независимый, рабочий, иллюстрированный).

12. Доступность. Есть ли возможность найти выпуск(-и) журнала в интернете или библиотеке.

13. Место доступа. Интернет-адрес или название библиотеки:

– Вики – ссылка на страницу о журнале в Википедии (если есть);

– сайт и доп. сайт – ресурсы, на которых можно найти оцифрованные копии изданий и/или какую-либо информацию по ним.

14. Комментарий.

В ходе работы были обнаружены сложности, связанные с недостаточностью материала в открытых источниках после 30-х гг. XX в., а также несостыковками в указателях. Так, например, журнал «Дрезина»: у К. Д. Муратовой обозначен петроградским, у С. И. Стыкалина – московским. Для ряда журналов нет возможности найти электронную версию или хотя бы цифровые копии (например, для журнала «Возрождение» была найдена только интернет-страница по одному из произведений, напечатанных в журнале). В подобных случаях, где возможно, давались ссылки на аукционы, где можно ознакомиться с отдельными страницами редких журналов и их описанием.

На данный момент осуществлена нормализация списка по 150 советским литературным журналам. Кратко охарактеризуем полученные данные. В распределении по местам изданий журналов лидируют Москва и Санкт-Петербург (Петроград-Ленинград), при этом московских журна-

лов в два раза больше, чем петербургских (рис. 1). Очевидно, что информация по этим журналам находилась гораздо легче, чем по региональным. Из последних лучше всего описаны алтайские и сибирские. Место издания трёх журналов не определено.



Рис. 1. Распределение советских литературных журналов по регионам²⁸

Интересно проследить, как распределяются журналы по общему количеству выпусков. Самым выпускаемым оказался журнал «Крокодил» (2 804 номера). Такой отрыв от остальных может быть обусловлен несколькими факторами: во-первых, «Крокодил» издавался на протяжении 86 лет (с 1922 по 2008 гг.), во-вторых, он выходил достаточно часто (3 раза в месяц), в-третьих, это «старейший советский сатирический журнал», пользовавшийся большой популярностью среди граждан [10]. Кроме того, немаловажно отметить, что этот журнал был создан при «Рабочей газете», считавшейся одной из знаковых периодических изданий того времени. Далее следуют журналы с трёхзначным суммарным количеством выпусков (рис. 2). Единственным экземпляром вышло 12 журналов. По трети журналов данных либо не удалось найти, либо они выпускаются до сих пор (27 изданий).



Рис. 2. Журналы, лидирующие по суммарному количеству выпусков в году²⁹

²⁸ На рис. 1 в группу «Другое» объединены те города, которые фигурируют не более 5 раз.

²⁹ Журнал «Крокодил» исключён ввиду слишком большого разрыва.

Как видно из рис. 3, большинство издавалось ежемесячно, однако важно отметить, что многие журналы закрывались быстрее, чем заявленная периодичность становилась фактической. Пятая часть выборки выходила нерегулярно.



Рис. 3. Распределение журналов по периодичности выпусков

Подавляющее большинство журналов носит не только литературно-художественный характер, но и общественно-политический. Кроме того, выделяются журналы на военную тематику (6 журналов). Два издания, по данным Президентской библиотеки³⁰, являлись рукописными: «Красный сапёр» – окопный рукописный журнал Сталинградского, Брянского, Донского и Западного фронтов, и «Заря» – литературно-художественный и общественно-политический сборник Литературного факультета Тульского государственного педагогического института. Также интересны журналы, обзорающие иностранную литературу («Восток», «Вестник иностранной литературы») и др.).

2. Биографическая база данных русских писателей

Задача включения в корпусные тексты максимального числа русских писателей, творчество которых приходится на изучаемый период, тесно связана с другой важной задачей, а именно с сохранением и популяризацией русского литературного наследия, значительная часть которого приходится на советский период. Представляется целесообразным для построения формальной модели литературной системы, а также для изучения авторских стилей учитывать также и некоторые социологические характеристики и особенности биографии писателей. Выдвигается предположение, что такие факторы, как, например, социальная среда, образование, профессия (основная) и др., вносят существенный вклад в формирование его стилистических навыков и мировоззренческих установок. Учёт этих факторов при количественном анализе художественных текстов позволит определить, действительно ли существует связь между социальными характеристиками авторов (и некоторыми важными чертами их биографии) и инди-

³⁰ URL: prlib.ru.

видуальностью его языка и стиля, а также какова действительная мера их влияния.

При формировании списка писателей и сборе биографических данных мы обращались прежде всего к следующим ресурсам:

- 1) Русский биографический словарь – электронная версия Энциклопедического словаря Брокгауза и Ефрона (1890–1907 гг.) и Нового энциклопедического словаря (1910–1916 гг.)³¹;
- 2) Литературная энциклопедия: в 11 т. (1929–1939 гг.)³²;
- 3) Краткая литературная энциклопедия (1962–1978 гг.)³³;
- 4) История русской литературы конца XIX – начала XX в.: библиографический указатель [7].

Главным образом эти ресурсы были полезны для поиска информации по относительно известным писателям. Далее в качестве дополнительных источников использовались данные по Википедии, словарям писателей, биографическим словарям, литературным сайтам, библиотекам и т. п.

Первая версия базы данных содержит три основных модуля: 1) список авторов, 2) биографические данные, 3) источники биографических данных [8]. На данном этапе не было жёстких требований к заполнению полей с биографической информацией, т. к. задача заключалась в сборе этой информации для накопления материала и оценки перспектив его оптимальной нормализации. Кроме того, для ряда писателей универсальный способ сбора данных практически не мог быть выработан, т. к. информация о них отсутствует.

Отметим, что биографический модуль представлен как одна из частей Корпуса русского рассказа (рис. 4). Интегрированы в информационный ресурс данные по 130 персоналиям.

Пётр Стефан Стефанович Гривь-Артыцкий

ФИО писателя: Пётр Стефан Стефанович
Псевдоним: Гривь-Артыцкий
Дата рождения: 1858 (по другим сведениям — 1859)
Дата смерти: 5 апреля 1937
Основные места проживания: Углич
Эмиграция: нет
Социальное происхождение (домы): крестьянин
Образование: Санкт-Петербургское реальное училище А. Крылова, гимназия Б. Мал. астрономический факультет Санкт-Петербургского университета
Профессии: работал в обсерватории Породина, затем потерял зрение
Пол: мужской
Другие комментарии: слепой, потерял зрение, 18 декабря 1935 года вступил в Союз писателей Санкт-Петербургской области, с 7 апреля 1938 года пенсионер и 18 годам исполнилось — предельный возраст по статье 30-10 УК РСФСР (зачисленная пролетарка и интеллигент). Псевдоним: Гривь
Интернет-ссылки: <http://slonov.ru/obitki/Grivya/Poid-323>

Рис. 4. Пример представления информации о писателе на сайте Корпуса русского рассказа

³¹ URL: rulex.ru/be.htm.

³² URL: feb-web.ru/feb/litenc/encyclp/le1/le1-7653.htm.

³³ URL: feb-web.ru/feb/kle/kle-abc/ke9/ke9-1581.htm.

Расширение полей базы данных для проведения более масштабных социологических исследований нам представляется возможным за счёт добавления (в некоторых случаях реструктурирования) следующих полей.

1. Семейное положение (замужем/женат/нет).

2. Родители. Дополнение информации о роде, национальности, происхождении отца и матери. Таким образом можно установить национальность и происхождение самих авторов.

3. Национальность. Сформирован список из 39 национальностей (по данным на 1890-е гг.), который по факту может пополняться.

4. Образование. Деление на общее (да/нет) и высшее (да/нет). Список для заполнения полей может быть переформатирован на основе данных об учебных заведениях на 1917 г. (127 позиций), а также типах учебных заведений (35 позиций, например: военная академия, военная прогимназия, земская школа, духовное училище, курсы, лицей, народный университет, университет и др.). Для учебных заведений актуальна проблема переименований. Из этого выводится необходимость дополнительного преобразования – указания на название, данное заведению при основании, и сопоставления его с современным (табл. 1). Это необходимый шаг ввиду разнородности представления этой информации на разных ресурсах, а также того, что не для всех писателей было зафиксировано название учебного заведения, но был известен его тип.

Таблица 1

Список учебных заведений начала XX в. (фрагмент)

Тип учебного заведения	Название при основании	Современное название
университет	Императорский Московский университет	Московский государственный университет
университет	Императорский Дерптский университет	Тартуский университет
училище	Императорское училище правоведения в Петрограде	–
лицей	Демидовский юридический лицей в Ярославле	Ярославский государственный лицей
учебное отделение	Учебное отделение восточных языков при Азиатском департаменте Министерства иностранных дел в Петрограде	–
институт	Восточный институт во Владивостоке	Государственный Дальневосточный университет

5. Причина смерти. Ориентируясь на данные ВОЗ, расширенная классификация причин смерти может включать такие причины, как аппендицит, бешенство, нападение, корь, самоповреждение, утопление и др. Однако в виду специфики материала и сопоставимого с ним исторического контекста представляется необходимым упоминание в некоторых случаях

не конкретной болезни, а скорее обстоятельств смерти (возраст, каторга, повешение, расстрел, несчастный случай, ссылка и др.). Этот список, на наш взгляд, необходимо расширить и уточнить, отталкиваясь от собранных данных.

6. Место работы. Вспомогательная таблица содержит поля: годы работы, где работал, должность, примечание. Характеристика периодов профессиональной деятельности может быть полезной с точки зрения сопоставления тематической направленности текстов писателей с политикой издательств, где они авторствовали, с одной стороны, и соотношения смежности занимаемых ими в разные периоды должностей с историческим контекстом – с другой.

Особый вопрос в контексте анализа биографических данных занимает не только их сбор, но и нормализация. Приведём несколько примеров.

7. Место рождения, место смерти. На основе данных переписи населения 1897 г. был сформирован список городов (906 позиций) с делением на губернии и соответствующими современными названиями (где возможно) (табл. 2). Если город уже не существует (был затоплен или вошёл в состав более крупного), то в первом случае для современного названия ставился прочерк, во втором – указывалось новое название. Этот вид преобразования кажется уместным для визуализации, которая может быть осуществлена на основе данных путём применения геоинформационных (GIS) технологий.

Таблица 2

Список городов конца XIX в. (фрагмент)

Город	Губерния	Исторический регион	Современное название	Регион	Страна
Мариегамн	Або Бьёрнеборгская губерния	Финляндия	Маарианхамина (Мариехамн)	Провинция Аландские острова	Финляндия
Красный Яр	Астраханская губерния	Центральная Россия	Красный Яр	Астраханская область	Россия
Темрюк	Кубанская область	Кавказ	Темрюк	Краснодарский край	Россия
Ростов-на-Дону	Область Войска Донского	Центральная Россия	Ростов-на-Дону	Ростовская область	Россия
Царское Село	Санкт-Петербургская губерния	Центральная Россия	Пушкин	Пушкинский район Санкт-Петербурга	Россия

8. Профессии. На основе собранных данных сформирован список из 66 профессий, например: актёр, беллетрист, библиотекарь, грузчик, кинемеханик, композитор, кочегар, кузнец, писатель и др.

Значительная по объёму выборка параметров хоть и проблематична, но должна позволить, на наш взгляд, более объективно оценить влияние биографических параметров, например, на сюжет или тематику произведе-

ний, упоминаемые в них локации, создаваемые автором социальные портреты героев. Кроме того, компиляция подобного ресурса в контексте проблематики Корпуса русского рассказа, включающего в себя тексты как популярных, так и малоизвестных авторов, намечает как одно из направлений исследований изучение феномена известности писателя через совмещение моделей его биографического и литературного путей.

3. База данных персонажей русского рассказа

На материале Корпуса-300 была проведена разметка всех персонажей, которые произносили хотя бы одну реплику. Информация об имени, поле, возрастной группе, социальном происхождении, семейном положении, профессии и других характеристиках была занесена в базу данных (табл. 3). Также было отмечено, является ли персонаж главным или второстепенным героем. В результате был получен список из 2 190 персонажей [14].

Таблица 3

Пример заполнения информации о персонажах (А. И. Куприн «В цирке»)³⁴

Персонаж	Полное имя	Профессия, положение в обществе и т. п.	Семейное положение	Пол	Возраст	Социальное происхождение
Луховицын	–	врач при цирке	–	М	–	вероятно высокое
Арбузов (гл. герой)	Никита Ионыч Арбузов	борец	–	М	–	–
Антонио	Антонио Батисто	акробат, приятель Арбузова	женат	М	–	–
Директор	–	директор	–	М	вероятно пожилой	вероятно высокое
Гришутка	–	помощник коридорного	–	М	молодой	низкое
Генриетта	–	жена Антонио	замужем	Ж	–	–
Шталмейстер	–	шталмейстер	–	М	–	–

Предварительные исследования на выборке Корпус-300 позволили сделать, например, наблюдения о социальном положении литературных персонажей и их профессиональной деятельности. Прослеживание распределения персонажей по социальному статусу проводилось в трёх категориях: высокий (доля персонажей этого статуса ко всей выборке – 12,74 %), средний (45,98 %) и низкий (40,64 %). Любопытно, что при анализе хронологической динамики можно заметить, что доля персонажей с высоким социальным статусом резко снижается от первого периода к третьему (21,63 → 14,55 → 3,55 %), в то время как доли героев среднего и низкого социального статуса увеличиваются (42,69 → 45,30 → 49,39 % и 35,53 → 38,33 → 46,94 % соответственно). Что касается определения рода деятельности и профессии героев русской малой прозы, эти поля удалось заполнить для 57,58 % всех персонажей. Самыми распространёнными оказались

³⁴ Опушен столбец «Характеристика автором», содержащий авторское описание персонажа, ввиду значительного объёма.

военнослужащие (11,69 %), рабочие (8,95 %), работники конторы (3,52 %), студенты разных специальностей (3,29 %) и прислуга (2,88 %).

Интересно, как социодемографическая картина мира русского рассказа изменится после 1930-х гг., в период Великой Отечественной войны, изменений в политическом устройстве страны и, наконец, развала СССР. Также кажется целесообразным проследить, как биографические факты из жизни писателей влияли на репрезентацию тех или иных – по статусу, гендеру или профессии – героев в их произведениях. Другое важное направление, которое открывается при наличии подобной разметки, – это исследование речевой дифференциации, влияния социальных факторов на речь героев. Прямая речь персонажей представляет собой не что иное, как имитацию устной речи. Особенности её графического представления в зависимости от статуса персонажа могут служить предметом социолингвистических исследований на базе художественных текстов.

Заключение

Создание литературных ресурсов, содержащих помимо самих текстов биографическую информацию о писателях и персонажах художественных произведений, существенно расширяет возможности изучения национальной литературы в социологическом аспекте, а также языка, стиля и особенностей авторов с учётом их индивидуальных биографических и социологических характеристик. На данном этапе нами предпринята попытка структурировать и проанализировать имеющийся материал о важном источнике художественной литературы – советских литературных журналах; частично была структурирована информация об изданиях исследуемого периода. Кроме того, ведётся работа по компиляции и систематизации базы биографических данных русских писателей XX в. На основе аннотированного подкорпуса осуществляется разметка персонажей рассказов с особым вниманием на указанный для них автором социальный статус, что впоследствии позволит проводить сопоставительный анализ текстов с учётом данных об авторах и особенностям изданий, в которых они публиковались. Предлагаемая методика описания литературных текстов – в добавление к стандартным методам корпусного анализа текстов – будет способствовать построению динамической модели русской литературы на протяжении столетия, а также может стать образцом для построения цифровых гуманитарных ресурсов не только для национальной литературы, но и для других видов искусства: живописи, музыки, театра, кинематографа.

Публикация подготовлена в рамках работы по проекту «Русская литература в социальном измерении: компьютерная платформа СОЦИОЛИТ» в 2023 г.

Список источников

1. Мартыненко Г. Я. Методологические проблемы создания компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX в.) / Г. Я. Мартыненко и др. // Компьютерная лингвистика и вычислительные онтологии. СПб.: ИТМО, 2018. № 2. С. 97–102.
2. Мартыненко Г. Я. О принципах создания корпуса русского рассказа первой трети XX в. / Г. Я. Мартыненко, Т. Ю. Шерстинова, Т. И. Попова и др. // TEL 2018: матер. XV Междунар. конф. по компьютерной и когнитивной лингвистике (Казань, 2018). С. 180–197.
3. Тынянов Ю. Н. Архаисты и новаторы / Ю. Н. Тынянов. Ardis Publishers, 1929.
4. Мартыненко Г. Я. Методы математической лингвистики в стилистических исследованиях / Г. Я. Мартыненко. СПб.: Нестор-История, 2019.
5. Bordoni C. Introduzione Alla Sociologia Della Letteratura / C. Bordoni. Pisa: Pacini, 1974.
6. Moretti F. Distant Reading / F. Moretti. London: Verso, 2013.
7. Муратова К. Д. История русской литературы конца XIX – начала XX в.: библиограф. указ. / К. Д. Муратова. 1963.
8. Шерстинова Т. Ю. Биографическая база данных русских писателей (к созданию корпуса русского рассказа XX в.) / Т. Ю. Шерстинова // Корпусная лингвистика. 2019. С. 439–447.
9. Муратова К. Д. Периодика по литературе и искусству за годы революции 1917–1932 / К. Д. Муратова; вступ. ст. С. Д. Балухатого. Л.: АН СССР, 1933. 344 с.
10. Стыкалин С. И. Советская сатирическая печать 1917–1963 гг. / С. И. Стыкалин, К. И. Кременская. М., 1963. 333 с.
11. Дикушина Н. И. Литературные журналы и газеты СССР: краткая лит. энц. / Н. И. Дикушина; гл. ред. А. А. Сурков. М.: Советская энциклопедия, 1962–78. Т. 4: Лакшин – Мураново, 1967. С. 271–306.
12. Казак В. Лексикон русской литературы XX в. / В. Казак. М.: Культура, 1996. 493 с.
13. Коровин В. М. Литературные журналы в России / В. М. Коровин // Энциклопедия Кругосвет. 2008. URL: krugosvet.ru/enc/kultura_i_obrazovanie/literatura/LITERATURNIE_ZHURNALI_V_ROSSII.html.
14. Иванова О. Ю. Дистрибуция персонажей малой русской прозы (на материале рассказов 1900–1930 гг.): ВКР бакалавра / О. Ю. Иванова. СПб.: НИУ ВШЭ, 2021. 52 с.

М. В. Шилова,
специалист

Томский государственный университет, Томск, Россия
e-mail: marvsh@yandex.ru
ORCID: 0009-0001-1660-5608

ВИЗУАЛИЗАЦИЯ ИСТОРИЧЕСКИХ ТЕКСТОВ: ВЫЯВЛЕНИЕ ГЕОГРАФИЧЕСКИХ СВЯЗЕЙ ЦИФРОВЫМИ СРЕДСТВАМИ

Аннотация. Цель исследования – охарактеризовать существующие цифровые инструменты экспликации географических данных из исторических текстов на картах. В исследовании использованы экспериментальные и сравнительные методы исследования. Полученные результаты сравниваются с затраченным временем. Ожидается, что исследования помогут охарактеризовать современное положение в области цифровой визуализации пространственных данных и определить возможный вектор дальнейшего развития в этом направлении.

Ключевые слова: *Digital Humanities (DH), цифровая гуманитаристика, историческая информатика, GeoAI, историческая картография*

Актуальность

В настоящее время происходит развитие цифровых решений для автоматической обработки и анализа текстов, методов обработки естественного языка. Также развиваются и геоинформационные системы. *Web-сервисы WolframAlfa* [1] и *VoyantTools* [2] разрабатывают технологии отображения данных из текстов на картах. *ArcGIS* [3] в описании своих продуктов заявляет о способности обрабатывать естественные языки. Активно ведутся разработки по направлению *GeoAI*. Например, в августе 2023 г. НАСА и *IBM Research* презентовали первую базовую модель геопро пространственного искусственного интеллекта с открытым исходным кодом. Она включает данные наблюдения за Землёй. *HLS Geospatial FM* доступна в *Hugging Face*, общедоступном репозитории моделей машинного обучения с открытым исходным кодом [4].

Цель

Цель данного доклада – дать характеристику существующему цифровому инструментарию для эксплицирования географических данных из исторических текстов на картах.

Задачи исследования

Провести эксперимент по выбору оптимальной программы для визуализации географических данных на карте в установленных ограничениях – на территории РФ, бесплатно и быстро. Выявить существующие сложности в использовании существующего ПО. Дать рекомендации по уточнению вектора развития в данной сфере.

Методы

В исследовании использованы экспериментальный и сравнительный методы исследования. Будут изучены некоторые существующие цифровые решения и проведены эксперименты по визуализации данных и сравнению полученных результатов и затраченного времени.

Статистика поисковых запросов

Для начала проследим, какова ситуация в сфере программного обеспечения, используемого для создания географических карт, в т. ч. исторических. Одними из самых популярных специализированных сервисов для составления карт являются *QGIS* и *ArcGIS*. По данным *Google Trends* и Яндекс *Wordstat*, количество запросов по этим программам устойчиво держится на уровне 30 тыс. и 15 тыс. запросов в месяц соответственно.

Выбор программного продукта для визуализации географических данных

Поставим экспериментальную задачу – сделать карту фрагмента романа Умберто Эко «Маятник Фуко» [5]. Выбранный фрагмент удобен тем, что уже разбит по годам, но всё же является образцом неструктурированного текста с историческими данными. В прил. 1 представлен предварительно обработанный фрагмент текста на русском языке. Была проведена следующая обработка.

1. Удалены все строки, за исключением тех, в которых упоминается граф Сен-Жермен. Представим, что мы хотим проследить его географический след.

2. Жёлтым цветом выделены реальные географические места, где появлялся Сен-Жермен, в т. ч. по слухам.

3. Голубым выделены те локации, которые не нужно учитывать в визуализации.

4. Добавлена одна важная локация – Пруссия в 1776 г.

5. Синхронизированы локации между русским и английским переводом. Например, в русском переводе указан ландграф Гессе, а в английском – *landgrave of Gessen*, что означает место, а не фамилию, как можно было бы решить исходя из русского перевода.

Визуализация на карте

Начнём с выбора подходящего приложения. Исходим из того, что опыта пользования подобными программами нет, а время ограничено.

Проект *Хронокон* [6]. Не требует скачивания и установки, работает в браузере. Позволяет создать интерактивную карту с таймлайном, подвижными объектами и примечаниями. Бесплатно. Без программирования. Доступно в России. Проект некоммерческий, поддерживается средствами и силами автора – когда средства и силы закончатся, проект закроется, карта исчезнет. Как можно скачать и хранить сделанную карту – не ясно. Субъективная предвзятая оценка времени, которое потребуется на создание карты – 2 дня на первую карту с учётом изучения, 1 день на последующие карты похожей сложности. Также обратим внимание на дружелюбность пользователю. В данном случае пользователь – гуманитарий. Этот

параметр складывается из интуитивности управления, дизайна и из того, насколько привычны пользователю категории, с которыми придётся работать. В данном случае процесс спроектирован как приятное хобби.

Яндекс Карты [7]. Нет таймлайна и подвижных объектов. Бесплатно. Без программирования. Позволяет загружать массивы данных в специальных форматах *XLSX*, *CSV*, *KML*, *GPX* или *GeoJSON*. Позволяет распечатать карту, сохранить её как рисунок, выгрузить данные для просмотра в других программах – например, *Google Earth*. Субъективная предвзятая оценка времени, которое потребуется на создание карты: 2 дня на первую карту с учётом изучения, 1 день на последующие карты похожей сложности. Процесс механический, рутинный, управление несложное. Средний уровень комфорта, т. к. дизайн утилитарный и результат такой же.

ArcGIS [3] – платная, есть бесплатная пробная версия 21 день. Заявляется функционал нейросети – обработка естественного языка и извлечение сущностей. Заблокирована на территории России и Белоруссии.

ArcGIS StoryMaps [8] – бесплатная, но также заблокирована.

Python Folium [9]. Бесплатная, с открытым кодом. Есть понятный учебник. Нужны минимальные навыки программирования. Субъективная предвзятая оценка времени, которое потребуется на создание карты – 3 дня на первую карту, 0,5 дня на последующие. Придётся иметь дело с базами данных, библиотеками. Для непрограммистов интуитивность управления отсутствует.

Global Mapper [10]. Платная программа, нет браузерной версии, есть мобильная версия. Для того чтобы получить пробную версию, требуется регистрация. Но чем ограничивается пробная версия – сроком или функционалом, не указано.

Google Maps [11]. Приблизительно то же, что у «Яндекс Карт». Только более интуитивно.

MapBox [12]. Платная. Есть пробная версия. Что в неё входит, не уточняется. Очень сложная система для широкого круга профессионалов. На изучение и составление первой карты уйдёт 3 дня, и 1 день на то, чтобы сориентироваться в разнообразии продуктов и решений. Последующие карты, вероятно, будут требовать 1 день.

QGIS [13]. Требуется скачивание и установка. Бесплатная с открытым кодом. Есть подробные руководства. Не интуитивная, требует длительного изучения. Предположительно, потребуется 3-4 дня для того, чтобы создать первую карту.

StoryMap [14]. Бесплатный онлайн-сервис для создания геопрезентаций на несколько слайдов. Можно получить визуализацию, которую можно будет разместить на своём сайте. Управление не сложное, но не вполне интуитивное. Займёт 1,5 дня на первую презентацию, и 1 день на следующую.

Voyant Tools [2]. Бесплатный онлайн-сервис для анализа текстов. Есть инструмент *DreamScape*, который позволяет визуализировать на карте данные из неструктурированного текста. Есть возможность ставить пометки на карте. Есть анимация связей и варианты выбора проекции карты. Распознаёт русский язык. Однако и в русском, и в английском, делает ошибки – распознаёт не все географические названия, некоторые распознаёт ошибочно. Поэтому подготовить текст всё же придётся. По предварительной оценке работа с этим сервисом займёт 2–3 ч в первый раз, и 1 ч для следующих работ. Но здесь, как и в других случаях, не учтено время подготовки текста к визуализации.

WolframAlpha [1]. Платный онлайн-сервис аналитических инструментов, в т. ч. и отображения данных на карте. Есть бесплатный функционал, которого может быть достаточно для нашей цели. Платная версия – 7 \$/мес. Работа может занять около 2 ч в первый раз, и 10 мин в следующий. Но результат в бесплатной версии довольно схематичный и в низком разрешении. Интегрирован с *ChatGPT*.

Palladio [15]. Бесплатный онлайн-сервис для визуализации данных. Специально для историков. Требуется навык работы с табличными данными. Прилагается подробное руководство для того, чтобы освоить такую работу. Первое использование может занять 2 дня, большая часть времени уйдёт на перевод текста в нужный формат.

OpenStreetMap [16]. Бесплатный онлайн-сервис с открытым кодом, созданный силами сообщества волонтеров. Продолжает редактироваться. Для того чтобы начать работу, нужно зарегистрироваться. Этот сервис предназначен для создания и редактирования карт местности, но не для визуализации историй.

NodeGoat [17]. Онлайн-сервис для работы с данными. Есть бесплатная версия с ограниченным функционалом. Подробное руководство. Параметрическая модель, где требуется создание объектов (в нашем случае Сен-Жермен) с подробным описанием. Далее можно делать различные типы визуализаций с введёнными данными. Предположительно займёт 3 ч на первое использование и по 1,5 ч на последующие. Предоставляет учётную запись в течение 48 ч – в случае положительного решения после рассмотрения заявки.

Scribble Maps [18]. Бесплатный сервис (есть платная версия). Можно начинать без обучения и без регистрации. Метод создания карт – простой поиск и расстановка меток. Есть возможность сохранить карту в различных форматах – как растровых, так и табличных. Сделать первую карту может занять от 1 ч.

Google Earth [19]. Бесплатный онлайн-сервис. Возможно делать свои проекты-исследования и делиться ими с другими. Увлекательный процесс, не обещающий быть быстрым и простым с первого раза.

Для того чтобы сделать карту, выбраны три сервиса – *Voyant Tools* (VT), *WolframAlfa* (WA) и *Scribble Maps* (SM) – как самые быстрые и простые в управлении (по субъективной предварительной оценке).

Процесс визуализации данных

После анализа возможностей выбранных программ потребовалось очистить текст от лишних слов, оставить только даты и нужные географические названия. Поэкспериментируем и с русским, и с английским текстом.

VT в русском языке не показал страны, только города. Хотя в июле 2023 г. распознавал и страны, и города. Вместо стран впишем их столицы. Сервис не позволяет редактировать введённый текст, также нельзя удалить текст из окна, только перейти на стартовую страницу. При загрузке исправленных данных в четвёртый раз на карте показался лишь один город. Смена браузера не помогла исправить ситуацию. Удалось откатить к предыдущему вводу данных. Английский текст был распознан, но не все города показаны на карте. Результат – на рис. 1.

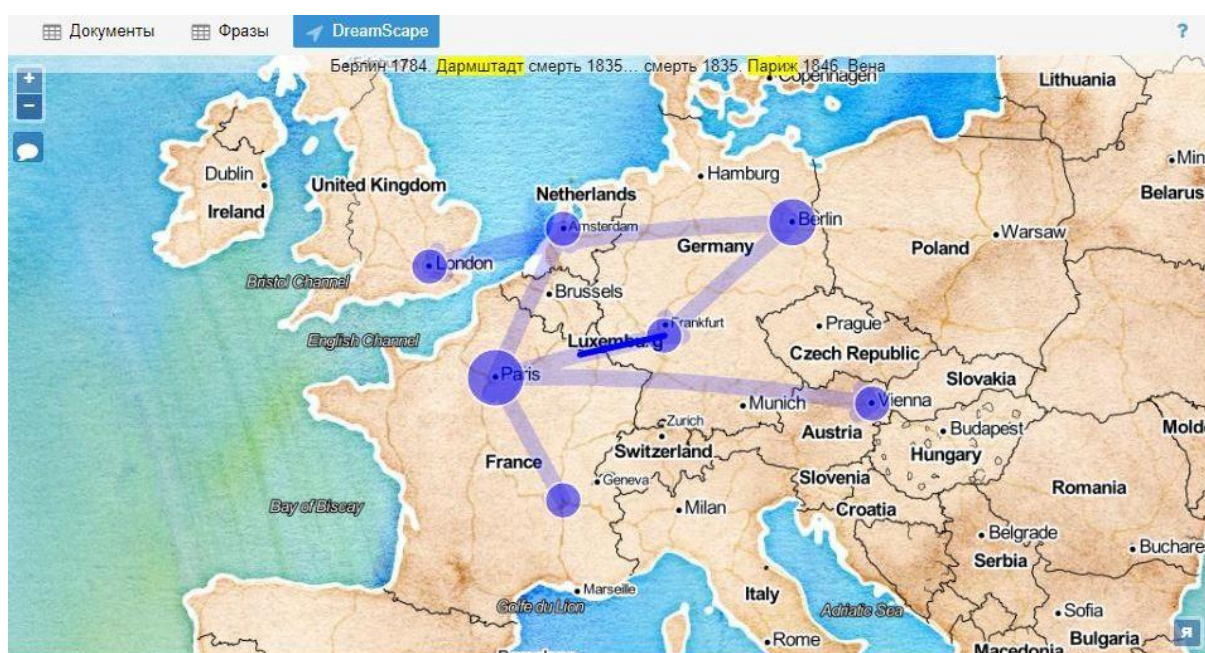


Рис. 1. Результат *Voyant Tools*

Потрачено 40 мин, но задача не решена, т. к. на сервисе возникли неполадки. В окне предупреждения было написано, что это экспериментальный инструмент, поэтому он может работать некорректно.

WA. Для этого сервиса также подходит русский язык. Есть подсказка, что сервис понимает естественный язык. Из 12 городов в запросе *Lyon* не был распознан. На карте были отображены следующие три города – Лондон, Амстердам и Париж. Потрачено 30 мин, задача не решена. Результат – на рис. 2.

SM. Удалось сделать карту за 30 мин. Но большой набор данных займёт больше времени на обучение, т. к. потребуется загрузить подготовленные данные. Результат – на рис. 3.



Рис. 2. Результат WolframAlfa

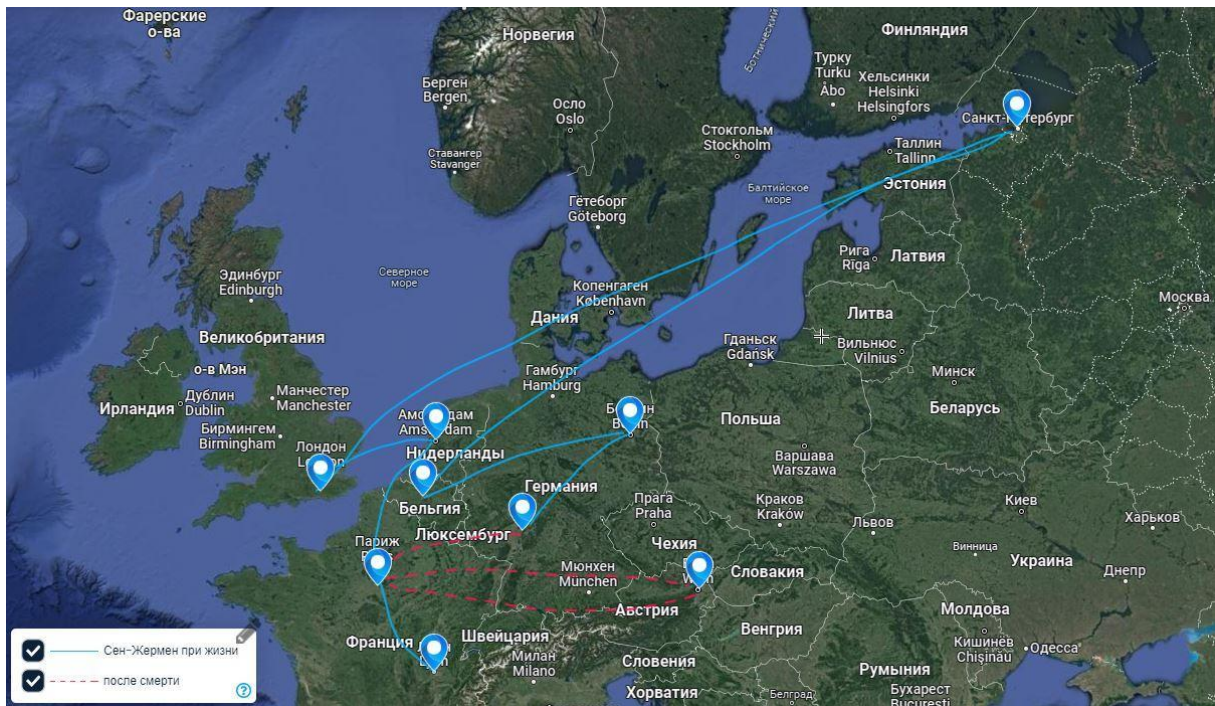


Рис. 3. Результат Scribble Maps

Средства ИИ

На территории России заблокирован доступ к *ChatGPT* от создателей *OpenAI*. Но есть большое количество альтернативных решений. Проверим некоторые из них.

Было проверено четыре генерационные нейросети, ни одна из них не справилась с задачей.

Выводы

Существующие на сегодняшний день цифровые инструменты для визуализации пространственных данных на картах имеют следующие характеристики, препятствующие их широкому использованию:

- 1) требуют специального обучения;
- 2) требуют специальных знаний, таких как опыт работы с наборами данных;
- 3) методы составления карт затратны по времени;
- 4) требуют оплаты;
- 5) недоступны на территории РФ;
- 6) ненадёжны;
- 7) некоторые не имеют функции создания таймлайнов.

Есть ещё одна сложность, субъективная. В ходе эксперимента выяснилось, что предварительная оценка временных затрат и сложности в несколько раз превосходит реальные показатели. Сравним предварительные оценки «Яндекс Карт», которые были в начале списка, и *SM*, которые появились в самом конце, когда *VT* и *WA* не помогли решить задачу.

«Яндекс Карты» были оценены в 2 дня работы, а *SM* – в 1 ч. Может показаться, что вторая программа выглядит намного проще первой. Но это не так, ведь по управлению они построены на одинаковой логике. Просто к моменту, когда был открыт сервис *SM*, сама задача была уже частично решена в процессе работы с *VT* и *WA*, а также был накоплен первичный опыт работы с географической визуализацией. И может оказаться, что иррациональный страх перед неизвестным создаёт предубеждение и искажает предварительную оценку временных затрат.

В нашем эксперименте получилось, что разница между информированной оценкой (*SM*) и оценкой из состояния неизвестности («Яндекс Карты») различается в 16 раз (если считать день за 8 рабочих часов). По этим причинам исторические тексты и лекции не так часто сопровождаются картами, как это было бы полезно для читателей и студентов.

Для улучшения ситуации следует рассмотреть два пути:

- 1) популяризировать работу с табличными форматами, ввести обучение таким форматам в курсы бакалавриата, магистратуры и даже школьного обучения;
- 2) разрабатывать инструменты с обработкой естественного языка, развивать направление, заданное в сервисе *VT*.

Заключение

Владимир Николаевич Владимиров, один из самых значительных деятелей в сфере *Digital Humanities* в России, писал в своей монографии 2005 г., что в освоении стандартного коммерческого программного обеспечения для нужд историков есть значительный резерв. Историческая информатика, только пережившая период становления, нуждается и в оценках её роли в историческом исследовании и образовании, и в скорейшем полноценном внедрении [20].

С тех пор прошло 18 лет (колоссальный срок по меркам *IT*), но значительных изменений в этой сфере не произошло. Визуализация на картах по-прежнему не является обычным навыком любого историка. Хотя пространственно-эксплицитные модели способны выявить такие связи в данных, которые невозможно найти, просто читая текст с перечнем географических названий [21].

Например, из нашего экспериментального кейса можно сделать вывод, что таинственный граф Сен-Жермен вовсе не так вездесущ, как это могло показаться при прочтении книги целиком. За 77 лет жизни он был замечен в семи городах, расположенных недалеко друг от друга. Конечно, здесь не хватает ещё и таймлайна, чтобы проанализировать временные промежутки, а также знаний о возможных скоростях перемещения между городами в то время. Но это задача для следующего исследования.

Благодарность А. В. Бочарову, С. А. Меркулову и Н. С. Пугачёву.

Список источников

1. WolframAlpha. URL: wolframalpha.com.
2. Dreamscape – Voyant Tools. URL: voyant-tools.org/docs/#!/guide/dreamscape.
3. ArcGIS: Mapping & Analytics Software and Services. URL: esri.com/en-us/arcgis/about-arcgis/overview.
4. НАСА и IBM выпустили модель Geospatial AI Foundation с данными наблюдения за Землёй // Хабр. 2023. URL: habr.com/ru/news/752440.
5. Эко У. Маятник Фуко / У. Эко. Litres, 2012. 1 172 p.
6. Хронокон – История на карте. URL: chronocon.org/ru/about.
7. Конструктор карт Яндексa. URL: yandex.ru/map-constructor/?from=maps_login.
8. ArcGIS StoryMaps. URL: storymaps.arcgis.com.
9. Python R. Python Folium: Create Web Maps From Your Data – Real Python. URL: realpython.com/python-folium-web-maps-from-data.
10. Global Mapper // Blue Marble Geographics. URL: bluemarblegeo.com/global-mapper.
11. Google Карты. URL: google.com/intl/ru/maps/about/mymaps.
12. Mapbox. URL: mapbox.com.
13. QGIS. URL: qgis.org/en/site.
14. StoryMap. URL: storymap.knightlab.com/edit/?id=sen-zhermen.
15. Palladio. URL: hdlab.stanford.edu/palladio-app/#/upload.
16. OpenStreetMap. URL: openstreetmap.org.
17. NodeGoat. URL: nodegoat.net/usecases.
18. Scribble Maps. URL: scribblemaps.com/create.
19. Google Earth. URL: earth.google.com/web/@0,0,0a,22251752.77375655d,35y,0h,0t,0r.
20. Владимирова В. Н. Историческая геоинформатика: геоинформационные системы в исторических исследованиях / В. Н. Владимирова. АлтГУ, 2005.
21. Janowicz K. GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond / K. Janowicz et al. // International Journal of Geographical Information Science. 2020. Vol. 34. No. 4. Pp. 625–636.

Приложение

1707. Рождается Клод-Луи де Сен-Жермен, если он действительно рождается.
1743. Первое появление на публике графа Сен-Жермена. **Лионская** ложа изобретает ступень Рыцаря Кадоша, чьё назначение – месть за тамплиеров.
1758. **В Париж** прибывает Сен-Жермен и предлагает свои услуги королю в качестве химика, специалиста по крашению и краскам. Он бывает у Помпадур.
1760. Сен-Жермен и его двусмысленная дипломатическая **Голландии**. Он должен скрываться, в **Лондоне** его арестовывают и затем отпускают. Дом Пернети основывает Общество иллюминатов **в Авиньоне**. Мартинес Пасквинес учреждает Орден Кавалеров Избранных Каменщиков Мира.
1762. Сен-Жермен **в России**.
1763. Казанова встречается с Сен-Жерменом **в Бельгии**. Там он именуется Де Сюрмон и превращает простую монету в золотую.
1776. Сен-Жермен, под именем графа Уэллдона, представляет химические проекты Фридриху II **Прусскому**. Создаётся Общество Филафетов для объединения всех герметических обществ. Ложа Девяти Сестёр, в которую входят Гильотен и Кабанис, Вольтер и Франклин. Вейсгаупт кладет начало **баварским** иллюминатам. По некоторым данным, он был посвящен неким **датским** купцом, Кельмером, возвратившимся **из Египта**, который будто бы – не кто иной, как таинственный Альтотас, учитель Калиостро.
1778. Сен-Жермен встречается **в Берлине** с Домом Пернети. Виллермоз учреждает Орден Добродетельных Кавалеров Града Святого. Строгое Послушание Тамплиеров сливается с Великим Востоком при условии, что будет соблюдаться **Шотландский** Исправленный обряд.
1784. Сен-Жермен якобы погибает, устраивая для ландграфа **Гессенского** фабрику по производству красок.
1835. Каббалист Эггингер утверждает, что встретил Сен-Жермена **в Париже**.
1846. **Венский** литератор Франц Граффер публикует мемуар о встрече его брата с Сен-Жерменом в 1788–1790 гг. Сен-Жермен принимает посетителя, листая книгу Парацельса.
1936. **Во Франции** открывается Великий Приорат де Голь. В «Тетрадах полярного братства» Энрико Контарди ди Роджо рассказывает, как его посетил граф де Сен-Жермен.

В. Ю. Сеницын¹, В. С. Назаровская²,

¹ кандидат физико-математических наук, доцент

^{1,2} Российский государственный гуманитарный университет, Москва, Россия

e-mail: ¹ fpmrggu@yandex.ru, ² nazarovskaya2380@mail.ru

ТЕСТИРОВАНИЕ ЭФФЕКТИВНОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ПРИ ВЫПОЛНЕНИИ ЗАДАНИЙ ЕДИНОГО ГОСУДАРСТВЕННОГО ЭКЗАМЕНА ПО РУССКОМУ ЯЗЫКУ

Аннотация. Представлено сравнительное исследование эффективности больших языковых моделей при выполнении заданий демонстрационного варианта контрольных измерительных материалов единого государственного экзамена 2023 г. по русскому языку. Обсуждаются полученные результаты тестирования больших языковых моделей и возможные перспективы внедрения их в образовательный процесс.

Ключевые слова: *искусственный интеллект, большая языковая модель, ChatGPT, YandexGPT, ЕГЭ, русский язык*

Большие языковые модели (*LLM*), впервые построенные около пяти лет назад, представляют собой предварительно обученные нейронные сети с очень большим количеством параметров. *LLM* предназначены для обработки и генерации текстов на естественных и искусственных языках и имеют в настоящее время очень широкие сферы приложений. Эффективность применения больших языковых моделей стремительно растёт. Так, компания *OpenAI* сообщает, что её языковая модель *GPT-4* уже сейчас обладает производительностью, сопоставимой с человеческой [1].

Системы, основанные на больших языковых моделях, принято считать не полностью контролируруемыми системами искусственного интеллекта. Поэтому особенно важен тот факт, что сейчас многие специалисты отмечают неизбежность решения проблем внедрения *LLM* в учебный процесс, а также связанную с этим необходимость трансформации системы образования [2].

Эффективность применения разных больших языковых моделей при решении различных задач требуется специально измерять, т. к. она заранее неизвестна. Например, исследования показали, что *ChatGPT-3.5* способна решать стандартные задачи курса теории вероятностей и математической статистики для студентов технических вузов посредством генерации кода на языках программирования *R* и *Python* с эффективностью более 80 % [3].

В данной работе получены оценки эффективности *LLM* при выполнении заданий демо-варианта ЕГЭ 2023 г. по русскому языку, который доступен на официальном портале ФИПИ. В исследовании участвовали четыре чат-бота, основанные на больших языковых моделях:

– *ChatGPT-3.5*, взаимодействие с которым осуществлялось через бесплатный аккаунт на сайте компании *OpenAI*;

– *ChatGPT-4*, доступ к нему был получен через премиум-аккаунт *OpenAI*;

– бесплатный *GPT*-бот в мессенджере *Telegram*, который можно найти по тегу *@AIGPT4bot*;

– *YandexGPT*, доступный через «Яндекс Браузер».

В табл. 1 представлены результаты тестовой части демонстрационной версии экзамена по русскому языку. Когда модель генерировала неправильный ответ на задание, она получала ещё одну попытку на исправление ошибки. Если со второй попытки модель находила правильный ответ, эксперт засчитывал полученный балл, но в итоговом подсчёте первичных баллов данная ситуация отражалась в виде интервального оценивания. В случае неудачной попытки исправить ошибку задание оценивалось в 0 баллов.

Таблица 1

Результаты тестовой части демо-варианта ЕГЭ 2023 г. по русскому языку

№ задания (баллы)	<i>ChatGPT 3.5</i> (бесплатный)	<i>ChatGPT 4</i> <i>plus</i>	<i>Bot</i> (<i>Telegram</i>)	<i>YandexGPT</i>
1 (1)	0	1	0	0
2 (1)	0	0	0	0
3 (1)	0	1	0	0
4 (1)	0	0	0	0
5 (1)	0	0	0	0–1
6 (1)	0	0	1	0
7 (1)	0	0	0	1
8 (3)	1	1	2	0–1
9 (1)	0	0	0–1	0
10 (1)	0	0	0	0
11 (1)	0	0	0	0
12 (1)	0	0	0	0
13 (1)	0	0	1	0–1
14 (1)	0	0	0	1
15 (1)	0–1	1	0	0–1
16 (1)	0	0	0	0–1
17 (1)	0	0	0	0
18 (1)	0	0	0	0–1
19 (1)	0	0	0	0–1
20 (1)	0	0	0	0
21 (1)	0	0	0	0
22 (1)	0	1	0	0–1
23 (1)	0	0	0	0
24 (1)	0	0	0	0
25 (1)	0	0	0	0
26 (3)	1	3	1–2	0–1
Количество первичных баллов	2–3 из 30 (7–10 %)	8 из 30 (27 %)	5–7 из 30 (17–23 %)	2–11 из 30 (7–37 %)

Все модели частично справлялись с трудными заданиями № 8 и 26 тестовой части варианта, в которых требовалось сопоставить, как в случае задания № 8, виды грамматических ошибок с представленными предложениями или различные выразительные средства языка с указанными примерами-фразами, как в случае задания № 26. Также почти все модели с успехом отвечали на задание № 15, в котором проверялось владение правилом написания «Н» и «НН» в разных частях речи.

В табл. 2 представлены результаты выполнения задания ЕГЭ № 27. Написанные моделями сочинения оценивались независимым экспертом по официальным критериям, указанным в демо-варианте ЕГЭ по русскому языку. При этом эксперт не знал, что проверял сочинения, написанные не школьниками, а языковыми моделями. Чат-бот *YandexGPT* отказался писать сочинение, объяснив свой отказ тем, что задание выходит за пределы его возможностей. Таким образом, итоговый интервальный результат модели от «Яндекса» составил 5–26 баллов из 100 возможных.

Таблица 2

Результаты сочинения демо-варианта ЕГЭ 2023 г. по русскому языку

Критерии оценивания ответа на задание № 27 (баллы)	<i>ChatGPT 3.5</i> (бесплатный)	<i>ChatGPT 4 plus</i>	<i>Bot (Telegram)</i>
К1. Формулировка проблем исходного текста (1)	1	1	1
К2. Комментарий к проблеме исходного текста (5)	3	3	1
К3. Отражение позиции автора по проблеме исходного текста (1)	0	1	0
К4. Отношение к позиции автора по проблеме исходного текста (1)	0	0	0
К5. Смысловая цельность, речевая связность и последовательность изложения (2)	1	0	0
К6. Точность и выразительность речи (2)	2	2	2
К7. Соблюдение орфографических норм (3)	3	3	3
К8. Соблюдение пунктуационных норм (3)	3	1	1
К9. Соблюдение грамматических норм (2)	2	1	1
К10. Соблюдение речевых норм (2)	2	2	2
К11. Соблюдение этических норм (1)	1	1	1
К12. Соблюдение фактологической точности (1)	0	1	0
Первичные баллы за задание № 27	18 из 24 (75 %)	17 из 24 (71 %)	12 из 24 (50 %)
Первичные баллы за весь ЕГЭ	20–21	25	17–19
Тестовые баллы за весь ЕГЭ	42–43	49	37–40

В табл. 2 для перевода в тестовые баллы применялась официальная 100-балльная шкала, доступная по URL: 4ege.ru/novosti-ege/4023-shkala-perevoda-ballov-ege.html. Согласно правилам ЕГЭ 2023 г. по русскому языку минимальный порог для получения аттестата составлял 10 первичных (24 тестовых) баллов, порог для поступления в вузы – 16 первичных (36 тестовых) баллов, а в подведомственные образовательные учреждения Минобрнауки – 19 первичных (40 тестовых) баллов.

Хуже всего модели справлялись с отражением позиции автора к проблеме текста, а также с выражением отношения к позиции автора, поэтому по критериям К3–К4 практически все модели получили 0 баллов. Также модели с трудом пишут семантически цельный текст, они не всегда соблюдают речевую связность, что отражено оценками по критерию К5. Однако все тестируемые модели соблюдали нормы литературного русского языка. Это свидетельствует о том, что языковая модель – это в первую очередь лингвистический инструмент, предназначенный для генерации текстов в соответствии с литературными нормами и грамматикой языка.

Самым высоко оценённым оказалось сочинение бота *ChatGPT-3.5*. Модель получила за него 18 из 24 баллов, что составляет 75 % от максимального результата. В приложении можно ознакомиться с тремя сочинениями, которые проверил эксперт, и составить своё собственное мнение о возможностях языковых моделей.

Лучшим результатом за весь экзамен отличилась модель *ChatGPT-4*, все правильные ответы на задания теста ею были получены с первой попытки. Важно отметить, что итоговые баллы всех чат-ботов значительно меньше среднего балла по стране за 2023 г., который составляет 66,43 [4]. Однако полученных ботами в процессе экзамена баллов достаточно для допуска к поступлению в вузы.

Худший результат был получен моделью *YandexGPT*. Однако можно отметить, что при решении тестовой части демонстрационного экзамена именно эта модель получила больше баллов с учётом повторного решения заданий. Стоит обратить внимание на то, что модель отечественного производства самая «маленькая» из всех участников тестирования, а также она имеет более жёсткие ограничения на генерацию ответов на запросы пользователей. При одинаково составленном запросе задания № 27 именно чат-бот от «Яндекса» отказался его выполнять. Эта модель всё же способна генерировать тексты на разные темы, что подтверждается различными источниками [5; 6]. Запрос на выполнение задания № 27 не был упрощён и адаптирован именно под эту модель, т. к. целью исследования являлась проверка возможностей нейросетей в сравнении с человеческими возможностями.

Три из четырёх оцениваемых моделей успешно сдали ЕГЭ по русскому языку, что указывает на значительный потенциал больших языковых моделей. Результаты тестирования можно сравнить с результатами ЕГЭ, которые получают абитуриенты. Однако структура единого государственного экзамена понятна для человека, тогда как моделям для успеш-

ной сдачи ЕГЭ по русскому языку необходимы дополнительные пояснения к условиям заданий. При этом качество выполнения заданий ЕГЭ по русскому языку чат-ботами свидетельствует о том, что большие языковые модели на данном этапе разработки и эксплуатации могут быть полезным инструментом в руках квалифицированного пользователя, но не способны пока серьёзно конкурировать с человеком в решении сложных задач.

Список источников

1. GPT-4 Technical Report // OpenAI. URL: cdn.openai.com/papers/gpt-4.pdf.
2. Ивахненко Е. Н. ChatGPT в высшем образовании и науке: угроза или ценный ресурс? / Е. Н. Ивахненко, В. С. Никольский // Высшее образование в России. 2023. Т. 32. № 4. С. 9–22. DOI: 10.31992/0869-3617-2023-32-4-9-22.
3. Синецын В. Ю. О применении ChatGPT в учебном процессе по математическим дисциплинам в вузе / В. Ю. Синецын, В. С. Назаровская // Информация и образование: границы коммуникаций (INFO'2023): сб. тр. № 15 (23). Горно-Алтайск: ГАГУ, 2023. С. 197–199.
4. ЕГЭ-2023 в цифрах: средний балл по разным предметам, стобалльники, двоечники и не только // Мел. Медиа про образование и воспитание детей. URL: mel.fm/blog/mikhail-lantsman/69758-yege-2023-v-tsifrah-sredny-ball-po-raznym-premetam-stoballniki-dvoyechniki-i-ne-tolko.
5. Нейросеть Яндекса пишет письма и сочиняет сказки: главное о YandexGPT // РБК Тренды. URL: trends.rbc.ru/trends/industry/6465e4d39a79474b00a85e06.
6. YandexGPT теперь может кратко пересказать текст // Коммерсантъ. URL: kommersant.ru/doc/6068983.

Приложение

Сочинение, написанное ботом ChatGPT-3.5

В прочитанном тексте В. И. Белова «Деревня» актуализируется проблема духовной привязанности человека к своей малой родине, к местам детства и юности. Автор поднимает вопрос о том, как важно для человека сохранять связь с местами, где он вырос, и как эти места оказывают влияние на его внутренний мир и понимание себя.

Пример-иллюстрация 1. В рассказе автор описывает свои воспоминания о детстве и молодости, проведённых в деревне, о том, как он стремился к высотам, смотрел на леса и прятал «мальчишеские богатства» в щелях кряжей. Этот образ подчеркивает сильную связь с деревней и её важное значение в формировании личности.

Пример-иллюстрация 2. В тексте также упоминается старинный хвойный бор, который шумит торжественно и мудро. Это место символизирует непреходящее и вечное в жизни, что указывает на значимость родных мест в жизни человека.

Эти два примера показывают, что родные места оказывают глубокое воздействие на внутренний мир человека и помогают ему сохранять связь с прошлым.

Автор выражает свою позицию, подчёркивая, что деревня и родные места остаются незыблемыми во времени, что они способны дарить тепло и вдохновение, обогащая личную и духовную сферы человека.

Считаю, что автор правильно поднимает важную проблему сохранения связи с родными местами. В нашем современном мире, насыщенном быстротечными событиями, сохранение воспоминаний и духовной связи с корнями важно для здоровья души и понимания себя. Это позволяет сохранять стабильность и ценность в своей жизни, даже если окружающий мир постоянно меняется.

Сочинение, написанное ботом *ChatGPT-4*

Одной из ключевых проблем, затронутых В. И. Беловым в представленном тексте, является противоречие между стремлением человека уйти от своего прошлого, от родных мест, и неодолимой тягой к возвращению, к корням.

Проблема иллюстрируется следующими примерами.

«Из этой сосновой крепости, из этих удивительных ворот уходил я когда-то в большой и грозный мир, наивно поклявшись никогда не возвращаться, но чем дальше и быстрее уходил, тем яростней тянуло меня обратно». Этот отрывок показывает, что, несмотря на решимость рассказчика уйти и не возвращаться, он все равно испытывает сильное влечение к родным местам.

«В тот день, когда я уходил из дому, так же, как и сегодня, вызванивали полевые кузнечики, так же лениво парил надо мной ястреб, и только сердце было молодым и не верящим в обратную дорогу». Этот пример подчеркивает, что природа остается неизменной, в то время как человек меняется, и его восприятие мира тоже.

Между этими примерами прослеживается смысловая связь: оба отрывка акцентируют внимание на противоречивых чувствах героя, разрывающего его между прошлым и настоящим.

Позиция автора заключается в понимании сложности человеческой природы, в которой прошлое и настоящее тесно переплетены, и одно не может существовать без другого.

Лично я согласен с позицией автора. Каждый из нас несет в себе память о прошлом, и даже если мы стремимся к новым горизонтам, наши корни остаются неотъемлемой частью нас. Невозможно полностью отрезать себя от прошлого, и это делает нас богаче, глубже и мудрее.

Сочинение, написанное *GPT*-ботом в мессенджере *Telegram*

В тексте «Родные места» автор задаёт проблему связи человека с родными местами и смирением перед мощью природы. Он описывает свою встречу с родными местами – деревней, лесом, рекой, и отмечает, какие воспоминания и эмоции они в нём вызывают. Автор обращает вни-

мание на то, как все эти спокойные и единые когда-то места меняются: деревни исчезают, на их месте появляются города, природа и общение с ней уходят на второй план.

Примеры-иллюстрации из текста, которые важны для понимания проблемы.

1. «Детство вписалось в мою жизнь далеким нервным маревом, раскрасило будущее яркими мечтательными мазками». Этот пример показывает, как сильное детское впечатление может остаться навсегда. Детство – это время, когда связь с родными местами наиболее сильна, и эти впечатления остаются с нами на всю жизнь.

2. «Нет, в здешних местах пожары не часты, и лет пятьсот уже не было нашествий. Может быть, так оно и надо?» В этом примере автор отмечает, что изменение в окружающей среде может быть неизбежным и не всегда отрицательным. Он задает риторический вопрос о необходимости изменений и оставляет его без ответа, оставляя каждому читателю возможность задуматься над этим вопросом самостоятельно.

Примеры-иллюстрации связаны между собой, т. к. оба они отражают проблему изменения родных мест и связи с ними. Первый пример показывает силу и продолжительность влияния родных мест, а второй пример поднят в контексте изменений, которые происходят в мире и в самом человеке.

Моя позиция по проблеме исходного текста заключается в том, что связь с родными местами и природой необходима для человека. Родные места, детство, природа – все они помогают нам сохранить связь с нашим прошлым, понять наши ценности и укрепить нас в настоящем. Вместе с тем, необходимо уметь приспособиться к изменяющемуся миру и принять его, т. к. изменения это неизбежная часть нашей жизни.

В тексте автор ностальгирует по прошлым временам, но не оспаривает необходимость изменений. Он описывает свою преданность родным местам и одновременно понимание, что мир меняется и нельзя остаться в прошлом. Эта двойственность взгляда на проблему делает текст более интересным и выводит нас на мысль о том, что каждому индивидуально необходимо найти баланс между прошлым и будущим, между своими родными местами и изменяющимся миром.

Е. М. Северина,
доктор философских наук, профессор
Южный федеральный университет, Ростов-на-Дону, Россия
e-mail: emkovalenko@sfedu.ru
ORCID: 0000-0001-6518-2771

ПРОЕКТ CHEKHOV DIGITAL: ЦИФРОВОЙ ИНДЕКС ИМЁН И НАЗВАНИЙ

Аннотация. В настоящей работе анализируется специфика процесса разработки цифрового индекса имён и названий реальных людей и объектов, упоминаемых в текстах произведений и писем А. П. Чехова, которая осуществлена на основе указателей академического издания произведений писателя. Создание цифрового индекса позволяет организовать семантический поиск по текстам произведений и писем писателя, редакционно-критическому аппарату цифрового издания *Chekhov Digital*.

Ключевые слова: *цифровое издание, проект Chekhov Digital, семантический поиск, цифровой индекс, указатель имён и названий*

Проект *Chekhov Digital* – это цифровое издание полного собрания произведений Антона Павловича Чехова, которое было напечатано и опубликовано в период с 1974 по 1982 гг. (30 томов, объём – около 46 000 страниц). Цифровой проект реализуется в формате стандарта *TEI/XML*, включает семантическую разметку значимых сущностей в текстах произведений писателя и в редакционно-критических материалах академического издания, что создаёт новые возможности для академических исследований в цифровом формате и использования литературных текстов в цифровых проектах [3]. Работа над проектом в настоящее время продолжается и осуществляется Центром цифровых гуманитарных исследований Института филологии, журналистики и межкультурной коммуникации ЮФУ совместно с Международной лабораторией языковой конвергенции НИУ ВШЭ и лабораторией филологии ЮНЦ РАН.

Полное собрание сочинений и писем А. П. Чехова в 30 томах (далее – ПССиП) [6] представляет собой академическое издание текстов произведений и писем писателя, включая ранние редакции и различные варианты текста. Подготовка этого издания была связана со скрупулезной текстологической работой с сохранившимися источниками для исправления цензурных, типографских, редакторских искажений, накопившихся в предыдущих редакциях. Были изучены архивные фонды, журналы и газеты, в которых писатель мог публиковаться, в т. ч. с использованием неизвестных псевдонимов и/или анонимно, а также неопубликованные записные книжки и дневники, найденные рукописи писателя.

Издание состоит из двух серий: Сочинения (т. I–XVIII), которые составляют тексты произведений писателя, и Письма (т. I–XII) – эпистоляр-

ное наследие А. П. Чехова. Кроме того, каждый том содержит незавершённые произведения писателя, варианты текстов и их первоначальные редакции. Тексты издания печатаются «по правилам современной орфографии и пунктуации, с сохранением индивидуальных особенностей, свойственных языку Чехова» [1, с. 7].

Цифровое семантическое издание *Chekhov Digital* следует подходу, реализованному в рамках электронного научного издания творческого наследия А. П. Чехова (ЭНИ «ЧЕХОВ») Фундаментальной электронной библиотеки «Русская литература и фольклор» (ФЭБ) [7], в котором основным элементом представления текстов является отдельное произведение как «самодостаточный фрагмент печатного издания (например, рассказы «Лошадиная фамилия» или «Унтер Пришибеев», входящие в состав IV т. академического Полного собрания сочинений и писем Чехова в 30 т.)» [7].

Редакционно-критический аппарат академического издания [6] играет важную роль не только в контексте его традиционного функционального назначения, но и для разработки семантического поиска в цифровом издании. Аппарат включает в себя историко-литературные и текстологические комментарии к текстам произведений писателя, описание истории их создания, цензурирования, источников текста, а также тех изменений, которые были внесены в основной текст на основе рукописей и авторизованных печатных изданий. В текст примечаний включены отзывы прижизненной критики, мнения современников о чеховских произведениях, информация о переводах на иностранные языки [1, с. 7–8]. Для этих материалов также разрабатывается семантическая машиночитаемая разметка, что позволяет исследовать их с использованием компьютерных методов.

Академическое издание произведений и писем А. П. Чехова [6] представляет собой не только собрание канонических текстов писателя и редакционно-критических материалов, но и набор упоминаний в текстах реально существовавших людей, дат, событий, ситуаций и т. п. Поэтому издание содержит ряд составленных специалистами указателей (индексов), которые в доцифровую эпоху давали возможность извлекать эти данные из текстов, а сейчас эти материалы представляют интерес с точки зрения организации семантического поиска в цифровом издании. Речь идёт в первую очередь об указателях имён и названий, упомянутых в текстах писателя, комментариях/примечаниях, которые встречаются в XIV/XV, XVI, XVII т. серии Сочинения, сводном указателе имён и названий (XVIII т.) и указателях имён и названий к каждому из 12 томов серии Письма. На основе этих указателей разрабатывается база данных имён и названий, упоминаемых в текстах писателя и примечаниях/комментариях издания, в которой содержится не только оцифрованный вариант указателей печатного издания, т. е. перевод результатов традиционных филологических практик в цифровой формат, но и дополнительная информация из других баз данных, таких как *Wikidata* (URL: wikidata.org), а следовательно, данные приобретают новые специфические свойства. Разрабатываемая база данных имён и названий – цифровой указатель (индекс), наследует функциональ-

ность печатного указателя, содержащего информацию о томах и страницах упоминания соответствующего имени/названия в текстах произведений ПССиП с учётом связи имен и названий друг с другом, что позволяет реализовать семантический поиск по текстам произведений писателя и редакционно-критическому аппарату. Подготовленные в таком формате данные становятся инструментом не только для новых филологических практик, но и позволяют получить представление о работе с большими коллекциями литературных текстов.

Цифровой индекс имён и названий писем содержит 10 836 записей, включая 3 039 записей о названиях и 7 797 записей об именах реальных людей, из них 683 записи (10 названий) – общие с записями из указателя имён и названий Л. Н. Толстого [8] (меньше всего в 1904 г. (XII т. писем) – 16 записей, а больше всего в 1887–1888 гг. (II т. писем) – 104 записи). Указатель имён и названий сочинений содержит 5 177 записей о названиях и 8 934 записи об именах реальных людей, собранных автоматически из указателей XIV–XVIII т. Оцифрованный индекс расширяет возможности доцифрового традиционного поиска по указателю за счёт «нечёткого поиска» по части слова/имени, поиска через просмотр всего списка, перехода на страницу соответствующего произведения, письма или примечания/комментария. При этом появляется возможность связать не только конкретную страницу с соответствующей записью в БД, но и найти выражение соответствующей сущности на странице.

Кроме того, на основе данных из оцифрованного индекса может быть построен частотный список встречаемости каждой сущности на страницах текстов произведений и писем писателя, в комментариях и примечаниях, что позволяет исследовать появление одних и тех же имён и названий в разных контекстах. Одновременное появление различных имён и названий в одном абзаце, на одной странице или в любом другом более широком или узком контексте дает возможность выявить взаимодействие между разными фрагментами текстов писателя, фрагментами примечаний/комментариев, что позволяет получить представление о социальных связях писателя, причем как в контексте взаимодействия писателя с другими людьми, например, адресаты и герои его писем, так и в контексте точки зрения А. П. Чехова на связи реальных людей и объектов друг с другом. Такого рода структуры предоставляют новые возможности для изучения идей и интересов А. П. Чехова, общих тенденций и тем его творчества и эпистолярия. Для каждого имени/названия из указателя можно выявить наиболее значимые связанные с ним имена.

Цифровой индекс также позволяет понять, каким образом создавался редакционно-критический аппарат академического издания, влияние различных редакторов на его формирование, выявить неточности и различия в подходах как отдельных томов издания, так и серий – сочинений и писем. Например, каждый том писем снабжён своим указателем имён и названий людей и объектов, упоминаемых в письмах и комментариях, в то время как для серии Сочинения принят другой подход – общий указа-

тель представлен в XVIII т. Полного собрания сочинений, но тома, в которых представлена нехудожественная проза, содержат указатели с дополнительным описанием имён и названий из XVIII т., в котором чаще всего указана краткая информация о человеке или объекте, с сокращениями имён/названий (насколько возможно). Например, в указателе XVIII т. есть следующее описание брата писателя «Чехов Ал. П. – Чехов Ал. П. (псевдонимы – Агафопод Единицын; А. Ед.; Ч (?); А. Седой; Алоэ)» [4, с. 507], а в указателе XVII т. – более полное описание «Чехов (псевдонимы – Алоэ, Агафопод Единицын, А. Седой) Александр Павлович (1855–1913), брат Чехова, писатель, журналист; автор ряда воспоминаний о Чехове» [2, с. 514]. Возможно, это связано с объёмностью, представленной в указателе XVIII т. информации и предполагает самостоятельную работу исследователя по поиску необходимых данных. Цифровой указатель решает эту проблему, однако требует «ручной» верификации автоматически собранной информации.

Цифровой индекс проекта *Chekhov Digital* [5] сохраняет все свойства традиционного индекса и в то же время расширяет исследовательские возможности за счёт компьютерного управления информацией и семантической поисковой системы проекта.

Исследование выполнено в рамках соглашения о научном сотрудничестве между Южным федеральным университетом (ЮФУ) и Национальным исследовательским университетом «Высшая школа экономики» (НИУ ВШЭ) («Зеркальные лаборатории НИУ ВШЭ»), проект № 6.13.1-02/250821-1 «Конвергенция языковых пластов русского языка в зеркале цифровых решений».

Список источников

1. От редакции // Чехов А. П. Полное собрание сочинений и писем: в 30 т. Сочинения: в 18 т. М., 1974–82. Т. 1: Рассказы. Повести. Юморески, 1880–82. М.: Наука, 1974. С. 5–8.
2. Петрова Г. А. Указатель имён и названий / Г. А. Петрова, Е. М. Сахарова // Чехов А. П. Полное собрание сочинений и писем: в 30 т. Т. 17. М.: Наука, 1980. С. 466–519.
3. Северина Е. М. Цифровые филологические практики: проект «Chekhov Digital» / Е. М. Северина, А. А. Бонч-Осмоловская, А. М. Кудин // Актуальные проблемы филологии и педагогической лингвистики. 2022. № 2. С. 153–165. DOI: 10.29025/2079-6021-2022-2-153-165.
4. Указатель имён и названий: (т. I–XVIII) // Чехов А. П. Полное собрание сочинений и писем: в 30 т. Т. 18. М.: Наука, 1982. С. 361–523.
5. Цифровой проект «Chekhov Digital». URL: chekhov-digital.sfedu.ru.
6. Чехов А. П. Полное собрание сочинений и писем: в 30 т. / АН СССР. ИМЛИ им. А. М. Горького. М.: Наука, 1974–83. URL: feb-web.ru/feb/chekhov/default.asp?/feb/chekhov/texts/che-te02.html.
7. ЭНИ «Чехов» / Русская литература и фольклор: фундаментальная эл. библиотечка. URL: feb-web.ru/feb/chekhov/default.asp.
8. Orekhov B. The 91st Volume – How the Digitised Index for the Collected Works of Leo Tolstoy Adds a New Angle for Research / B. Orekhov, F. Fischer // Digital Humanities 2018: Book of Abstracts / Libro de Resúmenes. Mexico: Red de Humanidades Digitales A. C., 2018. Pp. 465–466.

В. А. Баранов,
доктор филологических наук, профессор
Ижевский государственный технический университет им. М. Т. Калашникова,
Ижевск, Россия
Казанский федеральный университет, Казань, Россия
e-mail: victor.a.baranov@gmail.com
ORCID: 0000-0003-1730-6359

ДИСТРИБУЦИЯ И СЕМАНТИКА СЛАВЯНСКОГО СЛОВА В ИСТОРИЧЕСКОМ КОРПУСЕ

Аннотация. Представлены результаты автоматического анализа семантики двух слов – «звѣрь» и «убогъ» – в корпусе средневековых славянских рукописей X–XV вв. «Манускрипт» (URL: manuscripts.ru). Параметры и интерфейсы интернет-модуля дистрибутивного словаря позволили выявить неоднородность семантики анализируемых слов, сделать вывод об отличиях в её контекстуальной реализации в разножанровых подкорпусах.

Ключевые слова: исторический корпус, славянские средневековые рукописи, дистрибутивный метод, зверь, убогий

Одним из наиболее эффективных методов количественно-статистического автоматического анализа текстовых данных является дистрибутивный метод, предполагающий использование количественных характеристик как самой анализируемой единицы, так и окружающих её единиц. С помощью метода успешно решаются задачи автоматического поиска статистически устойчивых сочетаний – коллокаций и коллигаций, извлечения ключевых словосочетаний, выявления тематически и семантически близких языковых единиц, снятия лексической омонимии и мн. др. – как в коммерческих, так и в исследовательских целях. Поэтому в настоящее время практически все многофункциональные компьютерные системы, программы и библиотеки, предназначенные для автоматической обработки текстов, имеют средства извлечения, обработки, количественно-статистической оценки и визуализации синтагматических и парадигматических связей слов документов.

Теоретической базой исследования семантики лингвистических единиц с помощью дистрибутивного метода является дистрибутивная гипотеза о близости значений слов при их встречаемости в схожих контекстах (см., например, [3]). Предложено несколько методик анализа дистрибуции слов, в основе которых лежат количественные или статические сведения о совместном употреблении лексических единиц, позволяющие представить каждую как набор числовых значений (вектор) и сопоставить векторы между собой с целью выявления наиболее близких [4].

Значительные успехи в области автоматического анализа лексической семантики достигнуты при исследовании современных текстов. По мере увеличения объёма машиночитаемых ресурсов, содержащих исторические документы, расширяется текстовая база для изучения языка в его истории компьютерными методами.

В рамках проекта «Манускрипт: славянское письменное наследие» несколько лет назад был создан дистрибутивный словарь, предназначенный для извлечения из корпуса лексических единиц, близких по семантике. Для вычисления совместной встречаемости слов и построения их векторов использован метод *k-skip-n-gram*, а для установления близости между векторами – величина косинусного расстояния между ними.

Интерфейс словаря имеет два варианта – простой (URL: manuscripts.ru/mns/mns_evp.vec.simple) и расширенный (URL: manuscripts.ru/mns/mns_evp.vec.main)³⁵. В отличие от простого режима, в котором предлагается ввести маску анализируемого слова для получения списка семантических аналогов и визуализации связей в виде графа, расширенный имеет ряд параметров, которые позволяют изменить значения запроса: выбрать подкорпус, тип лингвистической единицы, указать количественную или статистическую меру вычисления совместной встречаемости слов, величину косинусного расстояния между векторами анализируемого слова и его семантическими аналогами, а также между последними, выбрать вариант раскладки графа и др. [1; 2].

Несмотря на относительно небольшой объём исторического корпуса (чуть более 3 млн текстовых форм), результаты автоматического анализа семантики слов, содержащихся в рукописях X–XV вв. корпуса, достаточно показательны, дают ожидаемые результаты, соотносимые с результатами автоматического анализа семантики слов современного русского языка, а в некоторых случаях, не соответствуя последним, позволяют увидеть специфику парадигматических отношений слов в языке средневековых славянских рукописных памятников. Проиллюстрируем сказанное примерами.

«Звѣрь» (параметры извлечения: весь корпус, количество аналогов – 15, статистика совместной встречаемости – *Log-Likelihood*). Наиболее близкими в историческом корпусе аналогами этого слова являются слова: «гадь» (0,14305), «пѣтица» (0,11892), «скоть» (0,11788), «земла» (0,10191), «христось» (0,09790), «грѣхъ» (0,09745), «плѣть» (0,09690), «страсть» (0,09571), «ровь» (0,09301), «пропасть» (0,09272), «образъ» (0,09180), «богъ» (0,09082), «змия» (0,08985), «доуша» (0,08965), «тѣло» (0,08887). Если близость «звѣрь» и первых трёх, а также «змия» понятна³⁶ – слова находятся в гипо-гиперонимических отношениях, то сближение с другими должно быть объяснено. Увидеть причину позволяет граф (минимальное косинусное расстояние – 0,08), который выявляет по крайней мере три группы близких по дистрибуции слов: 1) «гадь», «пѣтица», «скоть»,

³⁵ Процедуры и программы словаря, а также его веб-интерфейс созданы Р. М. Гнутиковым.

³⁶ Ср. аналоги этого же слова, например, в «Serelex: Поиск семантически связанных слов» (URL: serelex.cental.be/ru/ru-skipgram-librusec) или на других аналогичных ресурсах.

2) «грѣхъ», «тѣло», «богъ», «христось» и др. (с общим словом «земля»),
 3) «пропасть»; «ровъ»; «змия» (см. рис. 1).

Группировка аналогов указывает на неоднородность семантики анализируемого слова «звѣрь» в текстах X–XV вв.: 1) животное; 2) то, что греховно, телесно, плотско, противопоставлено божественному, душевному; 3) грехопадение(?).

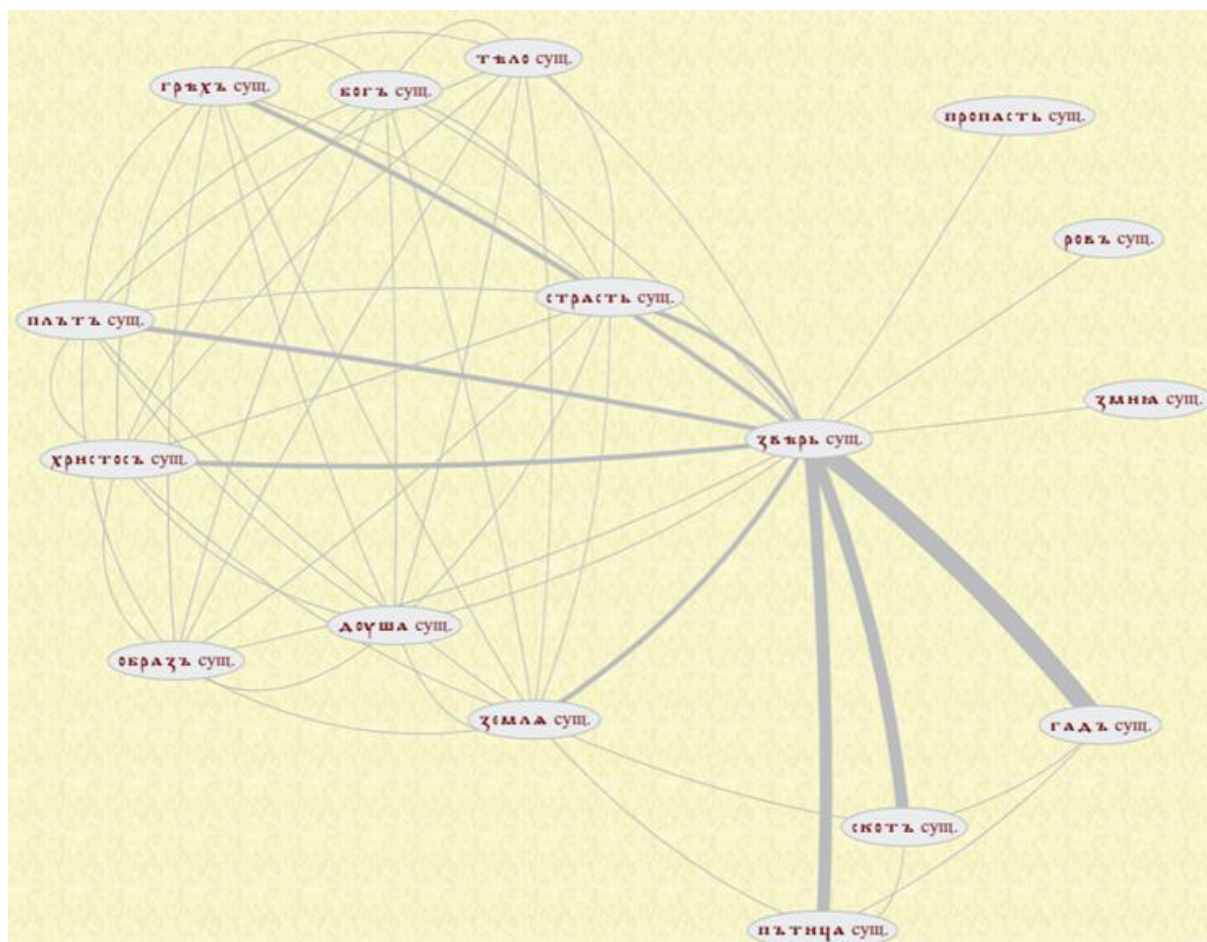


Рис. 1. Граф семантических аналогов слова «зверь»

Наиболее отчётливо значение, связанное с противопоставлением греховного и божественного представлено в новозаветных текстах Евангелий (для извлечения аналогов использован соответствующий подкорпус), в которых у слова «звѣрь» лишь два семантически близких слова – «сотона» и «ангелъ». На графе, построенном на основе выборки из ветхозаветных текстов Паримейника, также видны несколько подграфов (наименьшее косинусное расстояние – 0,15), наиболее отчётливые – «ковъчегъ», «пѣтица», «гадъ», «скоть», «родъ» и «человѣкъ», «земля», «богъ». В текстах Паренесиса Ефрема Сирина неоднородность семантического поля слова «звѣрь» на графе (наименьшее косинусное расстояние – 0,05) представлена как «животное» (подграф «пѣтица», «гадъ»), «испытание(?)» («гора», «поустыня», «градъ», «иосифъ»), «страх(?)» («образъ», «по-

мысль», «страхъ», «трепетъ»). Ни в старославянском словаре³⁷, ни в словаре И. И. Срезневского³⁸, ни в Древнерусском словаре XI–XIV вв.³⁹ иных значений, кроме «животное», нет. В то же время семантическое и тематическое разнообразие семантических аналогов, их группировка свидетельствуют о наличии у слова «звѣрь» не только дискретной семантики в старославянском и древнерусском языках, но и о том, что слово имеет частные, контекстно зависимые значения в отдельных текстах. Исходное значение находим в Полном церковнославянском словаре: «...“Звѣрь” в Апок., XIII гл. и др. означает антихриста...: иногда “звѣрь” значит вообще человека развращённого умомъ, и нравомъ безчеловечнаго, подобнаго дикимъ звѣрямъ...»⁴⁰. В современном русском языке реализуется обнаруженная в текстах X–XV вв., свойственная церковнославянскому языку, дискретность значения: «ЗВЕРЬ, м. 1. Дикое, обычно хищное животное. 2. перен. Чрезвычайно жестокий, бессердечный, грубый человек. ...», а также: «ЗВЕРИНЫЙ, прил. 1. Соотносящийся по знач. с сущ.: зверь (1), связанный с ним. 2. Свойственный зверю (1), характерный для него. // перен. Жестокий, дикий, свирепый. ...»; ЗВЕРСКИЙ прил. 1. Свойственный зверю (1), характерный для него. // перен. Крайне жестокий, свирепый. // Вызывающий чувство страха. ...»⁴¹. Ср. перечень похожих слов в «Паспорте слова» Национального корпуса русского языка: звериный – волчий, дикарский, животный, змеиный, кошачий, овечий, скотский, тигриный, человеческий, шакалий; зверский – бесчеловечный, гнусный, жестокий, звероподобный, зверский, злобный, кровожадный, наглый, отвратительный, садистский⁴².

«Оубогъ» (параметры извлечения: весь корпус, количество аналогов – 15, статистика совместной встречаемости – *Pointwise Mutual Information*). Перечень извлечённых семантических аналогов: «нищъ» (0,32191), «божеи» (0,27711), «благъ» (0,27166), «зълъ» (0,26704), «правдънь» (0,26503), «богатъ» (0,26425), «мъртвъ» (0,26381), «вѣрьнь» (0,25880), «мъногъ» (0,25747), «простъ» (0,25445), «великъ» (0,24953), «страшнь» (0,24872), «живъ» (0,24676), «сватъ» (0,24532), «небеснь» (0,24428). Построение графа слова «оубогъ» при минимальном косинусном расстоянии между его семантическими аналогами, равном 0,34, демонстрирует два семантических подполя, соотносимые со словами «нищъ», «богатъ», «простъ», которые остаются вне подграфа, и всеми остальными, объединёнными в один подграф, см. рис. 2.

³⁷ См., например, Цифровой словарь старославянского языка (URL: gorazd.org/gulliver/?envLang=ru).

³⁸ Материалы для словаря древнерусского языка. СПб., 1893. С. 966. URL: etymolog.ruslang.ru/doc/sreznevskijE-K.pdf.

³⁹ Словарь древнерусского языка (XI–XIV вв.): в 10 т. / АН СССР, Ин-т рус. яз.; гл. ред. Р. И. Аванесов. Т. 3: добродѣтельно – изгечиса. М.: Рус. яз., 1990. С. 365–366.

⁴⁰ Дьяченко Г. Полный церковнославянский словарь. 1899. С. 199.

⁴¹ Новый словарь русского языка. Толково-словообразовательный. М.: Рус. яз., 2000. Т. 1: А–О. 1 209 с.

⁴² URL: ruscorpora.ru/word/main.

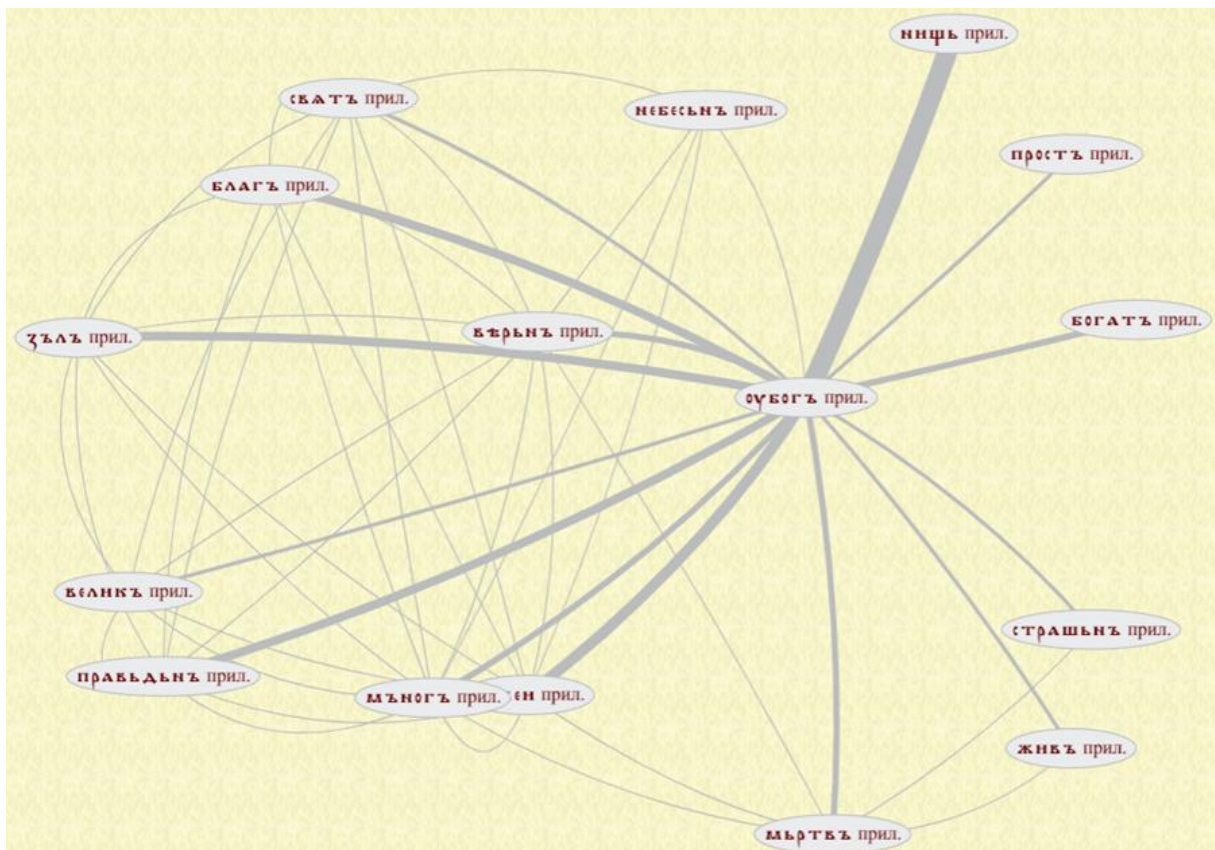


Рис. 2. Граф семантических аналогов слова «оубогъ»

Это свидетельствует по крайней мере о двух значениях слова «оубогъ». Действительно, в старославянском словаре, представленном на портале проекта «Гораздъ», слово имеет два значения – «бедный, нищий, убогий» и «убогий, бедный, ничтожный, униженный»⁴³. В глаголическом подкорпусе ряд из восьми семантических аналогов отчётливо демонстрирует второе значение слова: «правдънъ» (0,30894), «срамънъ» (0,25272), «нищъ» (0,23987), «неповинънъ» (0,23644), «оумилень» (0,20078), «правъ» (0,17337), «сильнъ» (0,14830), «сѣмѣренъ» (0,09691). В подкорпусе русских летописей при минимальном значении косинусного расстояния 0,01 аналоги «нищъ», «богатъ», «хоудии», «работънъ» (рабский, подчинённый), с одной стороны, и «сѣмирень», «странънъ», «лють», «правъ», «възлюбленъ», «дивънъ», «чърноризьскъ» – с другой, в графе противопоставлены друг другу, что демонстрирует дискретность значения «бедный, неимущий; нищий» и «лишённый; ничтожный, недостойный», фиксируемую, например, и И. И. Срезневским⁴⁴. Дискретность значения слова «убогий» сохраняется и до настоящего времени, что демонстрируется группами семантических аналогов в современном русском языке, представленных на корпусных интернет-ресурсах: «нищий», «сирый», «жалкий», «немошный» и «примитивный», «дебильный», «унылый», «тупой» и др.⁴⁵, – рядами похожих слов, например, в Национальном корпусе рус-

⁴³ URL: gorazd.org/gulliver/?envLang=ru.

⁴⁴ Материалы для словаря древнерусского языка. С. 1113–1114. URL: etymolog.ruslang.ru/doc/sreznevskijT-Ja.pdf.

⁴⁵ См., например, AdaGram (URL: adagram.ll-cl.org/about).

ского языка: «нищий», «бедный», «нищенский»; «сирый», «скудный», «неказистый», «неопрятный», «жалкий»; «заброшенный», «неуютный»⁴⁶, и синонимов на ресурсе «Карта слов»: «бедный», «сирый»; «жалкий», «неказистый», «невзрачный», «унылый», «неприглядный», «непривлекательный»; «несчастный», «пришибленный», «некрасивый», «ущербный», «испорченный», «бессильный», «извращённый», «искажённый» и др.⁴⁷, – а также в словарях современного русского языка: «УБОГИЙ¹ м. 1. Калека. 2. перен. Тот, кто пребывает в крайней бедности, в нищете; бедняк, нищий. УБОГИЙ² прил. 1. Немощный, увечный, искалеченный. // Жалкий, несчастный на вид. 2. перен. Пребывающий в крайней бедности, в нищете. // Свидетельствующий о крайней бедности, нищете; нищенский. 3. перен. Духовно ограниченный, крайне посредственный (о человеке). // Бедный по содержанию»⁴⁸.

Таким образом, автоматический анализ лексической семантики в средневековых славянских текстах результативен в нескольких отношениях: подтверждает имеющиеся историко-лингвистические сведения о слове, позволяет выявить незафиксированные парадигматические связи, а соответственно – неоднородность семантического поля исследуемой лексемы, показать реализацию семантики слова в различных текстологически однородных подкорпусах текстов. Сопоставление полученных сведений о средневековом слове с информацией о его функционировании и лексикографическими характеристиками в современном русском языке позволяет увидеть произошедшие в семантике слова изменения.

Работа выполнена при финансовой поддержке Российского научного фонда (РНФ), проект № 20-18-00206.

Список источников

1. Баранов В. А. Дистрибутивный словарь исторического корпуса «Манускрипт»: постановка задачи, материал, методы / В. А. Баранов // Актуальные проблемы филологии и педагогической лингвистики. 2022. № 2. Прикладная лингвистика: современные ракурсы и перспективы. С. 94–106. DOI: 10.29025/2079-6021-2022-2-94-106.
2. Баранов В. А. Дистрибутивный словарь славянских текстов X–XV вв.: параметры запросов и визуализация результатов / В. А. Баранов, Р. М. Гнутиков // Современная лингвистика: от теории к практике: матер. III Междунар. лингв. саммита (Казань, 2022) в 3 т. / под общ. ред. И. Э. Ярмакеева, Ф. Х. Тарасовой. Казань: КФУ, 2023. Т. 1. С. 99–103.
3. Sahlgren M. The Distributional Hypothesis. From Context to Meaning / M. Sahlgren // Rivista di Linguistica. 2008. Vol. 20. No. 1. Pp. 33–53.
4. Mikolov T. Word2vec: Tool for Computing Continuous Distributed Representations of Words / T. Mikolov, K. Chen, G. Corrado et al. 2013. URL: code.google.com/archive/p/word2vec/.

⁴⁶ URL: ruscorpora.ru/word/main.

⁴⁷ URL: kartaslov.ru.

⁴⁸ Новый словарь русского языка. Толково-словообразовательный. М.: Рус. яз., 2000. Т. 2: П–Я. 1 088 с.

А. Н. Михайлова,

магистрант

Национальный исследовательский университет «Высшая школа экономики»,

Пермь, Россия

e-mail: anast.mih@yandex.ru

ORCID: 0000-0001-9067-1622

АТТРИБУЦИЯ МЕМУАРОВ Г. К. ЖУКОВА И К. К. РОКОССОВСКОГО С ПОМОЩЬЮ МЕТОДА DELTA

Аннотация. Приведены результаты исследования авторства мемуаров советских полководцев Г. К. Жукова и К. К. Рокоссовского. Для атрибуции текстов использован цифровой метод *Delta*. В результате работы впервые с высокой степенью достоверности определено соавторство в мемуарах Г. К. Жукова, выделены главы, написанные непосредственно им самим. Доказано, что издание 1990 г. не имеет существенных отличий по соавторству от первого издания. Подтверждено, что дополнения к воспоминаниям К. К. Рокоссовского, опубликованные в 1997 г. его наследниками, соответствуют тому же автору, что и прижизненное издание.

Ключевые слова: *Жуков, Рокоссовский, авторство, мемуары, стилеметрия, Delta*

На период 1950–1960 гг. в Советском Союзе приходится очередной «мемуарный взрыв». Появляется общественный запрос на воспоминания крупнейших политических и военных деятелей. В результате этого в 1960-х гг. советские полководцы активно публикуют мемуары о Великой Отечественной войне. Не секрет, что эти издания подверглись значительному влиянию со стороны цензуры и военных чиновников. Авторство многих из них крайне спорно.

Наиболее скандальную славу получили «Воспоминания и размышления» Георгия Константиновича Жукова. Книга издана в 1969 г., вышла в 30 странах на 18 языках, пережила 15 изданий. При этом часто каждое следующее издание имело ряд заметных отличий от предыдущего. Отдельный интерес в этом отношении вызывает десятое издание, опубликованное в 1990–1992 гг. Подготовкой издания занимались дочь маршала, Жукова Мария Георгиевна, и его бессменный редактор, Анна Давыдовна Миркина. По их словам, книга была восстановлена по рукописи Жукова, но читатели отнеслись к ней скептически.

Внуки Константина Константиновича Рокоссовского также в 1997 г. опубликовали дополненные мемуары. И надо понимать, что негативная репутация мемуаров его коллеги Г. К. Жукова, как и советских полководцев в целом, интерполировалось и на мемуары Константина Константиновича в полной мере.

В рамках проведённого исследования были поставлены две задачи:

- 1) определить, написаны ли изданные в 1960-х гг. мемуары Г. К. Жукова и К. К. Рокоссовского маршалами единолично либо в соавторстве;
- 2) проверить, являются ли Жуков и Рокоссовский авторами изменений, внесённых в их мемуары в 1990 и 1997 гг.

Классическими гуманитарными методами ответить на эти вопросы не представляется возможным, а цифровыми ранее этот вопрос не решался.

В качестве методологической основы исследования использован стилеметрический анализ при помощи метода *Delta*. Он опирается на анализ наиболее частотных слов в тексте и уже хорошо зарекомендовал себя во многих литературоведческих исследованиях [4; 28]. В качестве инструмента использован пакет *Stylo* для *R* [26].

Корпус текстов для исследования составили воспоминания маршалов Жукова по изданиям 1969 г. и 2002 г. (по изданию 1990 г.) [11; 12], воспоминания маршала Рокоссовского по изданиям 1968 г. [22], выделенные в отдельный текст фрагменты, изъятые цензурой и опубликованные в 1997 г. [23], а также мемуары военачальников: А. П. Белобородова [2], А. В. Горбатова [6], П. И. Батова [1], маршалов И. С. Конева [17], А. И. Еременко [7–9], К. А. Мерецкова [15] и А. М. Василевского [3]. Дополнительно в корпус были добавлены тексты статей Г. К. Жукова и К. К. Рокоссовского, опубликованные в «Военно-историческом журнале» в 1965–1968 гг. [10; 13; 14; 18–21].

В качестве примера эталона стиля Г. К. Жукова принят оригинальный текст главы «Битва за Москву», опубликованный Мосархивом по рукописи маршала [5].

Результаты применения метода *Delta* представлены на дендрограмме расстояний для текстов мемуаров маршалов (рис. 1). Заметна значительная корреляция эталонного авторского текста Г. К. Жукова и его статей 1966 и 1967 гг. При этом непосредственно мемуары значительно удалены от них и очень близки между собой. Учитывая тот факт, что метод *Delta* может давать искажённый результат на текстах, существенно различающихся по объёму, отметим только, что все рассматриваемые статьи Г. К. Жукова, опубликованные в сборнике «Военно-исторический журнал», показали значительную близость с примером авторского стиля (сопоставимым по объёму).

Объединённое дерево, построенное на результатах десяти экспериментов, дало аналогичный результат.

Значимым результатом стал тот факт, что показали сильную стилеметрическую близость обе части мемуаров К. К. Рокоссовского. Такой результат позволяет считать, что мемуары, опубликованные маршалом в 1968 г., и дополнения, исключённые по требованию цензуры и опубликованные его наследниками в 1997 г., являются фрагментами одного текста.

Любопытно, что обе части мемуаров К. К. Рокоссовского определяются максимально близко, а статьи составляют два разных блока. Анализ текстов статей показал, что статья 1965 г. значительно отличается по фор-

ме (интервью), претерпела значительную обработку редакторами журнала. Статьи 1966 и 1968 гг. показывают значительную близость между собой.

Для того чтобы исключить искажения результатов, вызванные разницей в объёме текстов, тексты мемуаров Г. К. Жукова и К. К. Рокоссовского были разбиты на фрагменты от 5 до 15 тыс. знаков.

Помимо этого, для второй части исследования была добавлена статья с воспоминаниями А. Д. Миркиной [16], послужившая примером авторского стиля редактора мемуаров Г. К. Жукова.

К полученному корпусу глав мемуаров маршала Г. К. Жукова 1969 г. была добавлена гл. 11 «Ставка Верховного Главнокомандования» из издания 2002 г. Впервые эта глава появилась во 2-м издании мемуаров, и, по воспоминаниям академика А. Н. Яковлева [25], была добавлена под давлением военных чиновников. По этим же воспоминаниям, автором данной статьи был не Г. К. Жуков, а сотрудник АПН Вадим Комолов. В процессе подготовки настоящего исследования не было найдено примера его авторского стиля в необходимом объёме, однако работа в этом направлении ведётся.

Результаты применения метода *Delta* для текстов мемуаров маршалов Г. К. Жукова и К. К. Рокоссовского представлены на соответствующих дендрограммах (рис. 2 и 3).

Среди текстов Г. К. Жукова видим разделение на три большие кластера.

1. В один вошёл авторский текст маршала и все его статьи.
2. Во второй – статья А. Д. Миркиной.
3. В третий выделилась группа текстов, к которой *Delta* относит и главу «Ставка верховного главнокомандования».

Чётко прослеживается тематическое деление текстов.

Так, к стилю А. Д. Миркиной ближе всего первые две главы мемуаров, в которых описано детство и первые годы службы будущего маршала. Сама Анна Давыдовна уже в собственных мемуарах [16] неоднократно отмечает тот факт, что её тётка родом из той же деревни, что и маршал, делится воспоминаниями о предводителе всех мальчишек Егорке, ставшем потом легендарным Георгием Жуковым.

Авторскому стилю Г. К. Жукова ближе всего главы, посвящённые непосредственно боевым действиям. Туда можно отнести и воспоминания о начале войны, ликвидации Ельнинского выступа противника, борьбе за Ленинград и Москву, Сталинград, Берлинская операция и др. В своих воспоминаниях о времени написания мемуаров Анна Давыдовна приписывает маршалу такие слова: «...я же буду писать о главном – о событиях военных, крупнейших операциях Отечественной войны и их осуществлении» [16]. Полученный результат удивительным образом соответствует им.

В третью группу текстов вошли все главы, неоднозначные и сложные с политической точки зрения. Тут описаны события накануне Великой Отечественной войны, до сих пор являющиеся предметом дискуссий, капитуляция Германии и Потсдамская конференция, события гражданской

войны, роль Ставки Верховного Главнокомандования, добавленная уже во 2-м издании, и события на Халхин-Голе, которые согласно ряду воспоминаний, Г. К. Жуков описывать не хотел. Учитывая полученный результат, можно говорить о том, что и не писал.

В дополнение к указанному исследованию текст мемуаров Г. К. Жукова был проанализирован при помощи функции *rolling.classify()*, представленной в пакете *Stylo* (рис. 4). В цифровой текстологии она используется для выделения динамического авторского следа для работ, написанных в соавторстве. Получен результат, для предположения о трёх соавторах в мемуарах «Воспоминания и размышления» (издание 1969 г.) (Г. К. Жуков, А. Д. Миркина и неизвестный третий автор). Определены зоны, в которых авторский стиль Г. К. Жукова уверенно преобладает и практически не смешивается со стилем других соавторов. При этом зоны авторского стиля А. Д. Миркиной и третьего неизвестного соавтора часто пересекаются, что отражает активную совместную работу над рядом глав.

Такой результат позволяет считать, что мемуары Г. К. Жукова «Воспоминания и размышления» написаны в соавторстве с двумя другими авторами, одним из которых выступал официальный редактор книги А. Д. Миркина.

К сожалению, пример чистого авторского стиля К. К. Рокоссовского не выявлен. Однако, учитывая результаты анализа авторства статей Г. К. Жукова, опубликованных в «Военно-историческом журнале», можно предположить, что статьи К. К. Рокоссовского 1966 и 1968 гг. в этом же издании, показавшие общий авторский стиль, написаны преимущественно самим маршалом.

Результаты применения метода *Delta* для текстов К. К. Рокоссовского, представленные на дендрограмме (рис. 3), отражают деление текстов на три кластера:

- 1) главы, относящиеся к первой части книги (1–9) и статья, написанная в 1966 г.;
- 2) главы, относящиеся ко второй части книги (10–25) и статья, написанная в 1968 г.;
- 3) статья 1965 г., подвергшаяся значительному вмешательству редактора журнала и демонстрирующая иной авторский стиль.

Для выявления глубоких текстовых связей в цифровой текстологии используются методы сетевого анализа [27], исходя из предположения, что конкретные тексты могут быть представлены как узлы сети, а их явные отношения – как связи между этими узлами. Был выбран алгоритм укладки *ForceAtlas2*, встроенный в *GEPHI*, инструмент с открытым исходным кодом для управления сетью и визуализации.

Сетевая визуализация связи авторского стиля рассматриваемых текстов К. К. Рокоссовского и произведённый кластерный анализ (рис. 5) отразили деление текстов на три последовательных кластера, каждый из которых соответствует определённому временному периоду. Так, к первому кластеру были отнесены гл. 1–9 и статья 1966 г., ко второму – гл. 10–21 (за

исключением текста гл. 18–19) и статья 1968 г. Третий кластер составили гл. 18–19 и 22–25, а также статья 1965 г., показавшая сильную связь с гл. 18–19.

Если в примере с мемуарами маршала Г. К. Жукова полученные группы текстов имеют чёткий тематический принцип деления, то в текстах мемуаров К. К. Рокоссовского такого деления нет. В этом случае отражение нашло изменение стиля повествования с течением времени, либо смена редактора в период между 1966 и 1968 гг. Тем не менее все фрагменты книги сильно связаны между собой и отражают картину единого произведения, написанного преимущественно одним автором, с учётом допустимой доли вмешательства редактора.

В последней части исследования рассмотрены издания мемуаров Г. К. Жукова 1969 и 1990 гг. На самом ли деле текст мемуаров стал ближе к авторскому стилю Жукова?

Для ответа на этот вопрос была использована функция *classify()*, предложенная пакетом *Stylo* для *R*, содержащая в своей основе методы машинного обучения. Для каждой из глав двух изданий определён наиболее вероятный автор, произведено сравнение. Существенных изменений авторства в тексте не выявлено, за исключением трех глав, а именно:

- 1) гл. 10 «Начало войны» стала ближе к стилю Г. К. Жукова;
- 2) гл. 5 и 6 «Кавалерия РККА» и «Командование 3 и 6 корпусами» стали ближе к стилю А. Д. Миркиной.

В целом говорить о заметном смещении близости стиля мемуаров к авторскому стилю Г. К. Жукова оснований нет.

В результате проведённой работы можно сделать следующие выводы.

1. Дополнения к мемуарам К. К. Рокоссовского, впервые опубликованные наследниками маршала в издании 1997 г., написаны тем же лицом, что и первое их издание 1968 г.

2. Анализ текста мемуаров К. К. Рокоссовского «Солдатский долг» показал отсутствие явных соавторов. Работа с высокой долей вероятности написана одним автором.

3. В мемуарах Г. К. Жукова «Воспоминания и размышления» выявлено три разных авторских стиля, один из которых соответствует стилю самого Г. К. Жукова, второй – А. Д. Миркиной. Третий автор на сегодняшний день не известен.

4. Определены главы, с высокой долей вероятности написанные Г. К. Жуковым собственноручно, с минимальным вмешательством редакторов.

5. Издания мемуаров Г. К. Жукова «Воспоминания и размышления» 1969 и 2002 гг. не имеют значительных различий в авторском стиле. Говорить о заметном смещении авторского стиля к стилю Г. К. Жукова можно только в отношении гл. 10 «Начало войны».

Список источников

1. Батов П. И. В походах и боях / П. И. Батов. М.: Воениздат, 1974. С. 528.
2. Белобородов А. П. Всегда в бою / А. П. Белобородов. М., 1979. С. 398.
3. Василевский А. М. Дело всей жизни / А. М. Василевский. М.: Вече, 2014. С. 608.
4. Великанова Н. П. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» / Н. П. Великанова, Б. В. Орехов // Мир Шолохова. 2019. № 1. С. 70–82.
5. Г. К. Жуков в битве под Москвой: сб. док. / отв. сост. А. А. Гуров; ИВИМО РФ. М.: Мосгорархив, 1994. С. 210.
6. Горбатов А. В. Годы и войны / А. В. Горбатов. М.: Воениздат, 1989. С. 365.
7. Еременко А. И. В начале войны / А. И. Еременко. М.: Наука, 1964. С. 511.
8. Еременко А. И. Годы возмездия. 1943–1945 / А. И. Еременко. М.: Финансы и статистика, 1985. С. 424.
9. Еременко А. И. Сталинград / А. И. Еременко. М.: Воениздат, 1961. С. 504.
10. Жуков Г. К. В битве за столицу / Г. К. Жуков // Военно-исторический журнал. 1966. № 9. С. 55–65.
11. Жуков Г. К. Воспоминания и размышления / Г. К. Жуков. М.: АПН, 1969. С. 735.
12. Жуков Г. К. Воспоминания и размышления: в 2 т. / Г. К. Жуков. М.: Олма-Пресс, 2002.
13. Жуков Г. К. На Курской дуге / Г. К. Жуков // Военно-исторический журнал. 1967. № 8. С. 69–83.
14. Жуков Г. К. На Курской дуге (продолжение) / Г. К. Жуков // Военно-исторический журнал. 1967. № 9. С. 82–97.
15. Мерецков К. А. На службе народу / К. А. Мерецков. М.: Политиздат, 1968. С. 446.
16. Миркина А. Д. Не склонив головы / А. Д. Миркина // Маршал Жуков: полководец и человек: сб. в 2 т. М.: АПН, 1988. Т. 2. С. 47–78.
17. Конев И. С. Записки командующего фронтом / И. С. Конев. М.: Наука, 1972. С. 368.
18. Рокоссовский К. К. На берлинском и восточно-прусском направлениях / К. К. Рокоссовский // Военно-исторический журнал. 1965. № 2. С. 25–28.
19. Рокоссовский К. К. На волоколамском направлении / К. К. Рокоссовский // Военно-исторический журнал. 1966. № 11. С. 46–55.
20. Рокоссовский К. К. Победа на Волге / К. К. Рокоссовский // Военно-исторический журнал. 1968. № 2. С. 64–75.
21. Рокоссовский К. К. Севернее Берлина / К. К. Рокоссовский // Военно-исторический журнал. 1965. № 5. С. 36–41.
22. Рокоссовский К. К. Солдатский долг / К. К. Рокоссовский. М.: Воениздат, 1968. С. 384.
23. Рокоссовский К. К. Солдатский долг / К. К. Рокоссовский. М.: Межрегиональный фонд «Выдающиеся полководцы и флотоводцы Великой Отечественной войны 1941–1945 гг.», Воениздат, 1997. С. 479.
24. Рокоссовский К. К. Солдатский долг / К. К. Рокоссовский. М.: Воениздат, 2019. С. 448.
25. Яковлев А. Н. Сумерки / А. Н. Яковлев. М.: Материк, 2005. С. 672.
26. Eder M. Stylometry with R: a Package for Computational Text Analysis / M. Eder, J. Rybicki, M. Kestemont // The R Journal. 2016. No. 8. Pp. 107–121. DOI: 10.32614/RJ-2016-007.

27. Eder M. Visualization in Stylometry: Cluster Analysis Using Networks / M. Eder // Digital Scholarship in the Humanities. 2017. Vol. 32. No. 1. Pp. 50–64. DOI: 10.1093/llc/fqv061.

28. Burrows J. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship / J. Burrows // Literary and Linguistic Computing. 2002. Vol. 17. No. 3. Pp. 267–287. DOI: 10.1093/llc/17.3.267.

Приложение

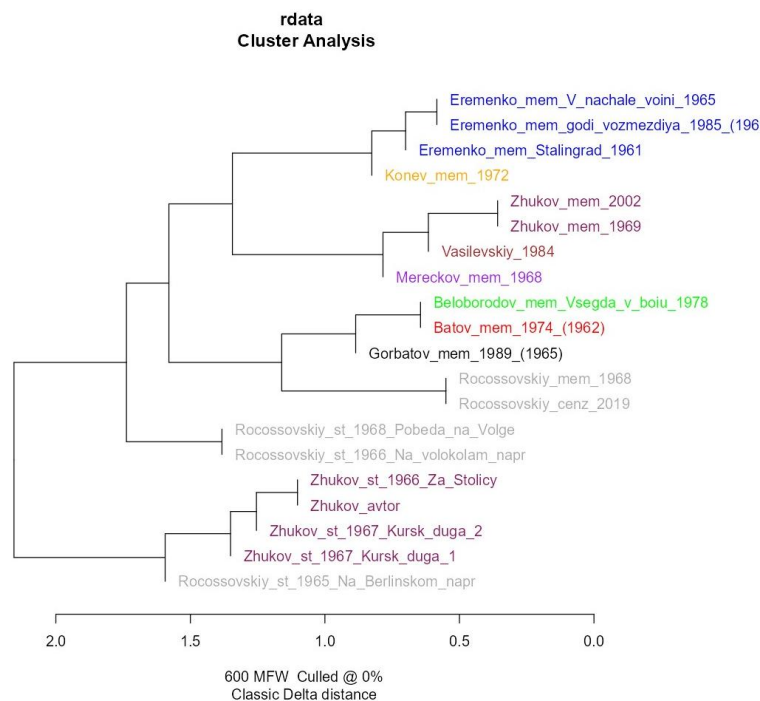


Рис. 1. Дендрограмма расстояний для текстов мемуаров советских полководцев

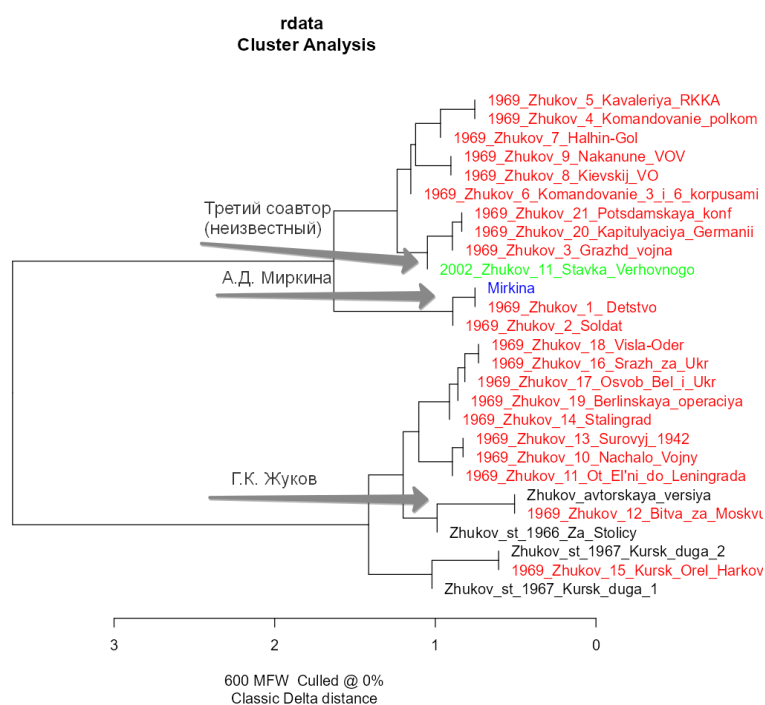


Рис. 2. Дендрограмма расстояний для текстов маршала Г. К. Жукова

rdata
Cluster Analysis

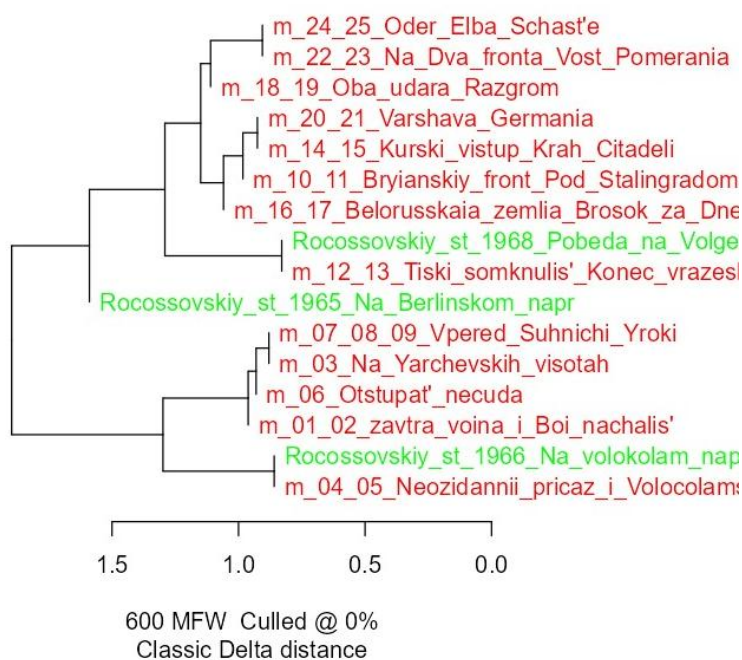


Рис. 3. Дендрограмма расстояний для текстов маршала К. К. Рокоссовского

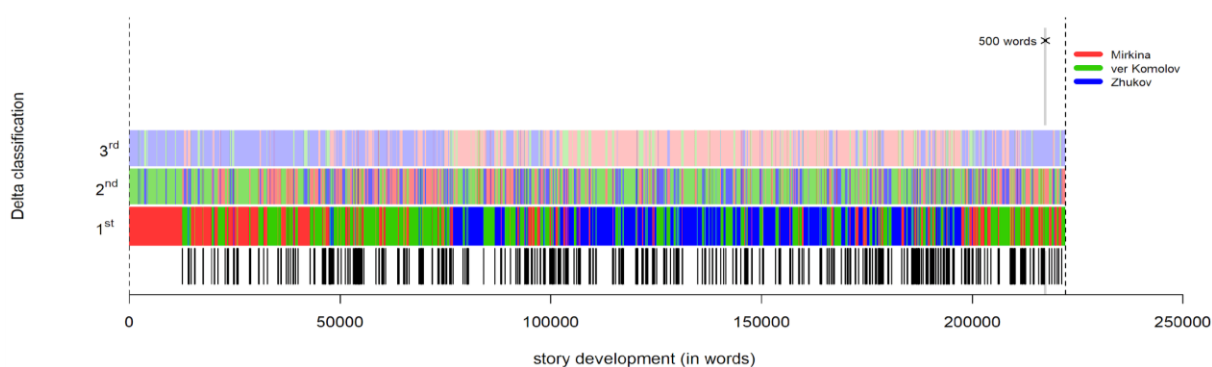


Рис. 4. Динамический авторский след трёх авторов в мемуарах «Воспоминания и размышления» (издание 1969 г.): Г. К. Жукова (синие участки), А. Д. Миркиной (красные участки) и неизвестного третьего соавтора, которым мог быть Вадим Комолов (зелёные участки). Результат применения функции *rolling.classify()*

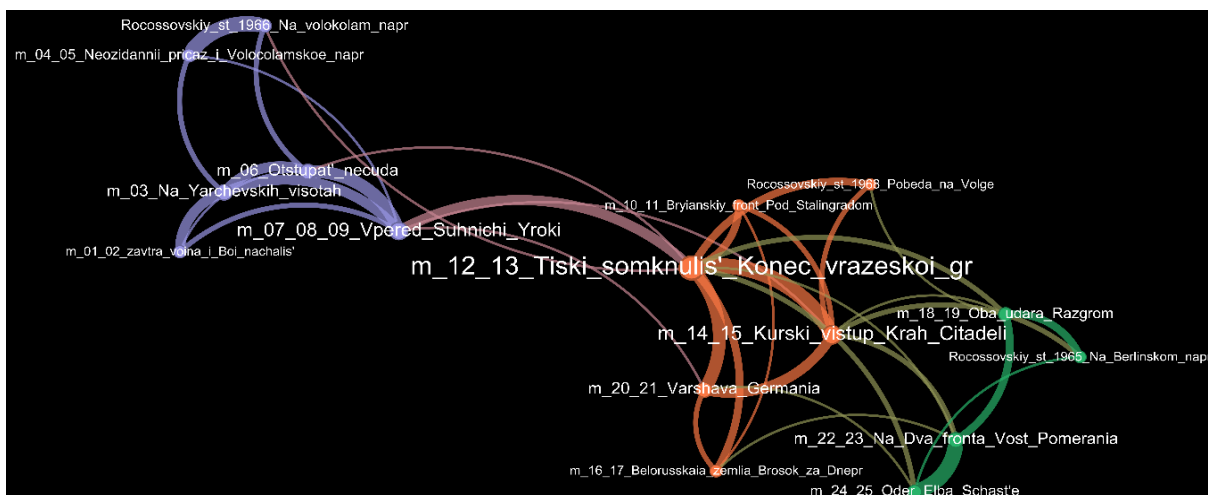


Рис. 5. Визуализация стилеметрической близости глав мемуаров К. К. Рокоссовского «Солдатский долг» (1997 г.) и его же статей через сети

УДК 93

Л. И. Бородкин,

заведующий кафедрой исторической информатики, член-корреспондент РАН
Московский государственный университет им. М. В. Ломоносова, Москва, Россия

НЕЧЁТКАЯ КЛАСТЕРИЗАЦИЯ С ЭЛЕМЕНТАМИ ОБУЧЕНИЯ В ИСТОРИКО-ТИПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

В докладе обсуждается один из подходов к решению историко-типологических задач, основанный на концепциях искусственного интеллекта.

Перспективный вариант реализации типологического анализа при изучении сложных совокупностей объектов социальной природы связан с применением методов многомерной кластеризации. Большинство этих методов основано на однозначном отнесении каждого объекта к тому или иному классу (типу). Выделяемые с помощью этих методов группы объектов, как правило, позволяют выявить специфику каждого типа. Наиболее распространённым методом здесь является кластеризация *k-means*.

Однако при этом в тени остаётся внутренняя структура кластеров, состав ядра типа и его «окружения». Сложным может быть и вопрос об однозначной принадлежности к какому-либо кластеру «пограничных» объектов (переходного типа). Адекватный инструмент для решения типологических задач с учётом указанной специфики объектов социальной природы даёт теория нечетких множеств (*ТНМ, Fuzzy Set Theory*). Как указывает основатель этой теории американский профессор Л. Заде, «её развитие в 60-х гг. обязано большей частью своим идеям задачам, относящимся к распознаванию образов (классификации). Однако по существу глубинная связь между теорией нечётких множеств и распознаванием образов основана на том обстоятельстве, что большинство реальных классов размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен» [1]. В этой связи при создании систем искусственного интеллекта нашла применение *нечёткая логика (FL)*, моделирующая метод рассуждений, который напоминает человеческие рассуждения, подходы к принятию решений. Эти подходы включают все промежуточные возможности принятия решений между «да» и «нет».

В данной работе предложена модификация алгоритма нечёткого кластер-анализа с использованием обучающей выборки, основанного на использовании *ТНМ* [2]. Обучающая выборка позволяет получить проблем-

но-ориентированную кластеризацию. Эта ориентация «заложена» в принципе отбора объектов для обучения. Применение фазсификации (преобразующей входные данные или чёткие числа в нечёткие множества) алгоритма кластеризации *k-means* в задачах аграрной типологии губерний позволяет расширить возможности интерпретации результатов многомерного анализа, введя в рассмотрение количественные оценки степени типичности объектов каждого кластера [3]. В настоящее время опубликованы уже десятки работ, в которых предлагаются различные алгоритмы кластеризации, имеющие целью определение степеней принадлежности объектов к кластерам – размытым (нечётким) множествам. Алгоритмы нечёткой кластеризации с обучающей выборкой можно рассматривать как естественное развитие соответствующих «традиционных» алгоритмов.

Значительно менее разработанными являются алгоритмы распознавания (кластеризации с обучением) нечётких паттернов (образов). При разработке такого алгоритма мы предполагаем, что «учитель» обладает априорным знанием о принадлежности определённых объектов (обучающей выборки) к тем или иным кластерам, т. е. речь идёт о наличии типичных, характерных для каждого кластера объектов («маркеров»), что позволяет сформировать «ядро» каждого кластера. Отметим, что кластеризация с обучением в данной постановке задачи не предполагает статистических оценок свойств обучающей выборки, присущих машинному обучению нейросетей, связанному обычно с большим объёмом данных. Здесь мы имеем дело с «малыми» выборками.

Предлагаемый нами алгоритм определяет значения степени принадлежности объекта обучающей выборки к кластеру не с помощью эксперта, а алгоритмически; при этом решающее правило является «нечётким» и в режиме экзамена. В данной работе обсуждаются результаты применения предложенного подхода для оценки пространственного соотношения двух путей аграрного развития Европейской России в начале XX в.

Многомерная проблемно-ориентированная кластеризация с обучением в задаче аграрной типологии

Одна из крупных задач при изучении экономической истории России конца XIX – начала XX в. связана с построением типологии аграрного развития её регионов. Существенный аспект этой типологии определяется пространственным распределением социальных факторов этого развития. Попытки «жёсткого» определения понятия «тип» в историко-типологическом исследовании нередко входят в противоречие с принципиальной «размытостью» границ между типами. Это обстоятельство и приводит к использованию концепций *ТНМ* в типологических задачах экономической истории.

Построение многомерной социально-экономической аграрной типологии губерний Европейской России в работах И. Д. Ковальченко и Л. И. Бородкина [4] проводилось на основе ряда признаков, включавших следующие признаки социального характера: 1) доля наёмных сельскохозяйственных рабочих по отношению к местным работникам; 2) число сель-

скохозяйственных рабочих в расчёте на десятину посева; 3) доля дворянской земли в удобной земле; 4) отношение проданных частновладельческих земель к общей их площади; 5) отношение арендованной крестьянами земли к надельной земле; 6) доля безлошадных и однолошадных в общем числе дворов; 7) доля дворов с четырьмя и более лошадьми в общем числе дворов; 8) подённая плата сельскохозяйственным рабочим в уборку урожая, коп.

Кластер-анализ данных о 50 губерниях в пространстве восьми указанных признаков выявил три макротипа социальной структуры аграрного развития губерний Европейской России на рубеже XIX–XX вв. (названных И. Д. Ковальченко крестьянским, помещичьим и помещичье-крестьянским). Общий характер соотношения выявленных типов социальной аграрной структуры показывает, что существует промежуточное по всем аспектам положение двух подтипов (промышленного и земледельческого) социальной аграрной структуры. Полюсы представлены, с одной стороны, прибалтийским и степным, а с другой стороны, западным и северо-восточным типами. Однако эти типы представляли полюса далеко не по одним и тем же аспектам. Анализ, проведённый И. Д. Ковальченко [5], показал, что наиболее существенные различия в социальной аграрной структуре губерний Европейской России связаны с двумя путями аграрной эволюции – помещичьим («юнкерским») и крестьянским («фермерским»). Эти пути (варианты развития) могли выступать как в чистом виде, так и переплетаясь в тех или иных комбинациях. Как было показано ранее в наших работах, прибалтийский и западный подтипы были разновидностями аграрной эволюции, в которой доминировало помещичье хозяйство. Степной и северо-восточный подтипы выражают аграрную эволюцию, в которой господствовало крестьянское хозяйство.

Результаты выявления социальной аграрной типологии губерний Европейской России на рубеже XIX–XX вв., полученные с помощью кластер-анализа (без обучения), подводят к вопросу: с каким типом социальной аграрной структуры – помещичьим или крестьянским – имеют больше сходства губернии третьего, помещичье-крестьянского типа социальной аграрной структуры, в котором наиболее заметно переплетались два пути аграрной эволюции. Один из возможных подходов к решению данного вопроса даёт предложенный нами алгоритм. Если включить в обучающую выборку несколько губерний, явно относящихся к помещичьему типу аграрной эволюции, и несколько губерний с выраженными чертами крестьянского типа, то можно определить, к какому из этих типов ближе каждая из рассматриваемых «промежуточных» губерний (на основе степени её принадлежности к каждому из двух кластеров). Так, если очередной объект «попадает» на середину дискриминантной оси (т. е. между центрами двух кластеров, определённых по обучающей выборке), то он относится к обоим кластерам с одинаковыми степенями (весами) принадлежности, равными 0,5. Чем ближе объект к центру кластера, тем ближе к 1 значение веса принадлежности его к данному кластеру.

В обучающую выборку были включены четыре губернии, в которых доминировал помещичий тип аграрной эволюции (Эстляндская, Курляндская, Лифляндская и Ковенская губернии) и шесть губерний (Херсонская, Таврическая, Екатеринославская, Донская, Самарская и Оренбургская) с явным преобладанием крестьянского («фермерского») типа. Таким образом, в восьмимерном пространстве указанных показателей была построена дискриминантная ось, на которой десять губерний из обучающей выборки чётко разделились на два соответствующих кластера. Затем программе предъявлялись поочерёдно все 50 объектов (губерний), заданных значениями восьми показателей. Далее вычислялись веса принадлежности каждого объекта к обоим кластерам.

Для интерпретации результатов распознавания введём пороговое значение 0,50 для веса принадлежности. Будем считать объект принадлежащим к данному кластеру, если вес его принадлежности к этому кластеру больше порогового значения. В соответствии с этим правилом в 22 губерниях из 50 доминировал помещичий, а в 27 губерниях преобладал крестьянский тип аграрной эволюции. Отметим отдельно Уфимскую губернию, которая характеризуется равными весами принадлежности к обоим типам (0,50). В число 22 губерний первого типа вошли все 12 губерний прибалтийского и западного подтипов (помещичий тип социальной аграрной структуры). Характерно, что из остальных десяти губерний с доминированием помещичьего пути почти все были отнесены по результатам «традиционного» кластер-анализа к земледельческому подтипу помещичье-крестьянского типа (исключение составляют лишь Петербургская и Полтавская губернии). Это Орловская, Пензенская, Бессарабская, Черниговская, Симбирская, Смоленская, Псковская и Пермская губернии). Показательно, что веса принадлежности этих десяти губерний к «помещичьему» типу аграрной эволюции оказались намного меньше, чем у 12 губерний двух указанных подтипов помещичьего типа социальной аграрной эволюции.

Для десяти указанных губерний помещичье-крестьянского типа значения весов принадлежности к «юнкерскому пути» не превышают 0,65; в то же время для 12 губерний помещичьего типа значения этих весов не опускаются ниже 0,77. В число 27 губерний с высоким весом принадлежности к крестьянскому типу аграрной эволюции вошли все 12 губерний степного и северо-восточного подтипов. Именно для этих 12 губерний характерны наиболее высокие веса принадлежности к «фермерскому» типу аграрной эволюции.

Полученные результаты дают некоторое (разумеется, приближённое) представление о соотношении потенциальных возможностей двух путей аграрной эволюции в Европейской России на рубеже XIX–XX вв. (с доминированием крестьянского «фермерского» типа). Важным представляется вывод о том, что алгоритмы нечёткой кластеризации с обучением дают адекватный инструмент для построения проблемно-ориентированной классификации/типологии в исследованиях социально-экономических систем.

Список источников

1. Заде Л. А. Размытые множества и их применение в распознавании образов и кластер-анализе / Л. А. Заде // Классификация и кластер. М., 1980. С. 208.
2. Borodkin L. Defining Agricultural Regions in Russia: Fuzziness in Multivariate Classification of Historical Data / L. Borodkin // History and Computing. 2009. No. 11 (1–2).
3. Borodkin L. I. Two Paths of Agrarian Evolution in European Russia: an Essay in Multivariate Analysis / L. I. Borodkin, I. D. Kovalchenko // Russian Review. 1998. Vol. 47. No. 4.
4. Ковальченко И. Д. Аграрная типология губерний Европейской России на рубеже XIX–XX вв. (Опыт многомерного анализа) / И. Д. Ковальченко, Л. И. Бородкин // История СССР. 1979. № 1.
5. Ковальченко И. Д. Методы исторического исследования / И. Д. Ковальченко. М., 1987.

Ю. В. Норманская

Институт языкознания РАН, Москва, Россия

Институт системного программирования им. В. П. Иванникова РАН, Москва, Россия

ВОЗМОЖНОСТИ «ЛИНГВОДОКА» ДЛЯ ВЫЯВЛЕНИЯ СТЕПЕНИ БЛИЗОСТИ ЯЗЫКОВ И ДИАЛЕКТОВ И ОПИСАНИЯ ГРАФИКО-ФОНЕТИЧЕСКИХ ОСОБЕННОСТЕЙ

В настоящее время наиболее разработанной с точки зрения наличия чётких алгоритмов, реализованных в виде компьютерных программ, является классификация языков по количеству родственных слов, между которыми можно установить регулярные соответствия в 100-словном или в 110-словном списке Сводеша. С. А. Старостин предложил специальную формулу скорости распада языков в зависимости от количества различий в списках базисной лексики, на основании которой создан алгоритм построения деревьев языкового родства, реализованный в СУБД *Starling* (URL: starling.rinet.ru), см. подробнее [1].

На платформе «ЛингвоДок» любой пользователь после регистрации может создавать свои словари и/или корпуса и анализировать материалы других пользователей, авторы которых разместили их в открытом доступе. В 2023 г. создана опция «Глоттохронологический анализ языков/диалектов», которую можно запустить в словарях во вкладке «Инструменты». Эта опция может применяться к любому набору языков, в словарях которых доступно более 50 слов из 100-словного списка М. Сводеша. Список Сводеша – Старостина [Старостин 2007: 784] выбран, поскольку лишь для него обоснованы и разработаны строгие семантические спецификации [5], позволяющие получить достаточно точные сравниваемые данные для разных языков. Согласно предложенной С. А. Старостиным глоттохронологии, см. подробнее [3], из подсчёта по формуле, представленной на рис. 1, сначала удаляются заимствования, родственные слова соединяются на платформе «ЛингвоДок» этимологическими связями, затем подсчитывается процент совпадений между списками двух идиомов и вычисляется время распада.

$$t = \sqrt{\frac{\ln\left(\frac{Nn(t)}{N_0}\right)}{-n\lambda^n \sqrt{Nn(t)}}$$

Рис. 1. Формула С. А. Старостина для подсчёта близости языков и диалектов, встроенная в «ЛингвоДок»

Эта формула, для которой С. А. Старостин подобрал экспериментальным путём лямбду, равную 0,05, даёт возможность определить время распада любого набора языков, см. подробнее [3].

На «ЛингвоДоке» также встроена функция создания графиков близости языков в форматах 2D и 3D.

Но большинство компаративистов, особенно на Западе, не вполне доверяет классификации языков, созданной на основе анализа 100-словных списков (о противоречиях, которые возникают при классификации по 100-словным спискам по сравнению с традиционным подходом, подробнее см. в [Беликов 2009]). При этом традиционный подход к генетической классификации языков, основанный в первую очередь на анализе общих фонетических и морфологических инноваций и используемый большинством учёных-компаративистов до сих пор, не был алгоритмизирован и компьютеризирован, а потому его применение могло варьироваться в зависимости от личности учёного.

На платформе «ЛингвоДок» в настоящее время в открытом доступе находятся около 2 тыс. словарей и корпусов, созданных на основании аудиословарей, собранных в полевых условиях в формате *.wav*, и архивных записей по уральским и алтайским языкам. Этот материал позволил нам приступить к разработке программ анализа данных для уточнения транскрипций языковых данных (ранее транскрибирование для коми языков было выполнено «на слух», без привлечения фонетических программ), этимологического анализа и построения классификаций на основании обчёта фонетических инноваций в близкородственных языках и диалектах.

Для определения степени близости диалектов друг к другу с точки зрения фонетических инноваций была разработана программа «Анализ когнатов в разных диалектах одного языка / в разных языках», которую тоже можно найти в любом словаре во вкладке «Инструменты». На первом этапе для каждого символа из транскрипции алгоритм обчисляет его соответствие в словах из других диалектов этого же языка, связанных этимологиями с настоящим словарём. Высвечивается меню, в котором автор отмечает, с какими словарями он хочет проанализировать сравнения.

1. Обсчитываются корни, заранее соединённые этимологическими связями, исходя из того, что в них первый гласный (сочетание гласных) соответствует первому гласному (сочетанию гласных), первый согласный (сочетание согласных) соответствует первому согласному (сочетанию согласных), второй согласный – второму. На выходе получаем для каждой пары идиомов список соответствий. У автора словаря есть возможность скачать его, проанализировать, проверить правильность транскрипций и этимологий, которые привели к нестандартным рядам соответствий, и внести корректировки в транскрипцию и этимологию. Далее алгоритм перезапускается повторно уже на материале, выверенном автором.

2. Алгоритм оценивает, есть ли фонемы, у которых два и более соответствий во втором диалекте. Если две или более фонем из рассматриваемого словаря соответствуют одной фонеме из другого, рассматривается,

нет ли позиционного распределения между ними, не учтённого на первом этапе. По факту этого обсчёта система выдаёт в формате *Excel* список соответствий между двумя диалектами с возможными правилами распределения.

Эта программа позволяет в полуавтоматическом режиме обрабатывать большие массивы словарных данных (15–20 тыс. ед.) для выделения рядов соответствий и дополнительных распределений между ними на материале фонетических словарей диалектов одного языка и языков близкородственных. Эта функция необходима при обработке диалектных материалов; её обычно очень не хватает диалектологам для выяснения полного набора рядов соответствий в однотипном материале. Пока что эта функция «ЛингвоДока», насколько нам известно, не имеет аналогов в других системах.

В результате работы на платформе «ЛингвоДок» с 2012 г. было создано более 1 700 словарей и 440 корпусов текстов из полевых и архивных материалов. Географическое расположение мест, где были собраны материалы, можно посмотреть на сайте «ЛингвоДок» (URL: lingvodoc.ispras.ru/map_search/WLKzz). Все материалы были обработаны с помощью авторских программ «ЛингвоДока», созданных под руководством Ю. В. Норманской, и проэтимологизированы, а также обчислены с помощью программы «Анализ когнатов», которая позволяет установить фонетико-этимологическую близость языков и выявить все инновационные изоглоссы. В результате этой работы были выявлены языки, которые за последние 250 лет практически не претерпели существенных изменений – это карельский, мокшанский, эрзянский, коми-пермяцкий, коми-зырянский.

Но другие уральские языки – саамский, марийский, удмуртский, мансийский, западный и восточный хантыйский, ненецкий, селькупский – изменились весьма значительно. Привлечение архивных данных по этим языкам и их сравнение с материалами современных диалектов позволило выявить интересные особенности, которые не просто отсутствуют в современных диалектах, но и позволяют сделать важные выводы для уточнения праязыковой реконструкции, путей миграций населения, того, кто и когда создал литературный язык и какой диалект XIX в. лёг в его основу.

Анализ материала, проделанный нами на платформе «ЛингвоДок», лёг в основу трёх томов серии монографий «Кириллические памятники на уральских и алтайских языках» [2]. Кратко перечислим здесь значение основных результатов, которые были получены в этих книгах.

I том.

1. Первые кириллические книги позволяют увидеть, как менялась классификация диалектов.

1.1. Коми языки: есть несколько общепринятых, но при этом противоречащих друг другу классификаций коми диалектов/языков: часть исследователей считают коми-язьвинский язык диалектом коми-пермяцкого, другие полагают, что между тремя коми диалектами/языками есть примерно одинаковые различия, которые не позволяют два из них объединять

в особую группу. Анализ памятников XVIII и XIX вв. показывает, что в XVIII в. с точки зрения графико-фонетических особенностей коми-язьвинский и коми-пермяцкий были очень близки и, вероятно, образовывали отдельную группу, в отличие от коми-зырянского языка. К началу XIX в. коми-пермяцкий язык подвергся значительному влиянию коми-зырянского и стал к нему ближе, чем к коми-язьвинскому. Современные диалекты с фонетической точки зрения сохраняют ту же систему различий, которая возникла в XIX в.

1.2. Удмуртский язык: анализ книг второй половины XIX в. и современных диалектов с точки зрения графико-фонетических и лексических особенностей показал отсутствие специфической близости между центрально-южными и периферийно-южными диалектами, которая постулировалась предшествующими российскими исследователями.

1.3. Мансийский язык: было установлено, что в XVIII в. не было чётких диалектных различий между северными и западными мансийскими диалектами. Между северными и восточными диалектами середины XIX в. появляется первое различие: в ауслауте *-y переходит на востоке в -0/-g, а на севере – в -j/-x. Становится ясно, общепринятая мансийская диалектная классификация Л. Хонти построена на признаках, появившихся только в конце XIX в., и нерелевантна для более раннего периода.

2. Первые кириллические книги позволяют датировать время фонетических изменений.

2.1. Саамский язык: сопоставление гласных в прасаамском языке, в переводе первого Евангелия и в современном литературном языке показывает, что изменения системы вокализма: ПСаам. *o > oa, *õ > vū, ы, *ā > oā в литературном кильдинском произошли сравнительно недавно, а именно после 1878 г, когда Евангелие было переведено на саамский язык.

2.2. Марийский язык: прамарийские редуцированные гласные *ĩ и *й ещё сохраняются в первых луговых марийских букварях, созданных в 1870-х гг., в качестве «и», «у» соответственно. Но иногда в этих текстах в качестве дублетов зафиксировано начало перехода *ĩ, *й > ы, которое в XX в. стало их единственным рефлексом.

2.3. Удмуртский язык: было установлено, что большинство фонетических изменений в удмуртских диалектах произошли в XX в., например, ПУдм. *i > бавлинский диалект ə; *u > южные диалекты o и др. В первых книгах XIX в. эти инновации отсутствуют.

2.4. Мансийский язык: в середине XIX в. в восточном диалекте ещё сохранялись в качестве дублетов прамансийская *-y в ауслауте, k- перед гласными заднего ряда. В начале XX в. в этих позициях во всех словоформах *-y перешла в 0, а *k в x.

2.5. Хантыйский язык: ещё в начале XX в. сохранялось шесть гласных второго слога: ə, a, i, e, o, y, в конце XX в. o, y > ə. В восточных диалектах в первом слоге шесть гласных среднего ряда: ä, ö, ó, õ, э, ə, отмеченных в [DEWOS], появились только в XX в.

2.6. Ненецкий язык: в XIX в. в западных диалектах ещё не полностью закончился процесс деназализации – утраты сонорного носового в сочетании «сонорный носовой + шумный», а также утраты /х/ в интервокальной позиции.

3. Первые кириллические книги позволяют верифицировать для исчезнувших диалектов точность транскрипции разных авторов.

3.1. Мансийский язык: на исчезнувших западмансийских диалектах существует две коллекции текстов А. Каннисто и Б. Мункачи, транскрипции которых несводимы друг к другу. Кириллический словарь, проанализированный нами, позволил верифицировать бóльшую степень точности транскрипции А. Каннисто.

4. Первые кириллические книги подтверждают реконструкцию праязыка.

4.1. Саамский язык: проведённый анализ показывает, что система гласных графем в [Евангелие от Матфея 1878] в исконной лексике практически однозначно соответствует прасаамской реконструкции по [Lehtiranta 1989]. Интересно, что, несмотря на инновационные процессы дифтонгизации в литературном кильдинском, и сегодня есть примеры сохранения архаической прасаамской системы гласных – например, в воронинском говоре кильдинского саамского.

4.2. Селькупский язык: разноместное ударение в южных и центральных диалектах практически не отмечено в опубликованных словарях, грамматиках и текстах. В книгах Н. П. Григоровского, изданных в 1879 г., оно указано и полностью коррелирует с ударением, отмеченным в картежке А. П. Дульзона, которая была собрана в XX в. и содержит около 100 тыс. лексем из южных и центральных диалектов, и с ударением в современных тундровых ненецких диалектах. Очевидным образом, оно должно быть реконструировано для прасамодийского языка, см. подробнее монографию [Норманская 2018].

5. Памятники позволяют уточнить реконструкцию праязыка.

5.1. Марийский язык: вопрос о реконструкции редуцированных гласных на протяжении 70 лет был предметом дискуссии учёных, его исследованием занимались выдающиеся финно-угроведы: В. Штейниц, Э. Итконен, Г. Берецки, Л. П. Грузов [Steinitz 1944; Itkonen 1953; Bereczki 1992; Грузов 1964 (1, 2), 1966]. В. Штейниц, Э. Итконен и Л. П. Грузов считали марийские редуцированные архаизмом, а Г. Берецки – современной инновацией, произошедшей в некоторых (в первую очередь в горных) марийских диалектах под влиянием тюркских языков. Анализ первых луговых марийских букварей 1870-х гг. показывает, что в них представлены особые рефлексы марийских редуцированных гласных **ĩ* и **й*, которые исчезли в современном луговом марийском. Это является важным аргументом в пользу реконструкции прамарийских редуцированных гласных.

5.2. Мансийский язык: анализ показал, что необходимо уточнить прамансийскую реконструкцию Л. Хонти, выполненную на основе анализа более поздних материалов конца XIX – начала XX в.: реконструировать

ПМанс. *a вместо *j̄, *č̄ вместо *ć̄, поскольку именно так выглядят рефлексы этих звуков во всех проанализированных первых источниках.

5.3. Хантыйский язык: практически полная корреляция систем вокализма в васюганском, ваховском (восточном) и обдорском (западном) диалектах в XIX в. позволяет предположить, что схожая система должна быть реконструирована и для прахантыйского языка. Ранее прахантыйская реконструкция вокализма Е. А. Хелимского базировалась на данных ваховского диалекта в транскрипции [DEWOS], где присутствует шесть гласных среднего ряда: *ö, õ, ó, õ̃, ə, ə̃*, появившихся только в источниках XX в.

II том посвящён описанию двух первых книг «Материалы для ознакомления с наречием остяков Нарымского края» (1887) и «Беседы об истинном Боге и истинной вере на наречии обских остяков» (1900), созданных святителем Макарием (Невским) в рамках его миссионерской деятельности. В книге выполнено системное описание графико-фонетических, морфологических и синтаксических особенностей этих изданий. Оказалось, что эти книги созданы с высокой степенью точности и отражают некоторые архаизмы, не представленные в других доступных нам источниках по селькупскому языку, но реконструированные ранее для прасамодийского языка. Удалось показать, как значительно может меняться язык носителей того или иного региона: за 100 лет среднеобской диалект, который в книге св. Макария имел южно-селькупские особенности, в записях А. П. Дульзона и его последователей, сделанных в 1960–90-х гг., приобрёл черты центрального диалекта. При анализе морфологии были привлечены данные всех доступных источников по другим диалектам селькупского, в т. ч. материалы ещё неизданных картотек, и показано, как менялись ареалы распространения тех или иных аффиксов.

III том посвящён лингвистическому анализу фонетико-морфологических особенностей пяти удмуртских книг конца XIX – начала XX в.: «Ўгет. Святой Тихон дышетэм зеч кыльёс. Наставление христианское св. Тихона» (1878), «Букварь для вотяцких детей Сарапульского уезда» (1882), «Зеч кыльёс. Святой Тихонлэн зечлы дышетэм кыльёсыз. Наставление христианское святителя Тихона на вотском языке» (1891), «Закон божий: Книжка с картинками для маленьких детей на вотском языке глазовского наречия» (1912) и «Святой Феодорлэн улэмез. Житие св. Феодора. На вотском языке елабужского наречия» (1913). Все источники были подготовлены в рамках работы Переводческой комиссии Православного миссионерского общества, при этом никаких сведений об авторах и месте создания переводов на страницах изданий нет. В монографии все данные памятников сопоставляются со словарём удмуртских диалектов [6] и с данными по современным говорам. Интересно отметить, что по некоторым из вышперечисленных современных говоров нет столь подробной информации о рефлексации всех фонем и морфем. Но материалы Ю. Вихманна полностью соответствуют первым кириллическим книгам, когда этот диалект у него представлен. В частности, отражение гласных в словах книги «Зеч кыльёс. Святой Тихонлэн зечлы дышетэм кыльёсыз. Наставление христианское свяи-

теля Тихона на вотском языке» (1891) точно совпадает с малмыжско-уржумским диалектом в [6], причём речь идёт не о тенденциях, а о соответствиях в тех словах, которые ранее учёные считали нерегулярными отклонениями. Эта корреляция доказывает точность как словаря Ю. Вихманна, так и первых книг, составленных священниками с помощью носителей языка.

В книге М. П. Безеновой показано, что в конце XIX – начале XX в. наиболее сильно от современного литературного удмуртского языка отличались книги, написанные на говорах периферийно-южного диалекта: «Ўгет. Святой Тихон дышетэм жеч кыльёс. Наставление христианское св. Тихона» (1878), «Жеч кыльёс. Святой Тихонлэн жечлы дышетэм кыльёсыз. Наставление христианское святителя Тихона на вотском языке» (1891), в них представлено 8–10 (в зависимости от книги) инновационных процессов и 4–5 позиций, в которых сохраняется архаическое произношение. Северное, центрально-южное и промежуточное наречия были в тот период с точки зрения фонетики достаточно близки друг другу и незначительно отличались от литературного удмуртского. Интересно, что с точки зрения морфологии, наоборот, наибольшее количество инноваций имеет книга на северном наречии: «Закон божий: Книжка с картинками для маленьких детей на вотском языке глазовского наречия» (1912). Таким образом, материалы, собранные в исследовании М. П. Безеновой, позволяют в ряде случаев описать хронологию изменений в удмуртских диалектах и уточнить степень их близости.

Работа выполнена при поддержке РНФ № 20-18-00403 «Цифровое описание диалектов уральских языков на основании анализа больших данных».

Список источников

1. Бурлак С. А. Сравнительно-историческое языкознание / С. А. Бурлак, С. А. Старостин. М., 2005.
2. Норманская Ю. В. Кириллические памятники на уральских и алтайских языках. Т. 1–3 / Ю. В. Норманская. М., 2022.
3. Старостин С. А. Сравнительно-историческое языкознание и лексикостатистика / С. А. Старостин // Лингвистическая реконструкция и древнейшая история Востока. Ч. I. М., 1989. С. 3–39.
4. Starostin G. Preliminary Lexicostatistics as a Basis for Language Classification: a New Approach / G. Starostin // Journal of Language Relationship. 2010. No. 3. Pp. 79–116.
5. Kassian A. The Swadesh Wordlist. An Attempt at Semantic Specification / A. Kassian, G. Starostin, A. Dybo et al. // Journal of Language Relationship. 2010. No. 4. Pp. 46–89.
6. Wichmann Y. Wotjakischer Wortschatz / Aufgezeichnet Y. Wichmann; Bearb. T. E. Uotila, M. Korhonen. Helsinki, 1987.

B. V. Orekhov,

Candidate of Philological Sciences, Associate Professor, Senior Researcher
National Research University Higher School of Economics, Moscow, Russia
Institute of Russian Literature (Pushkin House) of the Russian Academy of Sciences,
St. Petersburg, Russia
e-mail: nevmenandr@gmail.com
ORCID: 0000-0002-9099-0436

HARD CASES IN CROSS-CORPORA DRAMA ANNOTATION

The paper deals with the observations that are made during the work on Dracor.org [1]. Originally, Dracor.org is a set of isolated corpora in different languages with a unified TEI markup that allows you to do quantitative drama studies [2]. But since the platform now includes a noticeable set of corpora in German, English, Russian and other languages, it is natural to try to study cross-corpora drama features. These features include characters that appear in different plays and even in different corpora. Usually these are historical personalities (e.g. kings), mythological figures (Achilles, Heracles), popular literary characters (Faust). The idea is to mark all the characters that are mentioned in the plays more than once. We must not forget that Euripides' Achilles (*Iphigenia in Aulis*) is a different character than Kleist's Achilles (*Penthesilea*). Technically, these characters are part of different fictional systems. But in the future we would like to explore the differences of interaction between these characters and the other components of the play. This will help to establish the very difference between fictional systems and the author's style. So we have to mark the heroes with the same name to be able to extract them automatically from the corpus.

All characters in Dracor have their ids, which depend on their names:

```
<person xml:id="Othello_Oth">  
<persName>Othello</persName>  
</person>
```

But id in the play is not suitable for cross-corpora annotation tasks. For example, the character "messenger" is found in many plays, and often has the same id, but this is not the case when we are going to mark these characters as the same cross-corpora character. In addition, characters in different plays may occasionally have the same names. That means we need to have a separate attribute for this kind of case.

We have chosen the attribute "ana" for the tag "person". Due to the TEI guidelines this attribute "indicates one or more elements containing interpretations of the element on which the ana attribute appears". This means that we interpret the character of the play as some kind of entity that may appear

in other plays. The most convenient option would be to associate the value of this attribute with the Linked Open Data. We use Wikidata, which provides the URI for the entity we find in different plays.

```
<person xml:id="phaedra" sex="FEMALE" ana="http://www.wikidata.org/entity/Q212818">
```

This approach works very well and allows us to identify the characters in different plays. But there are cases where we can see the inconvenience of Wikidata URI for the purposes of our cross-corpora markup. There are three types of such cases.

1. ***Umbra case***. In ancient plays, characters and their shadows are often found separately. Sometimes they are found within the same play. Then they have different dracor ids, but the URI on Wikidata is the same. See Tantalus and Tantaliumbra in *Thyestes* by Seneca.

```
<person xml:id="tantalus" sex="MALE" ana="http://www.wikidata.org/entity/Q515531">
<persName>Tantalus</persName>
</person>
<person xml:id="tantaliumbra" sex="MALE" ana="http://www.wikidata.org/entity/Q515531">
<persName>Tantali Umbra</persName>
</person>
```

and Δαρειός and Εἶδωλον Δαρείου in *Persians* by Aeschylus:

```
<person xml:id="eidolon_dareiou" sex="MALE" ana="http://www.wikidata.org/entity/Q44387">
<persName>Εἶδωλον Δαρείου</persName>
</person>
<person xml:id="dareios" sex="MALE" ana="http://www.wikidata.org/entity/Q44387">
<persName>Δαρειός</persName>
</person>
```

At the moment we have to mark these characters with a single id. But it's not the best solution.

2. ***Ivppiter case***. The purpose of cross-corpora markup is to tie together characters from different plays and different corpora at the same data object. From this point of view, it is obvious that Zeus and Jupiter are the same character, they perform the same function in the story about Hercules. But in Wikidata we find different URIs for Zeus (URL: wikidata.org/wiki/Q34201) and Jupiter (URL: wikidata.org/wiki/Q4649). Thus, our data is separated. The plan is to get all the necessary information about the characters using SPARQL queries and characters ids. Perhaps, this case could be solved by the additional parameter of SPARQL query (P460). But every time this parameter must be different (depends on gods, kings or someone else we deal with).

3. ***Talthybius case***. Talthybius has two relevant objects in Wikidata. First of all, it's a character from a myth (URL: wikidata.org/wiki/Q1367860). This URI allows us to connect the character of the tragedy *Troades* by Seneca and *The Trojan Women* by Euripides. But the Euripides tragedy character has his own URI in Wikidata (URL: wikidata.org/wiki/Q60607554). Technically,

its attribute should have this value, but it contradicts the idea of tying the characters of different plays together. Additional SPARQL parameter P1074 ("fictional analog of") could solve this problem. Another solution is to extend Wikidata by ourselves and add Talthybius to the parameter "present in work" (P1441) from Dracor to the certain play objects.

References

1. Skorinkin D. Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies / D. Skorinkin, F. Fischer, G. Palchikov // Компьютерная лингвистика и интеллектуальные технологии. 2018. Pp. 662–682.

2. Skorinkin D. Measuring the "Epification" of Drama / D. Skorinkin, F. Fischer // Applications in Cultural Evolution: Arts, Languages, Technologies. Tartu: University of Tartu, 2018. Pp. 46–48.