



цифровые
гуманитарные
исследования

**ЦИФРОВЫЕ ГУМАНИТАРНЫЕ
ИССЛЕДОВАНИЯ**

2025 № 2 (003)

ИНСТИТУТ РУССКОЙ ЛИТЕРАТУРЫ (ПУШКИНСКИЙ ДОМ) РАН
ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ. 2025 № 2.

ISSN (Online) 3034-4522

Главный редактор
Орехов Б. В.
(Москва–Санкт-Петербург)

РЕДАКЦИОННЫЙ СОВЕТ

Алиева О. В. (Москва)
Акимова М. В. (Москва)
Беляк Г. Н. (Санкт-Петербург)
Балакин А. Ю. (Санкт-Петербург)
Белоусова А. С. (Богота, Колумбия)
Бонч-Осмоловская А. А. (Москва)
Вдовин А. В. (Москва)
Володин А. Ю. (Москва–Красноярск)
Гагарина Д. А. (Бишкек, Кыргызстан; Эрланген, Германия)
Кижнер И. А. (Хайфа, Израиль)
Колозариди П. В. (Санкт-Петербург)
Ляшевская О. Н. (Москва)
Маслинский К. А. (Санкт-Петербург)
Павлова Л. В. (Смоленск)
Полилова В. С. (Москва)
Пучковская А. А. (Лондон, Великобритания)
Романова И. В. (Смоленск)
Северина Е. М. (Ростов-на-Дону)
Сенаторова Е. Е. (Нью-Йорк, США)
Скоринкин Д. А. (Потсдам, Германия)
Шерстинова Т. Ю. (Санкт-Петербург)

© Авторы статей, 2025;

.....
Санкт-Петербург

ОГЛАВЛЕНИЕ

ИССЛЕДОВАНИЯ

Александра Митюкова

Онтологическая модель для связывания метаданных музейных предметов 4

Ксения Анисимова

Анализ тональности русской драмы XVIII–XX вв. как инструмент моделирования художественной структуры . . . 24

Елизавета Сенаторова

Количественный анализ речи персонажей в экранизации романа Л. Н. Толстого «Анна Каренина» (реж. Александр Зархи, 1967 год) 48

ИНСТРУМЕНТЫ

Борис Орехов

Открытые компьютерные инструменты для решения задач оцифровки и анализа русскоязычного текста в области Digital Humanities 71

ИСТОРИЯ ЦИФРОВЫХ МЕТОДОВ

Андрей Володин

Цифровые гуманитарии: от академических племен к эпистемическому сообществу 84

ХРОНИКА

Мария Кешисян

Конференция «Актуальные ошибки гуманитарных наук» . . 117

РЕЦЕНЗИЯ

Софья Порфирьева

Рецензия на книгу «Герменевтика: компьютерная интерпретация в гуманитарных науках» Стефана Синклера и Джеффри Роквелла 124

ИССЛЕДОВАНИЯ

Александра Митюкова

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ДЛЯ СВЯЗЫВАНИЯ МЕТАДАННЫХ МУЗЕЙНЫХ ПРЕДМЕТОВ

В статье описан пример разработки онтологической модели для концептуализации метаданных английских пейзажных гравюр и предметов Сервиза с зеленой лягушкой, изображающих архитектурные достопримечательности Великобритании. Модель призвана объединить метаданные различных предметов искусства и архитектурных памятников и их контекстные данные и проиллюстрировать характер взаимосвязи между ними. Концептуальная часть онтологии выравнена с актуальными тезаурусами с целью унификации терминологии. Для экземпляров предусмотрено связывание данных с таксономиями и контролируемыми словарями, обеспечивающее не только их верификацию, но и открывающее широкие возможности для обогащения данных онтологии. Онтология разработана с учетом структуры концептуальной модели верхнего уровня CIDOC CRM и предусматривает разметку с ее элементами. Объекты, являющиеся уникальными каждый по отдельности, с помощью онтологии могут быть связаны в одном проекте, применимом для изучения, экспонирования и публикации комплекса предметов, хранящихся в различных учреждениях культуры.

Ключевые слова: семантическое моделирование, онтология предметов искусства, связанные данные, связывание метаданных, выравнивание сущностей, тезаурусы, таксономии, контролируемые словари

Разработка¹ онтологических моделей для описания художественных предметов осуществляется в мировой практике с середины XX в. Предпосылками для развития этой сферы послужили различные задачи, возникающие перед сотрудниками институтов памяти при каталогизации, систематизации и изучении артефактов [Юмашева 2023]. Преимущества онтологий в применении к описанию предметов искусства кроются в том, что онтологические модели обладают необходимыми свойствами высокой адаптивности и пригодны для решения проблем синхронизации данных различных типов и отраслей.

Информация о культурном наследии создает значительные трудности для формального обращения, по причине своей разнообразности и неоднородности, а неполнота информации о прошлом является ее неотъемлемым свойством. Кроме этого, в настоящее время возникает потребность в стандартизации описаний больших объемов данных из различных институтов памяти, в каждом из которых система каталогизации и описания имеет различный формат [Жлобинская 2013]. Эти трудности потребовали разработки универсальной системы, способной не просто объединить формальные описания предметов искусства, но сделать их универсальными, чтобы они могли обрабатываться автоматическими методами. Первым шагом на этом пути стали словари-тезаурусы и таксономии, контролируемые терминологию и унифицирующие ее как в интерлингвистическом плане, так и в семантическом [Barbot 2023]. Иерархическая структура таких тезаурусов позволяет достаточно однозначно определять терминологию для описания предметов как на физическом уровне, так и на иконографическом. Однако описание художественных предметов исторически формировалась также в и семантическом направлении. Эти три уровня описания искусства, выделенные еще Панофским² [Панофский 1999], обычно разделены в различных видах документации о музейных предметах: музейных карточках, научных статьях, аннотациях к выставкам и музейным каталогам, каждый из которых обращается к отдельному уровню описания. С переходом на электронные носители информации данные о музейных предметах из карточек и статей трансформировались в реляционные базы данных, позволяющие довольно успешно оперировать как иконографическими данными, так и физическими характеристиками. Однако семантическая часть довольно сложно сочетается с реляционным типом формирования данных. Для анализа и поиска отношений между событиями, субъектами и объектами, оценки контекстной информации, в том числе

в континуальном аспекте, требуется экспертное знание и оценка, имеющие, в свою очередь ограничения в виде человеческого ресурса. Анализ информации о контексте производства и бытования произведений требует комплексной научной работы, связанной в том числе, с работой с архивными документами, историческими источниками и специальной литературой. Стоит отметить, что степень изученности и глубина описания предметов не одинаковы, и имеющееся количество данных может различаться от экземпляра к экземпляру, даже в рамках одного учреждения хранения, не говоря уже о различных институциях. Это затрудняет научное исследование предмета и не позволяет соотнести очевидные сведения при изучении и описании художественного произведения. При рассмотрении нескольких подобных произведений зачастую нет возможности выявить очевидные закономерности, и для проведения каких бы то ни было параллелей требуются специальные знания и изрядное количество времени.

Преимущественная особенность онтологического моделирования в этом смысле состоит в том, что кроме определения основного свойства объекта, онтология подсвечивает и характер отношений между разными объектами. То есть связи внутри одной онтологической модели могут быть не только иерархическими, но настолько разнообразными, насколько этого требует характер, формат, объем и свойства описываемых явлений. Выход за рамки отношений гипонимии/гиперонимии, характерных для таксономий и тезаурусов, и является признаком и задачей онтологической модели [Иванов 2007]. Тезаурусы обычно служат базисом для создания онтологий, определяющим перечень классов-концептов, между которыми определяются свойства-связи, обеспечивающие семантический уровень онтологии. Таким образом, суть и смысл онтологического моделирования кроется именно в семантических связях, определяющих взаимоотношения между объектами онтологии. Эти связи также могут выстраиваться в иерархии и определять отношения между объектами на разных уровнях.

Построение онтологической модели, таким образом, позволяет отразить социальные, культурные, исторические или географические взаимосвязи, помогая соотнести неочевидные или скрытые отношения, дополнить неизученные данные или подтвердить ранее поставленные гипотезы. Помимо этого, онтологическая модель, используемая для хранения и оперирования данными, делает их доступными для обработки автоматическими компьютерными

методами, что упрощает операции при большом количестве или сложной структуре данных, минимизируя возможность ошибки.

Поводом для разработки описываемой онтологической модели стал интереснейший феномен европейской культуры XVIII в. — слияние трех различных видов искусства в уникальном памятнике — Сервизе с зеленой лягушкой, заказанном императрицей Екатериной II в 1770 г. у английского керамиста Дж. Веджвуда³. Веджвуд со своим партнером и соратником Т. Бентли в кратчайшие сроки (1773–1774) создал и расписал поистине беспрецедентный комплект — 952 предмета с более чем 1244 изображениями достопримечательностей Британии на них. Большая часть видов была заимствована с ландшафтных гравюр, которые были отпечатаны ранее или создавались непосредственно для декорирования предметов сервиза известными английскими художниками и граверами. Созданные ими гравюры, отпечатанные в Лондонских типографиях, стали источником изображений для истинного шедевра — сервиза, позволившего императрице «иметь всю Англию на своем столе». Гравюры, будучи самостоятельными произведениями искусства, изображали значимые архитектурные, исторические, парковые и природные объекты. Многие гравюры были созданы с написанных ранее живописных произведений, некоторые из которых не сохранились. Такое многоуровневое и разнородное заимствование стало основой для разработки модели, которая объединяет данные о разных по направлению предметах искусства, связанных единым контекстным полем.

Разрабатываемая онтологическая модель, призвана отразить и объединить метаданные трех различных комплексов памятников и проиллюстрировать взаимосвязи между предметами сервиза, гравюрами и архитектурными достопримечательностями. Объекты, являющиеся исключительными каждый по отдельности, с помощью онтологической модели теперь получили возможность быть объединенными в одном проекте, применимом для изучения, экспонирования или публикации подобных комплексов, хранящихся в различных учреждениях культуры. Построение онтологической модели, решает проблему разнородности описаний каждого отдельного вида предметов и вариативности форматов этих описаний в различных учреждениях хранения и позволяет отразить всевозможные виды взаимосвязей, помогая соотнести неочевидные или скрытые отношения. Онтологическая модель делает хранимую информацию доступной для обработки автоматическими компьютерными методами. Возможности онтологии позволяют расширить

модель для реконструкции исторического и социокультурного контекста: отражения социальных, культурных, политических отношений между персоналиями, событиями и предметами искусства. Разработка такой модели важна не только в применении к Сервизу с зеленой лягушкой, но и для всех других подобных феноменов и может быть применена в будущем для связывания метаданных и описания произведений искусства, где аналогичным образом соединяются различные направления искусства.

Для построения модели использовалась middle-out методология SAMOD (Simplified Agile Methodology for Ontology Development) [Peroni 2016]. Данная методология имеет итеративную структуру. Методология предполагает последовательную разработку отдельных узлов модели и пошаговое наполнение их данными для валидации концептуализации. После добавления данных, при необходимости, концептуальная часть может быть отредактирована с учетом специфики загруженных сущностей, количества и качества данных, их типов и структуры.

На первой стадии процесс разработки включал сбор данных о гравюрах, сервизе и изображенных объектах. Для сбора данных о гравюрах использовались тексты, указанные на гравюрах [Stijnman 2012], метаданные музейных предметов, материалы научных статей и публикаций [Бардовская, Ботт 2019]. Для сбора данных о предметах сервиза, кроме указанных научных публикаций [Williamson 1909; Воронихина 1988] были использованы данные Госкаталога музейного фонда, собранные автоматическими методами через открытый API Министерства культуры [Спецификация API]. Сведения об изображенных объектах получены из текстов, отпечатанных на гравюрах, сопоставлены с данными каталога предметов сервиза [Raeburn, Voronikhina, Nurnberg 1995] и сверены с различными геоинформационными системами с определением точных географических координат изображенных объектов [Geonames].

Этап обработки собранных данных заключался в определении основных узлов концептуальной части, характеризующих основные блоки информации, и общие связи, отражающие ключевые отношения между этими блоками (рис. 1).

Дальнейшее уточнение концептов подразумевало выделение подклассов, характеризующих как физические данные предметов, так и информацию, полученную о них в процессе изучения. Для наших задач представления гравюр, их прототипов — живописных произведений и предметов фарфора, на которые перенесены виды

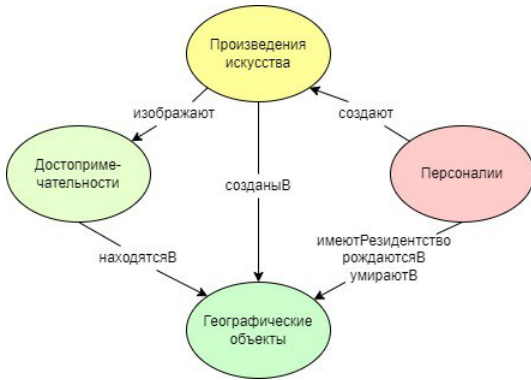


Рис. 1. Основные блоки классов онтологии и их свойства

с гравюр, был определен класс *Произведений искусства*. *Гравюры* и *Живописные произведения*, а также *Рисунки*, по общепринятому принципу деления произведений искусства, мы относим к подклассу *Изобразительного Искусства*, а *Фарфор* — к подклассу *Прикладного*. Кроме того, мы знаем, что прототипом гравюры может служить живописное полотно, в том случае, когда оно известно, как существующий или существовавший ранее (утраченный) физический объект искусства. Для минимум одной гравюры из определенной нами выборки, прототип существует в виде рисунка, также физического объекта, находящегося в собрании определенного культурного учреждения. Но прототипом для гравюры может быть рисунок или набросок, созданный автором непосредственно для производства гравюры. Зачастую такие рисунки не сохраняются или об их существовании неизвестно. При этом авторство рисунка-прототипа однозначно определено надписью на гравюре, указывающей, что рисунок создан неким автором (обозначения: *delineator*, *inventor*, *auctor*, *figurabat*) [Stijnman 2012]. Для таких случаев мы определим подкласс *Прототип*. Мы договоримся, что физически существующие рисунки мы будем включать в подкласс *Рисунок*, а гипотетические — в подкласс *Прототип*. Для реализации связи гравюры с ее прототипами: живописными или графическими объектами, введена связь *бытьПрототипом*, имеющая класс *Гравюры* в области значений и классы *Живопись*, *Рисунок* и *Прототип* в области определения.

Для подкласса *Прикладного искусства* мы определили следующий подкласс: *Фарфор*. В подклассе фарфора нам необходимо выявить все возможные виды предметов описываемого нами сервиса. Для получения данных о категориях предметов и прочих метаданных, описывающих фарфоровые изделия, было принято решение осуществить парсинг данных с сайта Госкаталога через предоставляемые Министерством культуры РФ открытый API [Спецификация API]. Для скачивания и обработки данных предоставленный запрос *cURL* был преобразован в запрос *request*. В вычислительной среде Google Collab произведено объединение скачанных файлов в формате *.json*, разделение столбца наименования на предметное слово, описание формы и прочую информацию⁴. В результате обработки удалось выделить ряд категорий предметов. Некоторые из них имеют уменьшительные формы, по которым принято решение о слиянии. При сравнении с исторической и научной литературой, описывающей состав сервиса [Williamson 1909], мы понимаем, что большая часть категорий предметов покрывается этими данными. Добавление очевидно отсутствующих видов предметов возможно после сверки наименований в каталоге сервиса и научных публикациях [Raeburn, Voronikhina, Nurnberg 1995].

Для всех сущностей классов произведений искусства определен ряд свойств данных, характерных для всех предметов искусства. Это размеры, форма, идентификаторы (инвентарные номера, каталожные номера), даты.

Общая схема реализации взаимосвязей предметов искусства и изображенных на них памятников может быть представлена диаграммой на рис. 2.

Класс *Достопримечательностей* создан для определения изображенных на предметах объектов. Он является областью значений для свойства *изображать*, областью определения в котором выступает *ПроизведениеИскусства*. Поскольку среди приблизительно 1000 имеющихся видов всего сервиса присутствуют разнообразные природные или рукотворные памятники, для первоначального деления была использована классификация, произведенная исследователями сервиса [Воронихина 1988]. После объединения некоторых из этих категорий, в конечном итоге список подклассов представлен пятью пунктами и определен следующими видами:

1. Виды парков и усадеб;

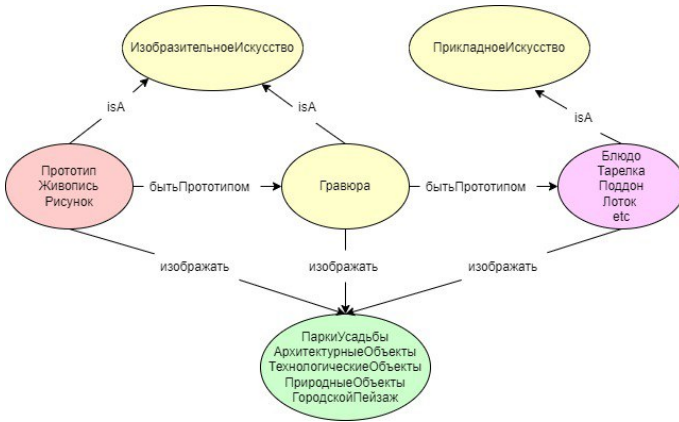


Рис. 2. Разделение класса Произведений искусства на подклассы

2. Замки, средневековые аббатства, приораты, церкви, руины древних дворцов;
3. Природные достопримечательности;
4. Технологические, промышленные объекты и мосты;
5. Городской пейзаж (Темза и Лондон).

Для сущностей класса достопримечательностей определены свойства данных, обозначающие их месторасположение: широту и долготу.

Класс *Географических объектов* имеет универсальный характер и используется и в качестве области значений (range) в самых разнообразных свойствах объектов (object properties) сразу у нескольких классов. Это указание на:

- место рождения и смерти персоналий;
- место расположения достопримечательностей, изображенных на предметах искусства;
- место производства произведений искусства (в том числе местонахождение фирм-производителей).

Связи, составляющие область значений *Географические Объекты* — это *находитьсяВ* для области значений *Организации* и *Достопримечательности*; *бытьСозданнымВ* — для *ПроизведенийИскусства*.

Подклассы для *Географических объектов* определены различными административными субъектами, представленными по степени их необходимости при описании той выборки данных, которую мы договорились исследовать. Это *Страны* и *Города*, связанные свойством *находитьсяВ*, а также *Графства Великобритании*. Административное деление Великобритании является одним из самых сложных и многоуровневых в мире. Вся территория крупных островов делится на Регионы (англ. Regions, или Government Office Regions) административно-территориальные единицы верхнего уровня. Каждый регион включает одну или несколько единиц уровня графств. Графство (shire, county) — основная административно-территориальная единица Англии; в настоящее время насчитывается 48 церемониальных графств. Для определения расположения природных и иных архитектурных объектов, церемониальные графства подходят лучше всего, так как изображенные объекты чаще всего имеют историческое расположение в определенных графствах, а принадлежность к тому или иному графству часто указывается в текстах на гравюрах.

Класс *Организации* определен для обозначения различных учреждений, тем или иным способом задействованных в производстве или хранении произведений искусства. Класс является областью значений для свойства *хранитьсяВ*, областью определения для него являются *ПроизведенияИскусства*. В классе *Организаций* выделены подклассы *Музеев* и *СоциальныхИнституций* для обозначения институтов памяти, в которых в настоящее время находятся описываемые предметы искусства. К подклассу *Социальных институций* относятся различные учреждения, где могут находиться на хранении объекты, значимые для нашей онтологии. Например, Консульство Великобритании в Финляндии, где хранится живописное полотно Т. Смита, являющееся источником для гравюры, изображающей долину Давдэйл, является сущностью класса социальных институций.

Класс *Персоналий* — важнейший из классов онтологии. Для класса *Персоналий*, определены связи, помогающие описанию биографических данных: *родитьсяВ*, *умеретьВ*, *иметьТитул/Звание*. Свойства данных: дата рождения, дата смерти в формате xsd:YEAR и титул/звание в строковом формате определяют точные

характеристики этих событий и свойств. Класс не имеет подклассов, так как все виды действий акторов реализуются посредством определенных свойств объектов. Так, для класса персоналий определена высшая связь *создавать*, имеющая область определения (domain) *Персоналии*, а область значений (range) — *Произведения-Искусства*. Эта связь применима тогда, когда участие в создании произведения искусства актора подтверждено, однако мы не можем однозначно идентифицировать специфику этой связи.

Для более конкретных действий определены связи нижней ступени иерархии: *гравировать*, *нарисовать*, *опубликовать*, *писатьМаслом*, *построить*, *проектироватьЛандшафт*, *произвести*, *создатьОфорт*. Эти связи также имеют область определения класс *Персоналий*, а областью значений для них являются конкретные виды произведений искусства, в зависимости от свойства: *Рисунок* для *нарисовать*; *Живопись* для *писатьМаслом*; *Гравюра* для *гравировать*, *опубликовать* и *создатьОфорт*; *ПаркиУсадьбы* и *АрхитектурныеОбъекты* для *построить* и *ПаркиУсадьбы* для *проектироватьЛандшафт*. Поскольку *ФирмыПроизводители* встречаются в самых различных сферах изготовления произведений искусства: это и печатные (литографии и типографии) производства, и, например, гончарная фабрика, — область определения и область значения для этого свойства выбрана более общая: *ФирмыПроизводители* и *ПроизведенияИскусства* соответственно. Этот принцип является общепринятым при построении онтологий художественных предметов, так как одно и то же лицо может выступать в разных свойствах одновременно или последовательно для одного или различных объектов. Для формирования корректных запросов о способе и авторе производства такая организация данных является оптимальной.

Из знаний о предметной области мы понимаем, что кроме категорий собственно художественных объектов, существуют также их комплекты, обладающие отдельными характеристиками. Для гравюр характерно объединение их в серии по тематике или авторству. В этом случае, обычно листы имеют схожий формат, одинаковое оформление, одного издателя. Даже если какие-то доказательства серийности отсутствуют на одном экземпляре гравюры (срезаны поля с надписями), по другим экземплярам гравюры определяется принадлежность листа к той или иной серии. Таким образом тиражность этого вида искусства обеспечивает полноту данных для исследования. Сведения о производстве серий гравюр часто находятся в рекламных или каталожных публикациях, от-

куда попадают в научный оборот. Гравюры из одной серии могут бытовать комплектом или по-отдельности. Кроме того, повторные издания серий могут осуществляться целиком или допечатываться избранными листами, в зависимости от целей и возможностей переиздания.

Комплектность фарфора в данном случае определяется категорией *Сервиз*. О принадлежности предметов искусства к одному сервизу свидетельствует одинаковое оформление изделий, повторяемость формы, схожая тематика декора, совпадение данных о месте и времени производства, фирме-производителе. Из научной литературы нам известно, что Сервиз с зеленой лягушкой делится на обеденную и десертную часть. С учетом этих особенностей в классе сервиза были выделены подклассы *Обеденной* и *Десертной* части. Для реализации принадлежности произведений искусства к комплекту определено свойство *быть Частью*. Определение и указание комплектности произведений искусства является важным для понимания контекста создания произведений, уточнения авторства, датировок, места производства.

Особое внимание уделено дифференциации текстовых данных, указанных на гравюрах, являющихся одним из главных источников информации об объектах разрабатываемой онтологии. Текстовые данные являются источником не только сведений о самой гравюре: авторах рисунка или живописного произведения — прототипа гравюры, автора гравюры, технике создания прототипа гравюры и самой гравюры, имени и адресе издателя тиража, месте и времени печати листа, номере листа в серии, если это предусмотрено; иногда и наименовании серии, которое печаталось на первом листе серии гравюр, дают нам широкий круг сведений об изображенном объекте, его местонахождении, владельце или заказчике гравюры (посвящение), его звании или титуле (герб), расширяя контекстное поле, подчеркивая важность связей между объектом и его окружением.

Для определения всех информационных составляющих на гравюре был создан класс *НадписиПодписиЗнаки*, включающий все встречающиеся на гравюрах виды текстов и обозначений (рис. 3). Важной особенностью текстовых данных является возможность их фиксации и транскрипции. Для включения этих данных в онтологию был использован текстовый тип данных, однако для точной фиксации важно и добавление изображений, позволяющих изучить и анализировать шрифт, особенности написания, пунктуации и пр.

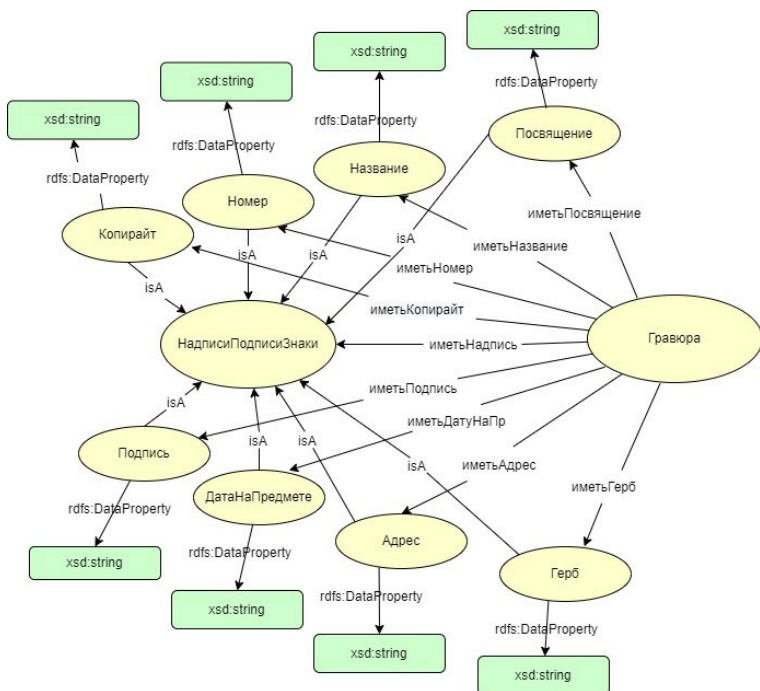


Рис. 3. Разделение класса НадписиПодписиЗнаки на подклассы

Основное преимущество онтологии заключается в возможности расширения ее структуры по мере обнаружения и интеграции новых сведений о личностях или предметах [Parundekar, Knoblock, Ambite 2010]. Расширение разрабатываемой онтологии подразумевается в основном в процессе добавления новых данных. Выбранная методология поддерживает эту возможность за счет своей итеративной структуры. Однако на данный момент никак не продумана возможность массовой загрузки данных. Этот этап станет возможным после разметки онтологии с концептуальной моделью высокого уровня CIDOC CRM. Такая разметка позволит не только однозначно определить типы данных для многих сущностей, но и использовать разработанные в других моделях, размеченных с CIDOC CRM, инструменты преобразования, интерпретации и массовой загрузки данных [DOREMUS 2018; ArCo 2019].

Значимое преимущество онтологических моделей — отсутствие каких-либо ограничений по отражению типа связей. Онтология не ограничивает, в отличие от реляционных баз данных, ни качество, ни количество этих связей. Для тех связей, которые имеют симметричную структуру, обратное имя свойства может быть приведено в круглых скобках и разработано таким образом, чтобы быть семантически значимым и грамматически правильным при чтении из диапазона в домен. Такой принцип помогает избежать создания лишних связей, между двумя конкретными классами. Этот способ формирования связей принят в больших концептуальных моделях верхнего уровня (CIDOC CRM).

Еще одна важная особенность, которую следует учитывать при формировании связей — их специфичность. Каждое свойство имеет определенное значение. Если значение свойства (даже незначительно) различается для разных классов, следует создать два (или более) свойства, например, в нашем случае такими связями являются *хранитьсяВ* и *находитьсяВ*. Первое применяется для обозначения места нахождения и хранения предмета искусства в определенном учреждении памяти, принадлежности ему; второе предназначено для обозначения физического нахождения объекта в административных границах географического региона.

Техническая реализация велась в редакторе онтологического моделирования с открытым доступом Protégé, который поддерживает разработку, хранение и экспорт онтологий в различных форматах, причем, как концептуальной ее части, так и данных, содержащихся в онтологии. Редактор позволяет выравнять разработанную модель с концептуальными моделями верхнего уровня, например, такими как CIDOC CRM [CIDOC CRM]. Подобные справочные ресурсы используются учреждениями культуры для представления и публикации коллекций и поддерживаются специальным комитетом при Международном совете музеев (ICOM) [Doerr 2003]. Кроме того, в редакторе предусмотрены механизмы для осуществления выравнивания онтологий и связывания сущностей с любыми контролируемыми словарями, тезаурусами и таксономиями, обеспечивающее терминологическую валидацию и унификацию. Разрабатываемая онтологическая модель, проектировалась с учетом и вниманием к структуре концептуальной модели верхнего уровня CIDOC CRM и ее расширений для описания библиотечных данных LRMOO (FRBRoo) [Le Boeuf 2012; Le Boeuf 2015]. Были изучены и рассмотрены существующие модели описания разнообразных коллекций: библиотечных материалов [Turcan-Verkerk

2020], нот [DOREMUS 2018], предметов фарфора [Wei 2020], посвященные специальным предметным областям [ArCo 2019; Icon 2023] и выровненные с CIDOC CRM.

Рассмотренные нами модели имеют свои специфические задачи и поэтому не могут быть использованы в аутентичном виде для концептуализации выбранной предметной области. Например, онтология DOREMUS служит для связывания метаданных однотипных предметов из разных институций, имеющих различные типы описания, а, следовательно, разный набор и полноту метаданных и поэтому направлена на преобразование и унификацию данных. При проектировании же нашей онтологии учитывался как опыт моделирования существующих моделей, так и специфика данных, для которых проектировалась онтология и задачей ее является определение и расширение контекстного поля за счет сведений о различных предметах. В конечном итоге разработанная модель предусматривает разметку с элементами CIDOC CRM [Bruseker, Carboni, Guillem 2017; Scharffe, Euzenat 2011] что является следующей стадией в плане работы над онтологией.

На данном этапе разработки осуществлено выравнивание классов и свойств с известными контролируемыми словарями, таксономиями и тезаурусами. Это сделано с целью унификации терминологии концептуальной части онтологии [Fonseca, Martin 2007]. Для выравнивания были выбраны общие тезаурусы, используемые в сфере архитектуры и искусства (Тезаурус по Архитектуре и искусству Исследовательского института Гетти ААТ). Выбор тезауруса основан на анализе терминологии концептуальной части и свойств. Для окончательного принятия решения о выборе тезаурусов был произведен автоматический анализ с помощью библиотек Python⁵ на основании исследования, проведенного Центром цифровых гуманитарных наук и культурного наследия Австрийской академии наук (Austrian Centre of Digital Humanities and Cultural Heritage of the Austrian Academy of Sciences) [Vocabulary Comparison 2023]. Был произведен анализ терминологии различных тезаурусов, как общей тематики, так и предметной области и на основании полученных данных о выявленных пересечениях с терминологией разрабатываемой онтологии, выбор был сделан в пользу Тезауруса по Архитектуре и искусству — ААТ.

Кроме того, произведено связывание сущностей онтологии с контролируемыми словарями. Связывание данных применяется для обеспечения валидации данных и верификации имен, наименований объектов, а также для контроля правильности перевода

терминов с одного языка на другой. Поскольку большая часть сущностей онтологии имеет первоначально англоязычную природу, тогда как онтология проектировалась на русском языке, потребовалось обеспечение точности перевода не только в области имен персоналий и наименований памятников, географических объектов, но и общей терминологии. Помимо этого, связывание данных обеспечивает очень важную для дальнейшего расширения онтологии возможность обогащения данных. Большинство известных контролируемых словарей пополняются и обогащаются. Связывание сущностей позволяет автоматическими методами переносить сведения из контролируемых словарей в онтологию, обеспечивая тем самым, практически бесконечную возможность для обогащения и пополнения данных онтологии. Для консолидации терминов был использован модуль, встроенный в редактор для работы с данными OpenRefine [OpenRefine Reconciliation]. Настраиваемый модуль позволяет автоматизировать поиск и консолидацию имен и объектов, присваивая сущностям соответствующий ID и ссылку на объект контролируемого словаря или тезауруса. Модуль позволяет производить поиск по всем четырем тезаурусам института Гетти, по объектам Wikidata, тезаурусу Geonames. Кроме того, настройки OpenRefine позволяют использовать модуль для связывания с другими тезаурусами через открытый API, если он предоставляется ресурсом. Более того, с помощью OpenRefine доступно автоматическое пополнение контролируемых словарей в случае необходимости, то есть данные, представленные в исследовательском проекте могут быть комплексно загружены в Wikidata со ссылкой на источник информации.

По итогам проведенного связывания все сущности онтологии получили те или иные ссылки на сущности контролируемых словарей (рис. 2).

Дальнейшая работа по наполнению модели и донстройке ее структуры может быть произведена после определения объема данных, подлежащих загрузке в онтологию, и их анализа. Эта работа требует вовлечения специалистов предметных областей и их взаимодействия по широкому спектру задач для поиска оптимальных решений концептуализации каждого вида объектов и определения связей, наиболее широко охватывающих все контекстное поле. Особенно необходимо уточнение терминологии как области прикладного искусства, так и архитектуры. Также планируется выравнивание онтологии с концептуальной моделью CIDOC CRM, которое позволит автоматизировать наполнение он-

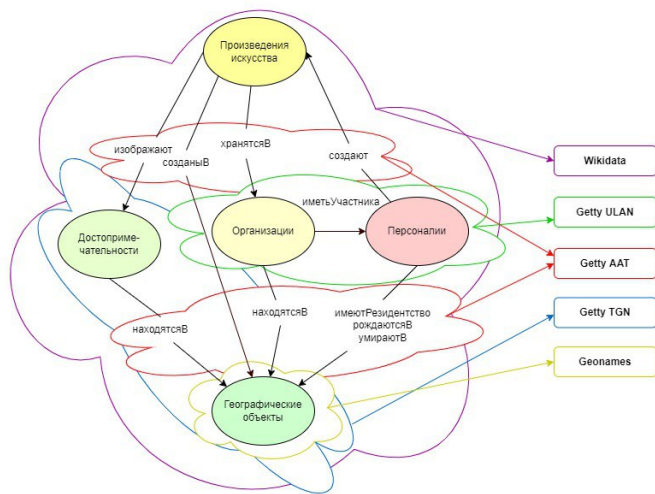


Рис. 4. Схема связывания сущностей онтологии

тологии данными, из различных институций, уже использующих эту модель для формализации и публикации своих объектов [LUX]. Наиболее сложно достижимым сегментом в дальнейшей работе является разработка системы подсчета точки съемки для дополнения геоинформационных данных расположения архитектурных и ландшафтных объектов, изображенных на гравюрах. В результате созданная модель впервые объединила данные о произведениях изобразительного и прикладного искусства, а также архитектурных объектах, связанных феноменом заимствования сюжетов. Предметы, хранящиеся в разнообразных учреждениях культуры различных стран и являющиеся национальным достоянием этих стран впервые объединены в онтологии, отражающей физические характеристики, иконографию коллекции изобразительных предметов, предметов прикладного искусства и архитектурных

объектов, включая их географическую привязку⁶. Применение таких моделей уже признано мировым сообществом специалистов и имеет значительную ценность для публикации музейных коллекций, объединения данных из разных учреждений и создания тематических цифровых платформ. Ценность таких моделей заключается в их гибкости и возможности расширения и признана мировым сообществом специалистов, занимающихся сохранением, представлением и распространением цифровых данных о культуре. Разработанная онтология может быть адаптирована для изучения других культурных феноменов, связывающих различные виды искусства.

Примечания

- ¹ В сокращенном виде отдельные положения настоящей статьи обсуждались на XIX конференции Ассоциации «История и компьютер». Тезисы доклада опубликованы в *Материалы Международной научной конференции «Современная историческая информатика: аналитика данных в исторических исследованиях» и XIX конференции Ассоциации «История и компьютер»*. М., 2024. С. 120–121.
- ² Имеется в виду иконографически-иконологический метод Эрвина Панофского (Erwin Panofsky, 1892–1968). Панофский следовал подходу, впервые примененному Варбургом (Abraham Moritz Warburg 1866–1929) в своих исследованиях, которые были направлены на осмысление художественных сюжетов и мотивов как свидетелей социокультурных феноменов. В своей работе Панофский изложил метод прочтения произведения искусства, который требовал различения произведения искусства на трех уровнях: Первичный, или естественный, предмет, который идентифицирует чистые формы, такие как конфигурация линий или изображения объекта, которые можно было бы назвать миром художественных мотивов. Перечисление этих мотивов относится к предыконографическому описанию произведения искусства. Второстепенный, или конвенциональный, предмет — это присвоение темы и концепции композиции художественных мотивов. Мотивы, воспринимаемые как носители вторичных, или условных, смыслов, можно назвать образами, а сочетание образов — это истории или аллегории. 15 Сюжеты изображения идентифицируются на этом уровне благодаря иконографическому анализу. Внутренний смысл или содержание — это интерпретация произведения искусства как признака чего-то другого, что выражает себя в бесчисленном множестве других признаков, и мы интерпретируем его композиционные и иконографические особенности как более конкретное свидетельство этого чего-то другого. Внутренний смысл определяется тем, как культурно-исторические события

отражаются в репрезентации, и этот смысл проявляется независимо от воли художника, который может его совершенно не осознавать. Панофский назвал этот этап иконологической интерпретацией.

³ Выражаю искреннюю благодарность Ирине Ефименко за научное руководство, ценные замечания и корректировку модели, а также за всестороннюю поддержку в ходе работы.

⁴ Код доступен по ссылке URL: https://colab.research.google.com/drive/1_LQ_M4tVtiFSJKkqb5OXK3sIG3WgGPO.

⁵ Код доступен по ссылке URL: <https://colab.research.google.com/drive/1cD-Jl6UX8fuQnRPieWmnSuXScSV3OAMY>.

⁶ Репозиторий проекта доступен по ссылке Landscape Engravings Ontology / Github.com. URL: <https://github.com/AMitiukova/LandscapeEngravingsOntology>

Литература

Исследования

Бардовская, Ботт 2019 — Бардовская Л. В., Ботт И. К. Английская видовая гравюра в Царскосельском собрании Екатерины Великой // Английский вкус императрицы: Царское Село Екатерины Великой. СПб: Русская коллекция, 2019. С. 121–134.

Воронихина 1988 — Воронихина Л. Н. О пейзажах сервиза с зеленой лягушкой / Музей-9. Художественные собрания СССР. М., 1988. С. 166–174.

Жлобинская 2013 — Жлобинская О. Н. Принципы и подходы к совмещению представления и доступа к библиотечным, архивным и музейным ресурсам: сборник научно-методических материалов рабочей группы Президентской библиотеки / [Жлобинская О. и др.; под общ. ред. Жабко Е. Д.]; ФГБУ «Президентская б-ка им. Б. Н. Ельцина». Санкт-Петербург: Президентская б-ка им. Б. Н. Ельцина, 2013. 505 с.

Иванов 2007 — Иванов В. В. Онтологический подход к созданию информационной системы по культурному наследию // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. 2007. № 2. С. 73–91.

Панофский 1999 — Панофский, Э. Смысл и толкование изобразительного искусства. Статьи по истории искусства. Санкт-Петербург, 1999. С. 45–49.

Спецификация API — Спецификация API / Портал открытых данных министерства культуры Российской Федерации. URL: <https://opendata.mkrf.ru/item/api>

Юмашева 2023 — Юмашева Ю. Ю. К вопросу о возможности объединения информационных ресурсов архивов, библиотек и музеев Российской Федерации. // История и архивы, 2023. 5(2). С. 65–83. DOI: 10.28995/2658-6541-2023-5-2-65-83.

ArCo 2019 — Carriero V. A., Gangemi A., Mancinelli M. L., Marinucci L., Nuzzolese A. G., Presutti V., Veninata C. ArCo ontology network and LOD on Italian cultural heritage // CEUR WORKSHOP PROCEEDINGS. CEUR-WS, 2019. Vol. 2375. P. 97–102. URL: <https://ceur-ws.org/Vol-2375/short3.pdf>.

Barbot 2023 — Barbot L., Raciti M., Ďurčo M. Use of vocabularies for metadata curation and quality assessment in Social Sciences and Humanities. Version 1.0.0 // DARIAH Campus [Event]. 2023. URL: <https://hdl.handle.net/21.11159/019595c3-501a-713f-b689-7de221fe9020>

Bruseker, Carboni, Guillem 2017 — Bruseker G., Carboni N., Guillem A. Cultural heritage data management: the role of formal ontology and CIDOC CRM // Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data, 2017. P. 93–131. DOI: 10.1007/978-3-319 65370-9_6.

CIDOC CRM — CIDOC CRM / Resources. URL: https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.4.pdf.

Doerr 2003 — Doerr M. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. // AI magazine, 2003. № 24. P.75–92. DOI: 10.1609/aimag.v24i3.1720

DOREMUS 2018 — Achichi M., Lisena P., Todorov K., Troncy R., Delahousse J. DOREMUS: A graph of linked musical works. / The Semantic Web–ISWC 2018: 17th International Semantic Web Conference. October 8–12, // Proceedings. Part II 17. Monterey. CA. USA., 2018. P. 3–19. URL: <https://www.eurecom.edu/publication/5565/download/data-publi-5565.pdf>.

Fonseca, Martin 2007 — Fonseca F., Martin J. Learning the differences between ontologies and conceptual schemas through ontology-driven information systems // Journal of the Association for Information Systems, 2007. № 8. P.4. DOI: 10.17705/1jais.00114.

Geonames — Geonames <https://www.geonames.org/>

Icon 2023 — Sartini B., Baroncini S., van Erp M., Tomasi F., Gangemi A. Icon: An ontology for comprehensive artistic interpretations. // ACM Journal on Computing and Cultural Heritage, 2023. № 16. P.1–38. DOI: 10.1145/3594724.

Le Boef 2015 — Le Boef P., A basic introduction to FRBRoo and PRESSoo, 2015. URL: <https://library.ifla.org/id/eprint/1150/>

Le Boeuf 2012 — Le Boeuf P. Modeling rare and unique documents: using FRBROO/CIDOC CRM. // Journal of Archival Organization, 2012. T.10. № 2. P. 96–106. DOI: 10.1080/15332748.2012.709164.

LUX — LUX: Yale collection discovery. URL: <https://lux.collections.yale.edu/>

OpenRefine Reconciliation — Getty Vocabularies OpenRefine Reconciliation / The Getty Institute. URL: <https://www.getty.edu/research/tools/vocabularies/obtain/openrefine.html>

Parundekar, Knoblock, Ambite 2010 — Parundekar R., Knoblock C. A., Ambite J. L. Linking and building ontologies of linked data. // The Semantic Web–ISWC 2010: 9th International Semantic Web Conference. ISWC 2010. — Shanghai, China. November 7–11, 2010. Revised Selected Papers. Part I. 9, 2010. P. 598–614. URL: https://link.springer.com/chapter/10.1007/978--3-642--17746--0_38.

Peroni 2016 — Peroni S. SAMOD: an agile methodology for the development of ontologies. // Proceedings of the 13th OWL: Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016), 2016. P. 1–14. URL: <http://www.essepuntato.it/papers/samod.pdf>.

Raeburn, Voronikhina, Nurnberg 1995 — Raeburn M., Voronikhina L. N., Nurnberg A. The Green Frog Service. London: Cocklegoose Press, 1995. 424 p.

Scharffe, Euzenat 2011 — Scharffe F., Euzenat J. Linked data meets ontology matching: enhancing data linking through ontology alignments. // Proc. 3rd international conference on Knowledge engineering and ontology development (KEOD), 2011. October. P. 279–284. URL: <https://www.scitepress.org/PublishedPapers/2011/36660/36660.pdf>.

Stijnman 2012 — Stijnman A. Engraving and Etching 1400–2000: A History of the Development of Manual Intaglio Printmaking Processes. London: Archetype Publications, 2012. 484 p.

Turcan-Verkerk 2020 — Turcan-Verkerk A.-M. Livre blanc Bibliissima+ / Observatoire des cultures écrites de l'argile à l'imprimé, 2020. Octobre. Zenodo. DOI: 10.5281/zenodo.6611722.

Vocabulary Comparison 2023 — Rastinger N. C., Carloni M., Illmayer K., Ďurčo M. «Vocabulary Comparison – Jupyter Notebook». Zenodo, 19 апреля 2023 г. DOI: 10.5281/zenodo.7845914.

Wei 2020 — Wei T. Terminology and ontology for cultural heritage: application to Chinese ceramic vessels. Diss. Université Grenoble Alpes, 2020. URL: <https://theses.hal.science/tel-03167916/>

Williamson 1909 — Williamson G. C. The imperial Russian dinner service: a story of a famous work by Josiah Wedgwood. London, 1909. 464 p.

Ксения Анисимова

АНАЛИЗ ТОНАЛЬНОСТИ РУССКОЙ ДРАМЫ XVIII–XX ВВ. КАК ИНСТРУМЕНТ МОДЕЛИРОВАНИЯ ХУДОЖЕСТВЕННОЙ СТРУКТУРЫ

Исследование посвящено описанию эмоциональной динамики как проявления художественной структуры русской драмы XVIII–XX вв. на основе автоматической разметки тональности реплик с использованием нейросетевых моделей BERT-архитектуры. Такие модели, дообученные даже на нехудожественных текстах, показывают удовлетворительные результаты при анализе тональности драматических реплик, что было проверено на ручную размеченной тестовой выборке. На основе такой автоматической эмоциональной разметки было показано, что динамика негативной тональности реплик совпадает с ключевыми этапами развития драматического действия, такими как завязка, кульминация и развязка. Автоматический анализ тональности можно назвать продуктивным инструментом моделирования художественной структуры: значимые изменения хронологической динамики на ключевые моменты развития классицистической формы, перехода к реализму, а от него — к модернизму, смену направления основного конфликта от внешнего к внутреннему и перехода от замкнутых драматических структур к открытым.

Ключевые слова: русская драма XVIII–XX вв., анализ тональности, эмоциональная динамика, художественная структура, BERT-модели, драматические структуры, BERT-модели

Введение

Ключевой элемент любого драматургического текста — реплика. Реплики — это высказывания, принадлежащие персонажам, прямая речь, части монологов и диалогов; то, что реализует продвижение сюжета и в немалой степени составляет его.

Чувства и эмоции, содержащиеся в репликах, с упрощениями, можно свести до тональностей, которые часто выражаются в дихотомических позитивно-негативных шкалах и которые извлекаются из текста автоматическими методами. Намекает на справедливость подобного упрощения, например, жанровое деление драмы, которое сводится к не менее черно-белым понятиям: трагедиям, комедиям и трагикомедиям.

Для такого рода литературы описание эмоциональной динамики, основанное на процентах реплик разных тональностей (прежде всего позитивной и негативной), может позволить отследить некоторые особенности художественной структуры драмы. На значимость эмоционального аспекта в тексте драмы указывали, например, Ярхо, Выготский, Волькенштейн [Волькенштейн 1960; Выготский, 1998; Ярхо, 2006].

Под таким углом имеет смысл рассмотреть русскую драму XVIII–XX вв., эмоциональный аспект текстов которой пристально еще не описывался.

Отдельный интерес вызывает применение инструментов анализа тональности к художественным текстам на русском языке. В немногочисленных исследованиях, где ставится такого рода эксперимент и при этом проводится формальная оценка результатов, отмечается, что согласованность ручной и автоматической разметки низкая [Sherstinova 2023]. Еще более малочисленными оказываются исследования, где для такой задачи использовались бы языковые модели архитектуры BERT. Примечателен этот факт в особенности потому, что для анализа тональности русскоязычных нехудожественных текстов эти модели давно и широко используются и показывают высокие результаты в рамках формальных оценок [Smetanin & Komarov 2021].

Использование BERT моделей для анализа тональности реплик русской драмы позволит оценить применимость этого инструмента для более широкой задачи анализа тональности художественного текста на русском языке, и для этого потребуется провести формальную оценку результатов. В случае удовлетворительного перформанса модели, — может получиться уловить те особенности

художественной структуры русской драмы, которые отражаются в ее эмоциональной динамике. Описать такие особенности и описывается целью описываемого эксперимента.

Эмоциональный аспект может оказаться жанрообразующим признаком, а конкретные показатели процентов тональностей окажутся количественным выражением таких различий. Анализ эмоциональной динамики в хронологической перспективе позволит оценить, насколько важным для художественной структуры драмы оказывается выражение персонажами эмоций и, следовательно, степень экспрессивности. На уровне композиции процент реплик, отнесенных моделью к значимым классам, позитивному или негативному, может оказаться сопоставим с общепринятыми этапами развития сюжетного действия: завязка, кульминация, развязка.

Корпус

Проект *DraCor* предоставляет доступ к корпусу русской драмы XVIII–XX вв. — *Russian Drama Corpus* или *RusDraCor*, который в этом исследовании использовался как выборка из довольно метафорической совокупности русской драмы [Fischer et al. 2019].

В корпус входят 212 пьес, размеченных в формате TEI. Самая ранняя, «Хорев» Сумарокова, относится к 1747 г., а самая поздняя, «Остров мира» Петрова, к 1947 г. В среднем на каждое десятилетие приходится 10 пьес, а единственное десятилетие, которое оказывается выбросом — 1850–ые, к этой декаде относятся 28 произведений (см. рис. 1).

В корпусе представлены пьесы 56 авторов. Наибольшее количество произведений в *RusDraCor* принадлежит Островскому — 38 пьес, по 14 пьес у Сумарокова и Чехова, а в среднем на каждого автора приходится по 4 пьесы.

Для 118 пьес, т. е. для чуть больше, чем половины корпуса русской драмы, указан “нормализованный” жанр¹ — комедия («Comedy») или трагедия («Tragedy»). Комедий в *RusDraCor* насчитывается 89, а трагедий всего 29, для других 94 пьес жанр не указан (см. рис. 1). Уже на моменте описания корпуса обнаруживается интересная деталь: комедии сильно преобладают над трагедиями.

Другой важный признак пьесы — количество актов. Больше всего в *RusDraCor* 5-актных пьес (27 %), затем идут 4-актные (18 %) и 3-актные (16 %). Одноактные пьесы и пьесы без указания на наличие актов при описании эмоциональной динамики объединились в одну категорию.

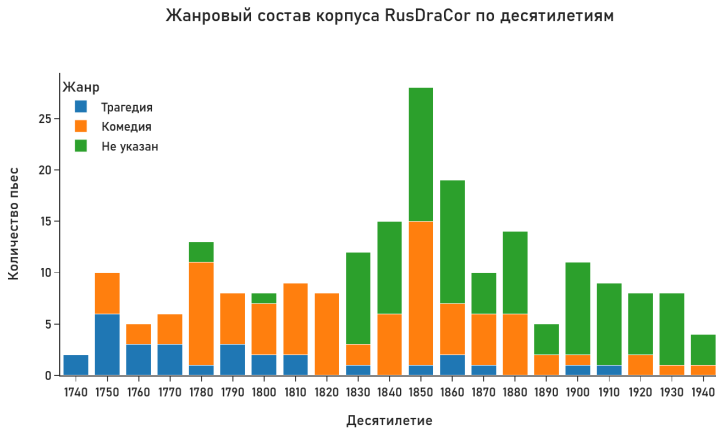


Рис. 1. Жанровое и хронологическое распределение корпуса RusDraCor

Из XML-файлов с пьесами были извлечены реплики с помощью языка программирования Python². Общее количество реплик составило около 117 тысяч³.

Критерии выбора BERT-моделей

Размеченных данных, на которых можно было бы дообучить модель архитектуры BERT для задачи эмоциональной разметки русской драмы, нет. Однако для анализа тональности текстов на русском языке существуют модели, дообученные на текстах из социальных сетей [Smetanin 2020]. С одной стороны, такие тексты неформальны, не обязаны быть последовательными, стилистически однородными и следовать языковым нормам; а с другой — тексты из социальных сетей имеют схожую с драматическими репликами природу, многое из того, что обозначают как «посты» в социальных сетях оказывается высказыванием от лица того или иного пользователя: комментарии, ответы и т. п. В том числе по этой причине есть основания полагать, что результаты применения таких моделей будут удовлетворительными, так что первым этапом в попытке оценить эмоциональную динамику русской драмы оказывается подбор уже готовой к использованию модели.

При выборе моделей для анализа тональности в первую очередь следует ориентироваться на метрики качества работы модели, чем они выше — тем лучше. Анализ тональности с BERT-моделями — это задача классификации, поэтому следует ориентироваться на метрики качества классификации.

Вторым критерием отбора модели становится набор выделяемых классов. Ограничиться двумя полярными классами означало бы получение крайне зашумленного результата — едва ли все или большинство реплик, пусть и драматических, можно отнести к однозначно позитивным и однозначно негативным, поэтому минимально допустимое число классов — три. Возможность модели выделить более чем три класса также отнесем к преимуществу, т.к. это может с одной стороны детализировать описание эмоциональной динамики русской драмы, если эти классы представляют собой непосредственное обозначение эмоций, а с другой — снизить зашумленность результатов, если среди этих классов будут такие, куда попадают неоднозначные варианты.

Кроме этого, модель должна быть описана в научной литературе и/или её использование — стандартная или популярная практика для специалистов, работающих с анализом тональности.

Выбранные модели

Под эти критерии подходит модель XLM-RoBERTa-Large, дообученная на датасете RuSentiment [Liu et al 2019; Rogers et al 2018; Smetanin & Komarov 2021]. В рамках исследования Комарова и Сметанина было показано, что эта модель показывает лучшие результаты для мультиклассовой классификации — F1-мера составляет 75.27 % [Smetanin & Komarov, 2021]. Модель способна определять 5 тональностей: позитивную («POSITIVE»), негативную («NEGATIVE»), нейтральную («NEUTRAL»), неясную («SKIP») и класс формул вежливости («речевых актов», «SPEECH ACT»).

Другая модель RuBERT-Based-Cased-Conversational, встроена в фреймворк DeepPavlov⁴, который предоставляет доступ к различным моделям и инструментам обработки естественного языка. Эта модель выделяет три класса, сообразно многим другим BERT-моделям для анализа тональности: позитивный («POSITIVE»), негативный («NEGATIVE»), нейтральный («NEUTRAL»). Согласно документации DeepPavlov, эта модель (но в предобученном виде) работает для классификации лучше остальных моделей, доступных в этом фреймворке⁵.

Методика формальной оценки автоматического анализа тональности русской драмы

Для оценки работы моделей использовались стандартные метрики качества классификации. Среди них прежде всего оценивались: точность выделения каждого класса (precision), полнота (recall), и взвешенная f1-мера, так как выбранные модели реализуют мультиклассовую классификацию, а самыми информативными классами в случае анализа тональности драмы оказываются позитивный и негативный, поэтому следует отдельно смотреть на метрики каждого класса.

В случае с драмой принципиально важным оказывается способность модели отличать негативный класс от позитивного и наоборот.

Метрики эти представляют собой результат сравнения автоматической и ручной разметки. Как уже было сказано, размеченных для анализа русских пьес в открытом доступе нет, поэтому для валидации результатов была размечена тестовая выборка пьес из RusDraCor.

Из всего корпуса автором исследования было размечено 10 пьес, что составляет около 5 % от общего числа произведений в нем. Такой объем вручную размеченных данных оказывается достаточным для *предварительной* оценки качества работы моделей. Для формирования репрезентативной выборки сформулированы следующие критерии подбора пьес для разметки:

- Отбирается по одной пьесе из каждого второго десятилетия (1740-е, 1760-е, 1780-е и т. д.).
- Распределение количества актов и жанров отобранных 10 пьес должно примерно совпадать с распределением этих признаков на всем корпусе.
- Авторы в выборке не должны повторяться и быть одними из самых частотных для этого десятилетия.

Список пьес, вошедших в выборку, представлен в таблице 1.

Название	Автор	Год	Акты	Жанр
Гамлет	Сумароков	1748	5	Трагедия
Корион	Фонвизин	1764	3	Комедия
Из жизни Рюрика	Екатерина II	1786	5	–
Урок дочкам	Крылов	1807	1	Комедия
Поездка в Кронштадт	Писарев	1823	4	Комедия
Шила в мешке не утаишь — девушки под замком не удержишь	Некрасов	1841	2	Комедия
Горячее сердце	Островский	1869	5	Комедия
Татьяна Репина	Чехов	1889	1	–
Балаганчик	Блок	1906	1	–
Противогазы	Третьяков	1924	3	–

Таблица 1. Тестовая выборка пьес корпуса RusDraCor, размеченная вручную, и их признаки

Выбор тональности реплики в процессе ручной разметки производился в соответствии с правилами разметки корпуса RuSentiment, на котором дообучались обе представленные выше BERT-модели [Rogers et al 2018]⁶. Пример разметки представлен в таблице 2. Для оценки качества модели RuBERT от DeepPavlov требуется разметка по трем классам, поэтому для преобразования пяти классов в три из ручной разметки удалялись реплики, которые вручную были размечены как неясные («SKIP») и («SPEECH»), что *могло завести результаты метрик качества работы этой модели.*

Реплика	Тональность	Пьеса	Автор	Год
Дед ваш Гостомысл скончался.	NEGATIVE	Из жизни Рюрика	Екатерина II	1786
Я думал лишь о ней, когда поехал с балу.	POSITIVE	Путешествие в Кронштадт	Писарев	1823
Ты когда мне кумом-то был? Когда у меня капиталу не было.	NEUTRAL	Горячее сердце	Островский	1869
Господи, помилуй!	SKIP	Татьяна Репина	Чехов	1889
Простите нас, сударь!	SPEECH	Урок дочкам	Крылов	1807

Таблица 2. Пример разметки реплик

*Результаты формальной оценки автоматического анализа
тональности русской драмы*

Метрики качества классификации XLM-RoBERTa-Large представлены в таблице 3. Взвешенная f1-мера составляет 73 %, что для задачи анализа тональности оказывается хорошим результатом. Для значимых позитивного и негативного классов реплик показатель точности можно назвать высоким: в более чем 72 % процентов случаев модель правильно определяет эти классы. То же самое нельзя сказать о полноте, которая для позитивного класса составила лишь 51 % — и это оказывается порогом, ниже которого результаты было бы сложно принять; для негативного класса полнота чуть выше и составляет 63 %.

	precision	recall	f1-score	support
NEGATIVE	0.78	0.63	0.70	783
NEUTRAL	0.79	0.90	0.84	1906
POSITIVE	0.72	0.51	0.60	277
SKIP	0.57	0.54	0.55	735
SPEECH	0.67	0.49	0.56	45
accuracy	0.74	3746		
macro avg	0.70	0.53	0.65	3746
weighted avg	0.74	0.64	0.73	3746

Таблица 3. Метрики качества классификации XLM-RoBERTa-Large

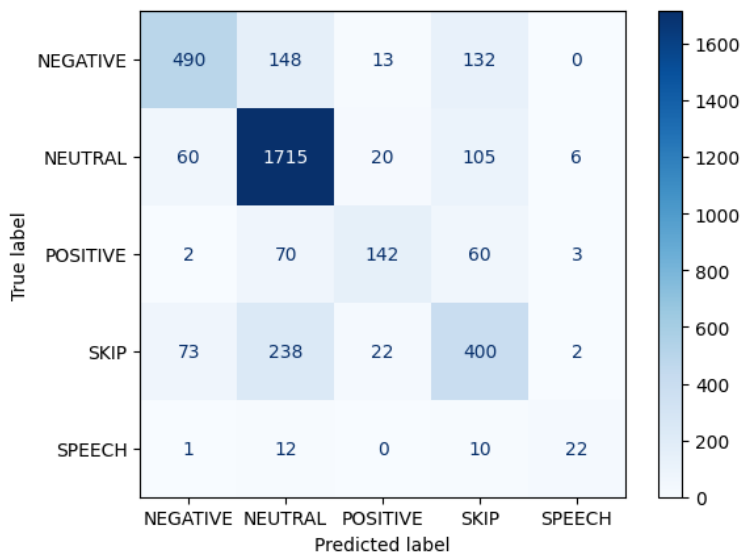


Рис. 2. Матрица ошибок модели XLM-RoBERTa-Large

На матрице ошибок XLM-RoBERTa-Large (рисунок 2) видно, что невысокие показатели полноты для позитивных и негативных классов связаны с тем, что реплики из них модель часто относит к нейтральному и неясному классам, что до определенной степени допустимо. Принципиально важным здесь оказывается то, что модель почти не относит негативные реплики к позитивным и наоборот — на 3746 реплик она допускает всего 15 таких ошибок, что говорит о высоком качестве работы этой модели для двух значимых классов.

Модель RuBERT-Base-Cased-Conversational из пакета DeepPavlov, дообученная на нескольких русскоязычных датасетах для анализа тональности, показала общую точность в 61 %. Для этой модели точность выделения негативного и позитивного классов составляет 42 % и 47 % соответственно. Эти классы модель выделяет случайно. Отдельно стоит отметить крайне низкую полноту у позитивного класса — всего 38 %.

Из матрицы ошибок этой модели (рисунок 3) видно, что она называет позитивные реплики негативными почти так же часто,

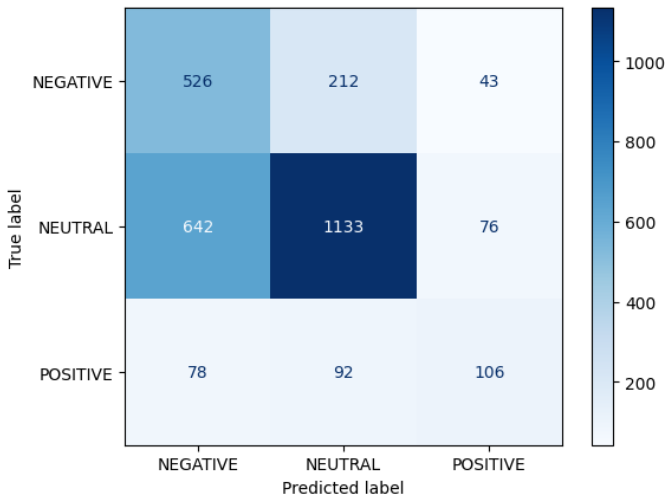


Рис. 3. Матрица ошибок модели RuBERT-Base-Cased-Conversational

как и правильно определяет их как позитивные, что *нельзя* назвать удовлетворительным результатом.

Из двух выбранных для эксперимента моделей приемлемые результаты показала только XLM-RoBERTa-Large, поэтому описание эмоциональной динамики строится на ее результатах.

Распределение тональностей реплик

Описание эмоциональных тенденций русской драмы будем основывать на значениях процентного соотношения реплик тональностей: для каждого жанра, десятилетия, автора, акта или для сочетания этих признаков.

Самый распространённая тональность — нейтральная: нейтральные реплики составляют более половины (58 %). Процент негативных — около 18 %, что делает этот класс вторым по численности. Общая доля позитивных реплик составляет всего 6 % от всех высказываний. Полное распределение тональностей представлено в таблице 4.

Ранг	Класс	Кол-во	% от общего числа
1	NEUTRAL	68253	58%

2	NEGATIVE	21595	18%
3	SKIP	18367	17%
4	POSITIVE	7150	6%
5	SPEECH	1677	1%
Итого (всего реплик)	117042	100%	

Таблица 4. Количество и доля реплик разных тональностей в датасете реплик на основе корпуса RusDraCor

Класс «SPEECH» оказывается слишком мало представленным, а класс «SKIP» аккумулирует в себе крайне разнородные реплики; поэтому показатели этих классов практически исключены из дальнейшего анализа, но включены в таблицы.

Эмоциональная динамика как проекция особенностей поэтики драматургических жанров

При детальном рассмотрении процентов реплик разной тональностей в разных жанрах, заметен большой разрыв в количестве нейтральных реплик в комедиях и трагедиях. Для комедий процент нейтральных реплик примерно совпадает с процентом нейтральных реплик по всему корпусу, а для трагедий она меньше на 18,7 % процентов. Доли остальных тональностей этих двух жанров отличаются не так сильно, за исключением неясного класса «SKIP», в трагедиях на 16–17 % больше реплик неясной тональности, чем в комедиях и в пьесах без указания жанра. Можно предположить, что доля нейтральных реплик в трагедиях уменьшилась прежде всего за счет обилия реплик, относящихся к классу SKIP, что соотносится с данными, представленными в матрице ошибок работы XLM-RoBERTa-Large (см. рис. 2).

Жанр \ Тональность в %	NEUTRAL	NEGATIVE	SKIP	POSITIVE	SPEECH
Комедия	59,8	17,2	15	6,6	1,4
Трагедия	47,5	21,4	26,5	3,8	0,8
Не указан	59	19,3	14	6	1,6

Таблица 5. Процентное соотношение реплик разной тональности по жанрам

Модель, результаты работы которой анализируются здесь, была обучена на современных текстах, которые в языковом плане отличаются от языка трагедий, представленных в RusDraCor. Например, лексика трагедий этого периода, второй половины XVIII-го века, могли привести к увеличению числа реплик неясной тональности.

Доля негативных реплик в трагедиях выше на 4,2 %, чем в комедиях, что при учете среднего удельного веса этого класса во всем корпусе оказывается заметным отличием (см. таб. 4, 5). Примерно также можно охарактеризовать долю позитивных реплик в трагедиях (3,8 %), которая оказывается почти в два раза ниже, чем в комедиях (6,6 %).

Разница в показателях между комедиями и трагедиями оказывается достаточно заметной, чтобы сделать предположение о том, что эмоциональный аспект, выражающийся в высказываниях персонажей, рассмотренный как признак, которым обладает текст или высказывание, может работать как «жанровый индикатор», т. е. по эмоциональному распределению или эмоциональной динамике оказывается возможным отличить один жанр от другого, что может говорить о проекции жанровых драматических систем или правил на уровень поэтики.

Однако полученное на признаках жанра и тональности для каждой из 117 тыс. реплик значение статистики хи-квадрат ($\chi^2 = 1429,93$) при чрезвычайно низком p-value (p 0,001; точнее, $p \approx 1,9110^{-303}$) указывает на наличие статистически значимой связи. Реальная значимость этой связи, измеренная с помощью коэффициента V-Крамера (этот шаг необходим), оказывается практически отсутствующей (0,08).

Крайне высокий показатель хи-квадрат и одновременно низкий коэффициент Крамера говорят о наличии некоторого паттерна, который при этом вряд ли оказывается статистически значимым. На практике это означает невозможность различения пьес разных жанров по количеству реплик разной тональности.

В контексте стилистическом это может указывать на ограничения метода: выбранные BERT-модели вряд ли способны уловить сложные эмоции вроде иронии или тем более эмоцию, выраженную контекстуально, как это часто происходит в комедиях (с этим, вероятно, связана высокая доля негативных реплик, как в комедиях, так и в трагедиях, и в целом значительно более высокая доля негативных реплик по сравнению с позитивными). На такие результаты также могла повлиять невысокая полнота работы модели.

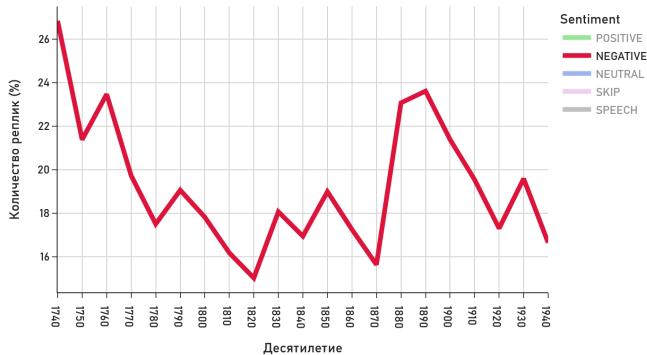


Рис. 4. Распределение количества негативных реплик по десятилетиям (%)

В таких условиях прямо говорить о выявлении какой-то жанровой эмоциональной проекции сложно, однако если отсутствие различия не связано с невысокой полнотой работы модели, схожесть эмоциональных динамик комедий и трагедий — занимательный предмет для отдельного филологического наблюдения.

Хронологическое распределение тональностей: открытые и замкнутые драматические структуры

Количественное описание эмоциональной динамики русской драмы в хронологической перспективе представляет собой модель, отражающую изменение экспрессивной составляющей драмы в контексте смены литературных тенденций и, соответственно, может оказаться проекцией структурных характеристик художественной системы драмы на поэтический уровень.

В этом контексте интерпретировать средние значения проблематично, особенно учитывая не слишком высокую полноту работы модели для негативных и позитивных классов, поэтому, помимо общей динамики, детально рассматривать будем те десятилетия негативных, позитивных и неясных тональностей, показатели которых отклоняются от среднего значения для каждой тональности более чем на одно стандартное отклонение ($\pm\sigma$).

Доли позитивной и негативной тональностей середины XVIII в. (1740–1760-е) выходят за пределы одного стандартного отклонения

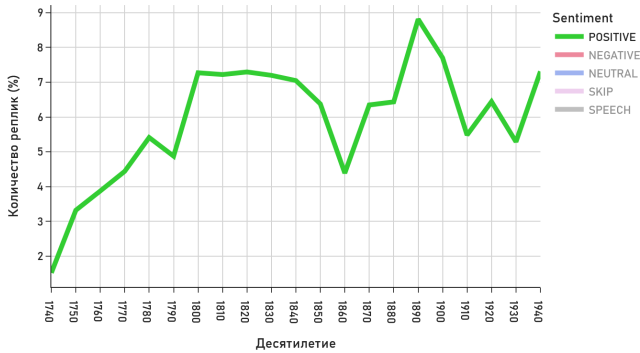


Рис. 5. Распределение количества позитивных реплик по десятилетиям (%)

(рисунки 2–3⁷). В случае с позитивной динамикой — в меньшую сторону, а в случае с негативной — в большую (кроме 1750-х). Это идейно сочетается с законами классицизма. Лаконичность классицистической композиции может правдоподобно выражаться в повышенном и более прямолинейном проявлении персонажами эмоций.

В динамике негативной тональности большими оказываются значения 1880–1890-х годов. Эти десятилетия представлены в корпусе преимущественно работами Чехова.

На 1880–ые гг. приходится 8 работ Чехова и 6 работ других авторов — 3 пьесы Островского и по одной пьесе Толстого Л. Мамина-Сибиряка и Пруткова. Из пьес Чехова этого десятилетия только «Татьяна Репина» имеет долю негативных реплик меньше среднего по всему корпусу. Из 14 пьес, у 9-ти процент негативных реплик выше среднего, самой негативной из всех оказывается пьеса «Трагик поневоле (Из дачной жизни)» Чехова с 40% негативных реплик. У остальных Чехова, кроме «Татьяны Репиной», процент негативных реплик составляет от 21% до 36%, что, опять же, оказывается выше среднего по корпусу и сопоставимо с долей негативных реплик в трагедиях середины XVIII в., хотя большинство чеховских пьес — комедии.

К 1890–ым относятся 4 пьесы Чехова и одна Мамина-Сибиряка. Наименее негативной пьесой этого десятилетия оказывается «Сва-

дьба» с 18,8 % негативных реплик, что уже немногим больше доли негативных реплик в целом. Самой большой процент негативных реплик (30,8 %) снова у пьесы Чехова — «Юбилей». Этот же период, 90-е годы XIX в., обращает на себя внимание в контексте позитивной динамики (рис. 4), на эти года приходится самая большая доля позитивных реплик — 8,81 %. Такое значение достигнуто также за счет преобладания пьес Чехова в этом периоде и за счет их «позитивности».

Чехов оказывается одновременно одним из самых негативных и самым позитивных автором всего корпуса.

Критики при жизни Чехова и после всегда говорили о некоем «настроении» чеховской драмы, и при этом будничности описываемых в ней ситуаций [Чудаков 1971]. Порой это «настроение» они описывали и в сопоставимых с терминологией этого исследования дихотомических положительно-отрицательных красках: «Его ум занимает отношение друг к другу двух мирозерцаний: одного — жизнерадостного, уравновешанного, полного веры и деятельного, другого — скептического, во всем сомневающегося, не видящего цели и смысла жизни...» [Полнер 1897]. А. Чудаков в книге «Поэтика Чехова» в этом же смысле отмечает, что Чехов в своих произведениях не отдает предпочтения бытовому или «возвышенным» ситуациям, но воспринимает мир целиком и притом не иерархично [Чудаков 1971]. Введение Чеховым бытовых разговоров и ситуаций в драматургическое действие, по-видимому, привело к некоей поэтической «полноте». Художественная система драматургии Чехова не экспрессивна напрямую, но из-за преодоления присущей более ранней драме ограничений, аккумулирует в себе те проявления драматургического конфликта, которые из текста предыдущего типа драмы были недостижимы. И хочется верить, что именно эта особенность чеховской драмы проявилась в результатах анализа тональности.

Первое десятилетие XX в. характеризует как большое количество негативных реплик, так и позитивных. Этот период представлен в RusDraCor разнообразно, пьесами следующих авторов: Блока (3 пьесы), Чехова (2 пьесы), Горького, Андреева, Мережковского, Л. Толстого а также Найдёнова, Бельского (1 пьеса). Самыми позитивными пьесами периода (11–12 %) оказываются пьесы Блока, Чехова и Андреева, а самыми негативными — Горького, Блока и Чехова. Блок, как представитель новой драмы и экспериментатор, также демонстрирует эмоциональную амбивалентность.

До сих пор при рассмотрении динамики негативной тональности мы обращались к десятилетиям, в которых доля негативных реплик заметно больше среднего. Теперь обратимся к тем из них, которые оказываются сильно меньше среднего — 1810-е и 1820-е гг. Этот период в корпусе представлен преимущественно комедиями таких авторов как Шаховской, Писарев и Хмельницкий. Комедии этого периода характеризует сильное французское влияние, в особенности Мольера, многие пьесы русских драматургов оказывались калькой французских пьес или по крайней мере были ими вдохновлены, одновременно эти пьесы зачастую высмеивали французоманию. Эти произведения подчинялись классицистическим канонам, хотя в этот момент уже происходил переход к сентиментализму (сентиментальные произведения в RusDraCor не представлены). Главным комедийным средством этого периода оказывалась сатира, высмеивались нравы высшего общества [Стенник 1982]. Настроения комедий того времени и языковая стилистика часто оказывались похожи на водевиль. В этом смысле неудивительным оказываются особенно маленькая доля негативных реплик.

Еще один период с выбивающейся низкой долей негативных реплик — 1870-ые. В RusDraCor 9 из 10 пьес, относящихся к этой декаде написаны Островским, его произведения являют собой реалистическую модель драмы, где, по Богдановой [Богданова 2024], человек полностью осознал себя как часть оформленного социума и основной конфликт построен на борьбе с ним. Замкнутость структуры драмы в России в момент Островского окончательное оформляется, в этот же момент русская драма окончательно становится самостоятельной и самобытной. В корпусе Островский оказывается самым представленным автором и, соответственно, из всех авторов на распределение тональностей реплик он повлиял больше всего. В такой ситуации при анализе тональности Островский становится своеобразным драматическим эмоциональным эталоном. Объективность и общность действия драмы реализма, фокус на борьбе с устоявшимся, но давящим на персонажа социумом, могут приводить к сокрытию описания внутренней борьбы персонажей, что и объясняет низкий показатель негативных реплик — это может оказаться отражением поэтических особенностей периода.

В этом смысле примечательным оказывается хронологическое соседство реалиста Островского с новой чеховской драмой. После драмы как подробного и достоверного описания жизни произошел переход к драме внутреннего конфликта [Богданова 2024]. Откры-

тость и широта допустимого новой системы, фокус на внутренних конфликтах, закономерно привели к возможности выражения этой уникальной в контексте развития драмы экспрессивности.

Негативная динамика как отражение этапов развития сюжета

Трехактные и пятиактные структуры — традиционные формы драматургического текста. И пятиактные, и трехактные пьесы оказываются самыми многочисленными в корпусе RusDraCor после одноактных пьес и пьес без деления на акты (см. таб. 1). Такие структуры, и их теоретические описания, характеризовали еще античную драму [Аристотель 2011; Забудская 2006].

Для русской драмы, которая оформилась как распространенный жанр под влиянием французской классической драмы в начале XIX в., и которая до второй половины этого же века развивалась под эгидой классицизма, который перенял структурные особенности античной драмы, рассмотрение эмоциональной актовой динамики может оказаться продуктивным для прояснения некоторых свойств композиции [Богданова 2024; Стенник 1982]. Негативная динамика, как было показано, во многом оказывается самой интерпретируемой, и, вероятно, этапы развития драматургического конфликта должны быть в той или иной степени отражены в количестве реплик этой тональности.

На рисунках 4–7 представлена негативная динамика для трехактных и пятиактных комедий и трагедий. Первое, на что стоит обратить внимание — размах. Для пятиактных комедий этот показатель составляет 6,65 %, а для пятиактных трагедий — 8 %; для трехактных комедий около размах составляет 3,9 % для комедий, а для трагедий — 6,67 %.

У комедий негативная динамика оказывается чуть менее выраженной, но её отличие от динамики трагедий все еще выделяется.

В пятиактных трагедиях от акта к акту процент негативности непрерывно растет и к концу пьесы достигает максимальных значений, тогда как в комедиях пиковые значения достигаются в середине пьесы, в момент кульминации, а под конец действия заметен явный спад. При этом начало пьесы, первые акты, в обоих случаях оказываются наименее негативными.

Негативная динамика трехактных комедий совпадает с динамикой пятиактных. Динамика трехактных трагедий похожа на комедийную динамику, хотя все еще характеризуется большим размахом. Содержательные выводы о таком различии сделать сложно в

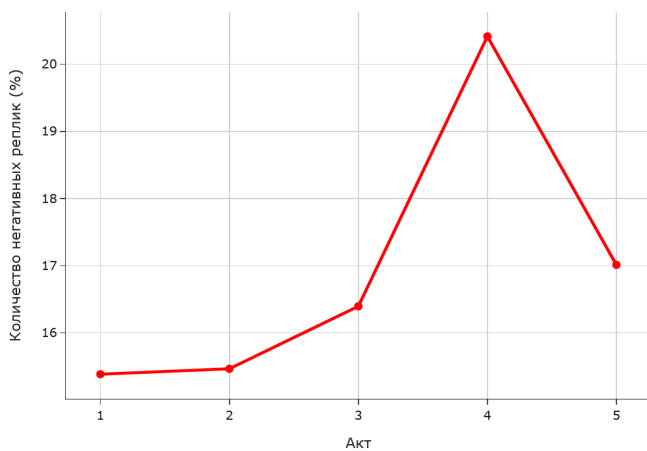


Рис. 6. Негативная динамика пятиактных комедий

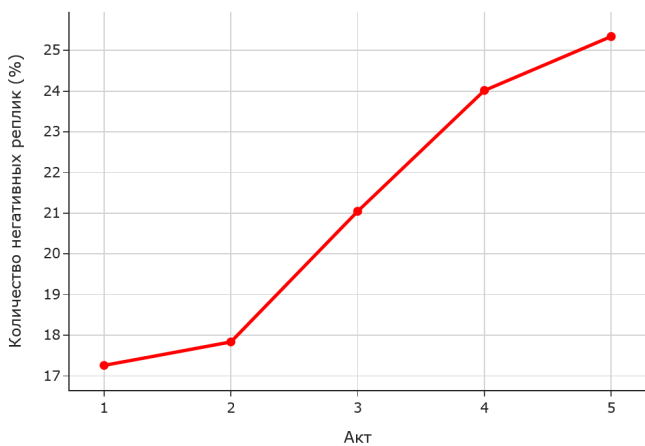


Рис. 7. Негативная динамика пятиактных трагедий

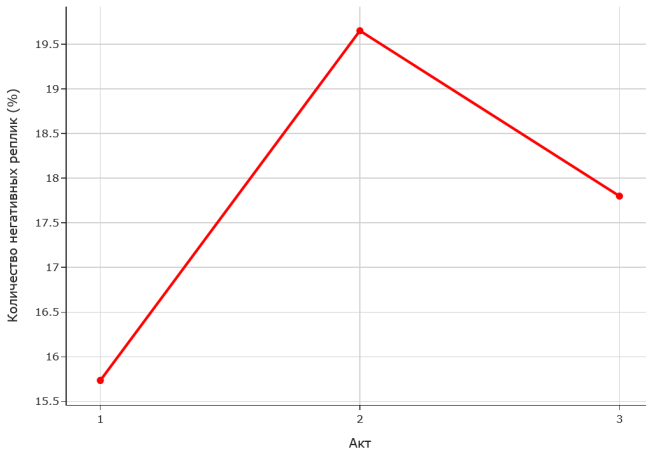


Рис. 8. Негативная динамика трехактных комедий

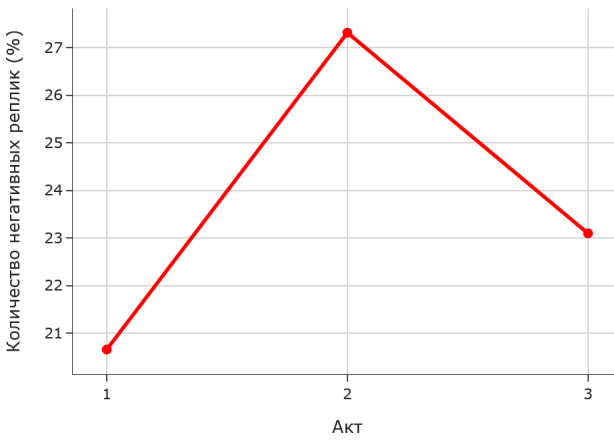


Рис. 9. Негативная динамика трехактных трагедий

связи с недостаточным количеством трехактных трагедий — их в корпусе всего 3, тогда как пятиактных трагедий 26, пятиактных комедий 24, а рехактных комедий — 18.

Хотя пятиактная и трехактные сюжетные структуры традиционно изображаются как график с «пиком» посередине, различия в динамике концовок комедий и трагедий, пятиактных и трехактных трагедий все еще не противоречит этой достаточно метафорической структуре. Повышение эмоционального накала может быть связано с той катастрофой, которая случается в конце трагедии и при этом все еще представляет собой сюжетную развязку, где накал страстей может в том числе спадать.

Такая структура сюжета характерна не только для драмы, подобное описывал Пропп и такое понимание стало нарицательным для европейской поп–культурной и художественной традиции. Примечательно, что первые попытки визуализировать сюжетную динамику оказывались метафорой накала эмоций [Whitcomb, 2010], и, хотя такой подход критиковался, такое понимание проекции развития сюжета стало хрестоматийным и аналогичную динамику мы прослеживаем в результатах анализа тональности русской драмы.

Заключение

Середина XVIII в. ознаменована появлением драмы в России как оформившегося жанра, перенявшего законы классицизма; начало XIX в. (в том виде, в котором оно представлено в RusDraCor) — это окончательное оформление классицистических драматических структур; пьесы Островского окончательно оформили русскую драму как самобытный и замкнутый, классический жанр, и в этот же момент Чехов кардинально изменил эту модель на разомкнутую, новую [Богданова 2024]. Малым количеством негативных реплик отличаются комедии 1810–1820-х годов, написанные под влиянием Мольера [Стенник 1982]. В 1870-х годах доминирует Островский, чьи реалистичные пьесы также демонстрируют низкий уровень негатива. Зато Чехов, чьи пьесы представлены в корпусе преимущественно в 1880–1890-х гг. — одновременно один из самых негативных и один из самых позитивных авторов, что отражает внутреннюю ориентацию и одновременно широкую восприимчивость его художественной системы.

Выбивающиеся из средних значений показатели негативной тональности по десятилетиям встречаются в моменты преобразования и предвосхищения этих преобразований. Эмоциональная

динамика драмы, полученная с помощью автоматического анализа тональности, может в какой-то степени отражать литературные трансформации — от классицизма до модернистской драмы, от замкнутых структур к разомкнутым и от внешнего конфликта к внутреннему.

Эмоциональная динамика русских комедий и трагедий, описанная с помощью автоматического анализа тональности, оказалась сопоставимой с качественными представлениями о композиции драмы и этапах развития сюжета художественного произведения.

Эмоциональную динамику можно назвать репрезентативным показателем, отражающим изменения драмы в истории, хотя природе этой связи стоит рассматривать в том числе с качественных и более детальных позиций. С формальных же позиций, модель типа XLM-RoBERTa-Large дала хороший результат в разметке тональностей реплик — взвешенную $f1$ -меру равную 73 %. Метрика эта рассчитана на тестовой выборке, размеченной лишь одним человеком, что можно отнести к слабым местам исследования. Хотя разметка тональности по своей природе субъективна и в идеале требует привлечения нескольких аннотаторов с последующим вычислением согласованности разметки, ограниченность ресурсов позволила на этом этапе выполнить лишь одну итерацию аннотирования.

Детекция любых эмоций — задача нюансированная. Распознавание тональностей BERT-моделями, приводит к крайне значительным упрощениям, что с одной стороны, конечно, сочетается со словами вроде комедия и трагедия, а с другой — не позволяет прямым образом *подтверждать* выделенные в литературе этапы развития драмы количественными показателями (поэтому в этой работе говорим лишь о возможности *сопоставления*), так как, сведение огромного множества иронических, саркастических или хоть сколько-то более комплексных чувств, которые безусловно присутствуют в репликах, к позитиву и негативу, приводит лишь к поверхностному описанию эмоциональной составляющей. Результаты эти, тем не менее, сочтем ценными по причине их целостности: эта поверхностная картина все же в каком-то виде изображает эмоциональный горизонт драмы.

С технической же точки зрения, подход с использованием уже дообученных BERT-моделей, несмотря на свою формальную эффективность на корпусе русской драмы, не является универсальным решением для языков с ограниченными данными для разметки. Невозможно точно смоделировать, какие именно домены

стилистически совместимы; этот выбор приходится оставлять либо на усмотрение лингвистической интуиции исследователя, либо решать экспериментальным путём. В этом исследовании интуиция о схожести реплик и постов в социальных сетях оправдалась.

Анализ тональности русских литературных текстов все же не обречённая на неудачу задача, а сложная, но многообещающая область для экспериментов.

Примечания

- ¹ Информация о жанре собиралась авторами корпуса Russian Drama Corpus на основе тегов в системе Wikidata: https://www.dracor.org/doc/odd#play_genre_normalized
- ² Код доступен в github-репозитории: <https://github.com/feredl/rusdracor-sentiment>
- ³ Полный датасет реплик можно найти на портале HuggingFace: https://huggingface.co/datasets/xnsmv/rusdracor_speech
- ⁴ <https://deepavlov.ai/>
- ⁵ <https://docs.deepavlov.ai/en/master/features/models/classification.html>
- ⁶ <https://github.com/text-machine-lab/rusentiment>
- ⁷ Здесь и далее на некоторых графиках вертикальная ось, которая демонстрирует значения в процентах, не везде начинается с 0 и заканчивается 100, но сделано это не для обмана читателя, а для наглядности. Динамика изменений в полном масштабе попросту не была бы различима, что, однако, не отменяет факта ее наличия, что косвенно подтверждается наличием условных десятилетних-выбросов по разным показателям (по методу стандартного отклонения)

Литература

Источники

Аристотель 2011 — Аристотель. Этика. Эстетика. Поэтика [Перевод] / пер. с др.-греч. М., 2011.

Полнер 1897 — Полнер Т. И. Драматические произведения А. П. Чехова // Русские ведомости. М., 1897. № 273.

Исследования

Богданова 2024 — Богданова П. Трансформации драмы в истории. Структуры порядка и структуры хаоса. М.: НЛЮ, 2024.

Волькенштейн 1960 — Волькенштейн В. Драматургия. М.: Советский писатель, 1960.

Выготский 1998 — Выготский Л. С. Психология искусства. Ростов-на-Дону: Феникс, 1998.

Забудская 2006 — Забудская Я. Л. Драма в пяти актах: Об истоках европейских драматургических форм // Индоевропейское языкознание и классическая филология. 2006. Т. X. С. 88–92.

Стенник 1982 — Стенник Ю. В. Комедия 1800–1820-х годов / Ю. В. Стенник // История русской драматургии: XVII – первая половина XIX века / отв. ред. Л. М. Лотман. Ленинград: Наука, 1982. С. 221–238.

Чудаков 1971 — Чудаков А. П. Поэтика Чехова. М.: Наука, 1971.

Ярхо 2006 — Ярхо Б. Я. Распределение речи в пятиактной трагедии / Б. Я. Ярхо // Методология точного литературоведения: Т. V / под ред. М. И. Шапира. М.: Языки славянских культур, 2006. С. 550–610.

Fischer et al 2019 — Fischer F., Börner I., Göbel M., Hecht A., Kittel C., Milling C., & Trilcke P. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama // Proceedings of DH2019: «Complexities». Utrecht, 2019. URL: <https://doi.org/10.5281/zenodo.4284002>

Liu et al 2019 — Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. — arXiv, 2019. — № arXiv:1907.11692. — URL: <https://doi.org/10.48550/arXiv.1907.11692>

Rogers 2018 — Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., & Gribov A. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian / A. Rogers [et al.] // Proceedings of the 27th International Conference on Computational Linguistics / ed. by E. M. Bender, L. Derczynski, P. Isabelle. — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. P. 755–763. — URL: <https://aclanthology.org/C18-1064/>

Sherstinova et al 2023 — Sherstinova T., Moskvina A., Kirina M., Karysheva A., Kolpashchikova E., Maksimenko P., Seinova A., & Rodionov R. Sentiment Analysis of Literary Texts vs. Reader's Emotional Responses // 2023 33rd Conference of Open Innovations Association (FRUCT). 2023. P. 243–249. URL: <https://doi.org/10.23919/FRUCT58615.2023.10143070>

Smetanin 2020 — Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access. 2020. Vol. 8. P. 110693–110719. URL: <https://doi.org/10.1109/ACCESS.2020.3002215>

Smetanin & Komarov 2021 — Smetanin S., & Komarov, M. Deep transfer learning baselines for sentiment analysis in Russian // Information Processing & Management. 2021. Vol. 58, №3. P. 102484. URL: <https://doi.org/10.1016/j.ipm.2020.102484>

Whitcomb 2010—Whitcomb S. L. The Study of a Novel. Boston, U.S.A
Kessinger Publishing, 2010.

Елизавета Сенаторова

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ РЕЧИ ПЕРСОНАЖЕЙ В ЭКРАНИЗАЦИИ РОМАНА Л. Н. ТОЛСТОГО «АННА КАРЕНИНА» (РЕЖ. АЛЕКСАНДР ЗАРХИ, 1967 ГОД)

Основная тема этой работы – сравнение романа Л. Н. Толстого «Анна Каренина» и его двухсерийной экранизации 1967 года, сделанной Александром Зархи. В центре нашего внимания – исследование близости литературного и кинематографического текстов по «плотности» прямой и косвенной речи в романе и в кинонарративе и тематическим моделям, созданным на основе текста оригинала и субтитров к фильму. Нами было выявлено, что в кинотексте присутствуют сцены молчания, которые, за исключением сцены покоса, так или иначе связаны с судьбой Анны. Можно предположить, что таким образом режиссер пытается передать идею одиночества Анны. При непосредственном сравнении текста романа и субтитров мы использовали метрику TF-IDF, позволяющую выяснить, насколько тексты вербально схожи. Помимо этого, мы построили тематические модели, позволяющие увидеть общие для документов темы. В итоге мы пришли к выводу, что реплики персонажей фильма А. Зархи переданы довольно близко к тексту литературного источника: особенные сходства, как лексические, так и тематические, есть у первой трети субтитров и первых двух частей романа, а также второй трети субтитров и четвертой части романа. Последние части романа не нашли такого же явного отклика в кинонарративе, что обусловлено затрагиваемыми в них темами: Толстой в большей степени фокусируется на философских и общественно-политических вопросах, отодвигая семейно-любовную линию на второй план. В то же время для киноадаптации важна именно сюжетная линия, связанная в первую очередь с Анной, а потому в субтитрах философские размышления, особенно характерные для восьмой части романа, опущены, так что фильм оканчивается гибелью главной героини.

Елизавета Евгеньевна Сенаторова
Columbia University
es4156@columbia.edu

DOI: 10.31860/cgi-2025-2-48-70

Ключевые слова: экранизация, Толстой, Анна Каренина, tf-idf, тематическое моделирование

Введение

Переложение¹ литературы на язык кинематографа стало одной из популярных практик уже в раннем кино. Этому способствовало не столько желание кинорежиссеров воплотить свое видение произведения в жизнь, сколько фактическое отсутствие профессиональных сценаристов, способных создать захватывающий сюжет [Буренина-Петрова 2017]. Учитывая литературоцентричность русской культуры, в дореволюционном кинематографе экранизации стали неотъемлемым жанром кинопроизводства. Однако вместе с этим появились попытки осмысления и теоретизации взаимоотношений литературы и кино. В 1926 году Борис Эйхенбаум так рассуждал на эту тему: «Перевести литературное произведение на язык кино — значит найти в киноречи аналогии стилевым принципам этого произведения» [Эйхенбаум 1926]. Тремя годами позже, его современник и единомышленник из ОПОЯЗа Юрий Тынянов так очертил проблему: «Самый подход к ней <литературной фабуле> должен измениться, потому что развитие и законы развития киносюжета — свои. Литературная фабула входит в кино не всеми особенностями, а некоторыми. Даже „инсценировка“ в кино „классиков“ не должна быть иллюстрационной — литературные приемы и стили могут быть только возбудителями, ферментами для приемов и стилей кино <...> Кино может давать аналогию литературного стиля в своем плане» [Тынянов 1977, с. 323]. Параллельно работая над сценарием для фильма по гоголевской «Шинели», Тынянов все же признавался, что иллюстрация литературного произведения – задача не из легких, ведь «у кино свои методы и приемы, не совпадающие с литературными. Кино может только пытаться перевоплотить и истолковать по-своему литературных героев и литературный стиль» [Тынянов 1973, с. 78]. Как писал позже Виктор Шкловский, попытка Тынянова создать на экране эквивалент гоголевского стиля все же не увенчалась успехом [Шкловский 1927, с. 15–16].

Итак, вопрос о «переводе» с вербального языка художественной литературы на визуальный язык кино в контексте русского кинематографа был актуален уже в 20-е годы XX века². В последующие

десятилетия дискуссия по этой теме развернулась в двух основных направлениях. С одной стороны, появился «текстовый подход», близкий представителям формальной школы и структуралистам Московско-Тартуской школы. Его последователи полагали, что кино эквивалентно³ литературе и в целом представляет собой некоторую ее разновидность. При таком подходе кино строится по принципу мимесиса, «текст выступает прототипом кинотекста, а литературный язык – прототипом киноязыка». Важно, однако, что эквивалентность подразумевает взаимность, а значит и литература должна обладать определенными кинематографическими элементами. Здесь можно вспомнить, например, идею В. Шкловского о принципе монтажа, лежащем в основе «Анны Карениной» [Шкловский 1981, с. 155] и позволяющем говорить о кинематографичности этого текста. В основе второго подхода к вопросу отношений между вербальным медиумом и визуальным лежала идея оппозиции литературы и кино на основе «разграничения изображаемого и вообразяемого», иными словами, невозможности трансмедиального перевода. Эта позиция была сформирована под влиянием Романа Jakobsona, считавшего, что грамматический строй текста не может быть передан при интерпретации и переводе с одного типа знака на другой [Jakobson 2001, с. 525]. В результате этого сам факт возможности экранизации ставится под сомнение [Буренина-Петрова 2017, с. 214].

И тут возникает вопрос: в центре полемики об экранизации оказывается проблема трансмедиального перевода – переноса содержания со страниц книги на киноэкран. Однако как быть с интермедиальным аспектом, то есть с адаптацией литературного текста в формат киносценария?

Материал исследования и постановка проблемы

В «Диалоге с экраном» [Лотман, Цивьян 1994, с. 139–144] Ю. Лотман и Ю. Цивьян кратко обращаются к роли звука в кинематографе. Представлен он может быть специально подготовленным саундтреком, шумовыми эффектами или речью, и каждое из представлений не только обладает собственной функцией, но и состоит в динамических отношениях с другими элементами (заимствуя функции или усиливая их). Например, саундтрек, включающий в себя как оригинальные композиции, так и например поп-песни, может транслировать определенные эмоции, настроение и мысли,

в то время как функцией звукового ландшафта может быть обозначение пространства, в котором происходит действие.

Речь в кино также является важным каналом передачи информации. Голос может принадлежать как человеку на экране (тогда это будет диегетический нарратор), так и тому, кого зритель не видит (в таком случае, нарратор является недиегетическим). В исследованиях экранизаций заметно усиливается интерес к компаративному анализу оригинального литературного текста и его медиальных преобразований – будь то адаптированный сценарий или звуковая дорожка фильма. В этом контексте мы, опираясь на цифровые методы, предлагаем сравнить реплики персонажей романа Толстого «Анна Каренина» с их киноверсией, а именно речью в двухсерийной экранизации романа, созданной режиссёром Александром Зархи в 1967 году.

Этот роман является одним из фаворитов среди текстов, получивших визуальное воплощение. Он насчитывает десятки экранизаций и визуальных интерпретаций по всему миру, однако его полной фильмографии до сих пор не существует. Когда-то *Tolstoy Studies Journal*⁴ предлагал актуальный до двухтысячного года список из 19 адаптаций, среди которых были полнометражные фильмы, телеспектакли и теле-сериалы. Хотя Википедия не вполне авторитетный источник, в статье об экранизациях «Анны Карениной» можно обнаружить расширенную фильмографию, насчитывающую 8 немых фильмов, 13 звуковых, 10 сериалов и 2 фильма-балета⁵. Наконец, в IMDb, одной из крупнейших кинематографических баз данных, насчитывается более 40 киноадаптаций, включающих короткометражные и полнометражные фильмы, сериалы и отдельные эпизоды с отсылками к роману, телефильмы, а также фильмы о непосредственном кинопроизводстве некоторых экранизаций.

Среди этого разнообразия киноверсий фильм Александра Зархи обрёл репутацию одной из наиболее «точных» или, скорее, близких к оригинальному тексту экранизации. Размышляя об экранизации другого толстовского текста, «Войне и мире» Сергея Бондарчука (1965), литературный критик и славист Кэрил Эмерсон отмечает, что главной заслугой этого фильма является его «преданность» тексту оригинала: «Я ценю то, что адаптация не верит в возможность вербального соревнования с Толстым. Она даже не пытается этого делать. В своей эпической картине „Война и мир“ Бондарчук почти полностью использовал фрагменты неискаженных толстовских диалогов. Это было великолепно: толстовская текстура нигде не была нарушена, а поскольку предполагалось, что образованные

русские зрители знают книгу наизусть со второго класса, не было никаких опасений по поводу „пересказа истории“» [Kokobov 2016]. Учитывая статус экранизации «Анны Карениной», можно предположить, что Зархи следовал этой же стратегии, стараясь вырезать как можно меньше оригинального текста из финальной версии киносценария. Сравнительный анализ речи персонажей в тексте романа и в его киноверсии позволит нам убедиться в этом или опровергнуть эту гипотезу.

Материалом для нашего анализа служат не киносценарий экранизации Зархи, а субтитры, фиксирующие реплики персонажей, произнесённые в фильме. Формат субтитров предполагает указание времени произнесения реплики и её текстового содержания. Однако в нём отсутствует атрибуция слов конкретным персонажам, а также разграничение диегетической и недиегетической речи. В силу этого очень сложно отследить, кто из героев говорит больше, между кем происходит больше взаимодействия внутри фильма, или чьи монологи, о которых писал Шкловский (по его мнению, весь толстовский роман строится на непонимании людей друг другом, на приеме «внутреннего монолога», где герои проецируют свой внутренний мир во внешний: «Анна всех винит и не оправдывает даже себя <...> Каренин оправдывается, как учреждение <...> отписывается встречными бумагами» [Шкловский 1981, с. 170]), длиннее и разнообразнее. Несмотря на это, с помощью количественных методов мы можем проследить интенсивность речи внутри фильма – насколько часто герои разговаривают, присутствуют ли сцены молчания, и как интенсивность речи, под которой мы понимаем количество произнесенных слов, изменяется на протяжении кинофильма. Более того, в нашем случае субтитры (или же переформатированный киносценарий⁶) являются тем, что Жерар Женетт назвал бы текстом "второго порядка"[Genette 1985, с. 5] – материалом, который *a priori* создается и воспринимается в сравнении с текстом оригинала [Hutcheon 2006, с. 6]. Вследствие этого, перед нами встает возможность проследить, как литературная основа преобразуется в кинотекст, насколько тексты близки с точки зрения затрагиваемых в них тем и используемой лексики. Таким образом, работа с субтитрами будет состоять из двух частей, первая из которых будет посвящена плотности кинонарратива, а вторая – сходству текста оригинала и субтитров.

Интенсивность речи в фильме

Субтитры в фильмах имеют следующий вид: номер реплики, время начала и конца реплики в формате HH:MM:SS,SSS (например «00:01:57,701 → 00:02:00,871») и сама реплика. Важно отметить, что обозначение времени в субтитрах не привязано к смене сцены или же говорящего. Так, например, субтитры к открывающей сцене первой части фильма выглядят следующим образом:

4

00:02:58,311 --> 00:02:59,972

Что Дарья Александровна?

5

00:03:00,246 --> 00:03:03,272

Приказала доложить,

что она уезжает,

6

00:03:03,449 --> 00:03:06,850

и пускай делают как им...

как вам, стало быть, угодно.

При этом во время просмотра фильма мы видим, что сцена беседы Стивы и камердинера Матвея снята одним кадром, и что слова «Приказала доложить, что она уезжает и пускай делают как им... как вам, стало быть, угодно» принадлежат только Матвею. Стало быть, можно предположить, что обозначение тайминга в субтитрах мотивировано временем, которое человек тратит на то, чтобы их прочесть, и местом, которое субтитры должны занимать в кадре. Для работы с субтитрами мы преобразовали исходные данные, получив таблицу (таблица 1), в которой в колонке «line» – реплика, в «seconds_start» и «seconds_end» – начало и конец фразы в секундах, в «start» и «end» – время начала и конца фразы в исходном формате, и в колонке «length» – в целях удобства восприятия округленное до целого числа количество секунд, отведенных под конкретную реплику.

Благодаря измерению времени, отведенного под реплику (последняя колонка), нам удалось выяснить, что в целом на речь персонажей в каждой части фильма 1967 года отводится около 30 минут (для обеих частей это половина экранного времени, поскольку части длятся чуть больше часа). Из вышеприведенной таблицы видно, что между репликами есть паузы, например, между первой и второй (строчки 0 и 1) присутствует интервал в 5 секунд. Та-

line	seconds_start	seconds_end	start	end	leng
Хорошо...					
Хорошо!	159.425	165.261	00:02:39	00:02:45	6.0
Ой, Боже мой!	170.570	172.595	00:02:50	00:02:52	2.0
Вам телеграмма, сударь.	174.841	176.672	00:02:54	00:02:56	2.0
Что Дарья Александровна?	178.311	179.972	00:02:58	00:02:59	1.0
Приказала доложить, что она уезжает, и пусть делают как им...	180.246	183.272	00:03:00	00:03:03	3.0
как вам, стало быть	183.449	186.850	00:03:03	00:03:06	3.0

Таблица 1. Таблица субтитров с обозначением времени реплики и количества отведенных под нее секунд

ким образом, 30 обычных минут речи равномерно распределяются по всему фильму.

В среднем за минуту экранного времени в первой части фильма произносится 53 слова, а во второй – 68 (немая сцена гибели Анны в конце картины из подсчетов исключена). На разных участках ленты, однако, плотность речи может меняться, в связи с чем и среднее число слов будет увеличиваться или уменьшаться. Для отслеживания динамики изменения среднего на протяжении экранизации мы построили регрессионную линию, которая пересекает график распределения. Так, становится видно, что к концу первой части среднее, а значит и число сказанных слов, незначительно, но все же уменьшается, в то время как к финалу второй части (рис. 1) оно постепенно возрастает.

Кроме того, если посмотреть на полиномиальную регрессию – аппроксимирующую кривую, описывающую график (Рис. 3), – мы увидим, что точки минимума в первой части достигаются в координатах, где речь персонажей практически отсутствует (20 и 55 минуты). Почти безмолвными потенциально могут быть сцены панорамных съемок природы или же сцены, когда герои только появляются в кадре и еще не успели сказать свои реплики. Таким образом, можно предположить, что малое число слов в кинотексте

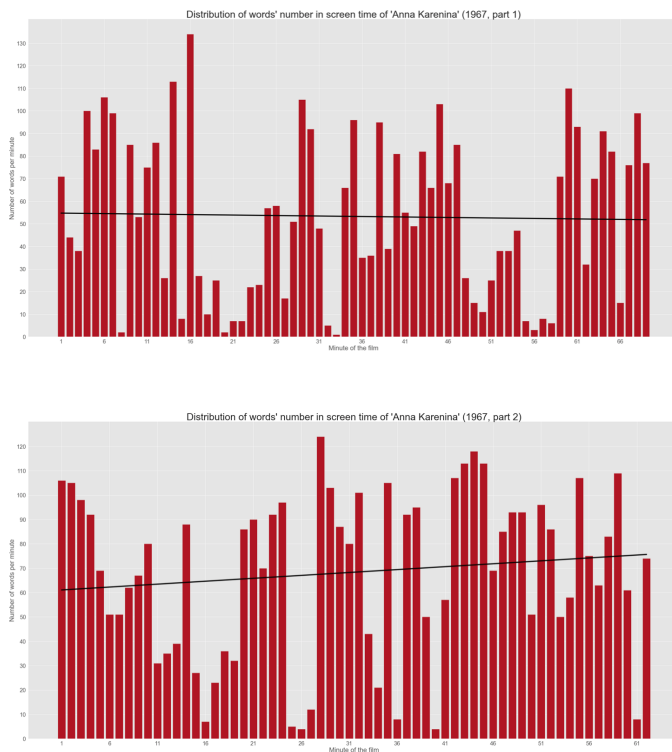


Рис. 1. Распределение количества слов по минутам экранного времени первой и второй части «Анны Карениной» (1967) и регрессионная линия. По оси X – минуты, по оси Y – количество сказанных слов

коррелирует со склейкой идейных фрагментов фильма, которых в данном случае должно быть три. Но неверно было бы утверждать, что трехчастная структура соответствует традиционному разбиению произведения на завязку, кульминацию и развязку. Есть ведь еще вторая часть, в которой ситуация несколько иная: минимум в ней лишь один и приходится он на 11 минуту, за которую персонажами сказано 50 слов.

Важно тем не менее отметить, что в конкретном случае минимальные значения кривой соответствуют в первой серии сценам бала и покоса (который отражен в экранизации сразу после сцены

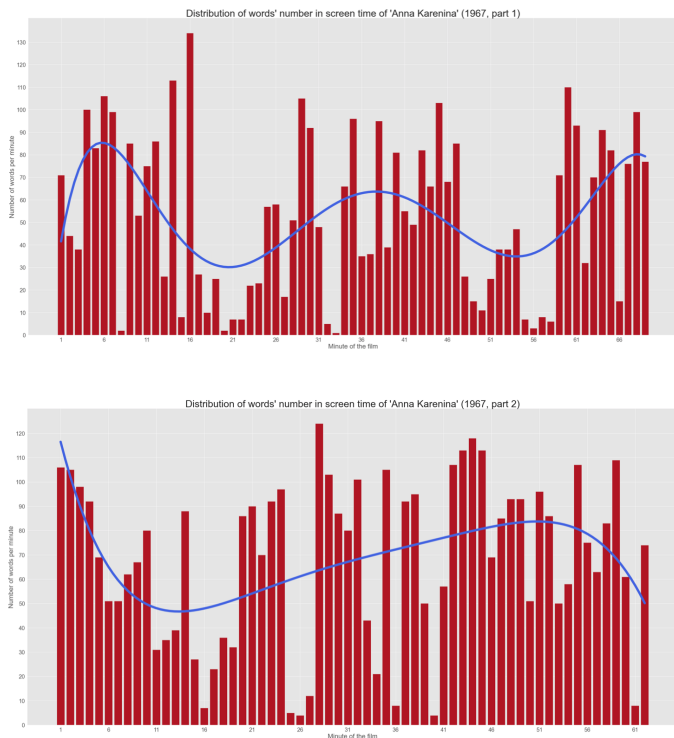


Рис. 2. Распределение количества слов по минутам экранного времени первой и второй части «Анны Карениной» (1967) и полиномиальная кривая. По оси X – минуты, по оси Y – количество сказанных слов

скачек и беседы Анны с Карениным) и сцене примирения Вронского и Каренина у постели умирающей Анны во второй.

Эти эпизоды бесспорно имеют большой вес для романа – Анна влюбляется, ее измена вскрывается, муж ее прощает. Наше предположение о монтажной склейке и съемках на природе также подтвердилось. Однако сложно сказать, что это фрагменты, действительно разбивающие нарратив на несколько частей (тем более, что параллельно существует еще одна сюжетная линия Левина, явно отраженная в экранизации Зархи). Кроме того, этими знаковыми сценами толстовский текст не исчерпывается, а потому стоит посмотреть на другие практически «безмолвные» временные отрезки,

которые регулярно появляются в фильме. Такими в первой части оказываются:

- 7 минута – встреча Анны и Вронского на вокзале в день Анниного приезда;
- 21–22 минуты – встреча Анны с истопником и проводником в поезде перед встречей с Вронским;
- 32–33 минуты – сомнения Каренина в супружеской верности Анны после вечера у княгини Бетси (VIII глава 2 части, соответствующая этой сцене, целиком состоит из описательного повествования и внутреннего монолога Каренина);
- 55–58 минуты – сцена покоса и кошмарного сна Вронского.

Во второй части фильма сцены, практически лишенные речи, следующие:

- 16 минута – попытка самоубийства Вронского;
- 25–27 минуты – жизнь Анны и Вронского в Италии;
- 36 минута – встреча Анны и Сережи в доме Каренина;
- 40 минута – сцена в театре, где все внимание светского общества приковано к Анне;
- 61 минута – ночь накануне самоубийства Анны, когда она решает «наказать» Вронского и в последний раз смотрит на него, спящего в своем кабинете (XXVI глава 7 части, которая, как и VIII глава 2 части, почти целиком состоит из внутреннего монолога Анны).

Как видно, все эпизоды, за исключением сцены покоса, связаны именно с судьбой главной героини. Не исключено, что таким образом Зархи транслирует идею одиночества Карениной, которая оказывается в безвыходном положении и не знает, к кому обратиться за помощью и с кем, что важно, поговорить («Она должна прийти к тем людям, которых она считает врагами. Она одинока» [Шкловский 1981, с. 119]).

Кроме того, в сценах, посвященных внутреннему монологу Каренина и попытке самоубийства Вронского, герои тоже испытывают если не отчаяние, то оставленность. Они, однако, позже находят

поддержку: Каренин – в словах и заботе графини Лидии Ивановны, а Вронский – в общении с княгиней Сорокиной и ее дочерью, а потому больше безмолвных сцен с ними нет. На фоне этого выделяется фрагмент о жизни Анны в Италии, в течение трех минут которого сказано около 20 слов. Одиночество в нем предстает в ином ключе – социальная изоляция не удручает Анну, а наоборот делает ее счастливой. По длине с этим фрагментом может лишь сравниться сцена непосредственно самоубийства героини, которая не отражена на графике, поскольку в ней в принципе отсутствует речь. Две минуты молчания, в течение которых Каренина находится наедине с собой, венчают киноадаптацию и являются своего рода апогеем одиночества главной героини.

Итак, благодаря визуализации интенсивности речи в экранизации 1967 года, нам удалось проследить, какие именно сцены лишены реплик персонажей. Некоторые из них практически не обращают на себя внимания при прочтении, как например внутренний монолог Каренина, однако экранизация, близкая событийно к тексту источника (соответствующие сценам главы в романе можно с легкостью обнаружить), «подсвечивает» их. Все они играют важную роль в развитии сюжетной линии произведения, и при этом они равномерно разбавляют непрерывающиеся беседы героев. Кажется, что в такие моменты тишины зритель остается лицом к лицу с героями и наблюдает за ними со стороны, что позволяет ему осмыслить случившееся и непредвзято судить о происходящих на экране событиях.

Построение тематических моделей на основе субтитров и текста романа

Есть вероятность, что сюжетная близость киноадаптации к роману подкреплена текстовой близостью экранизации к роману. Оценить содержание текстовых произведений в духе *distant reading* позволяет тематическое моделирование, см. в [Лейбов, Орехов 2022] обзор применений этого метода в сфере цифровых гуманитарных исследований. Прежде чем обратиться к измерению сходства вербальных уровней субтитров и художественного источника, мы сделали предварительную обработку наших данных. Корпус текстов, с которым мы работаем на этапе измерения сходства вербального уровня субтитров и художественного источника, состоит из 11 документов. Восемь из них соответствуют прямой и косвенной речи из частей, составляющих роман Толстого, а оставшиеся

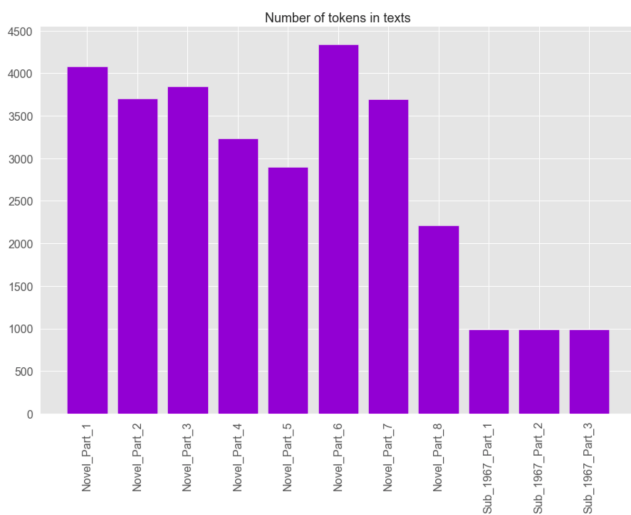


Рис. 3. Количество токенов в текстах нашего корпуса

три содержат субтитры обеих серий фильма, разделенные согласно принципу трехчастной структуры повествования на три одинаковых по длине документа (в токенах). Стоит отметить, что во всех анализируемых текстах слова были приведены к начальной форме, а стоп-слова, реплики на иностранных языках и именованные сущности (такие как имена персонажей и локации) были удалены. В результате мы получили корпус объемом 30944 токенов (соотношение количества токенов внутри документов разное, рис. 3).

Далее тексты были векторизованы (представлены рядом чисел), и для каждого слова в конкретном документе была рассчитана статистическая мера TF-IDF, позволяющая оценить важность слова для какого-либо документа относительно всех остальных документов. Полученная матрица показателей размером 11×5428 (где первое число отвечает за количество документов, а второе – за количество уникальных токенов) была перемножена на себя же, только транспонированную, в результате чего получилась уже матрица 11×11 , отражающая близость текстов числом от 0 до 1 (0 – тексты не похожи, обозначено красным цветом, 1 – тексты идентичны, обозначено синим цветом).

На визуализации нашей матрицы (рис. 4) видно, что непосредственно части романа довольно близки между собой, у них показа-

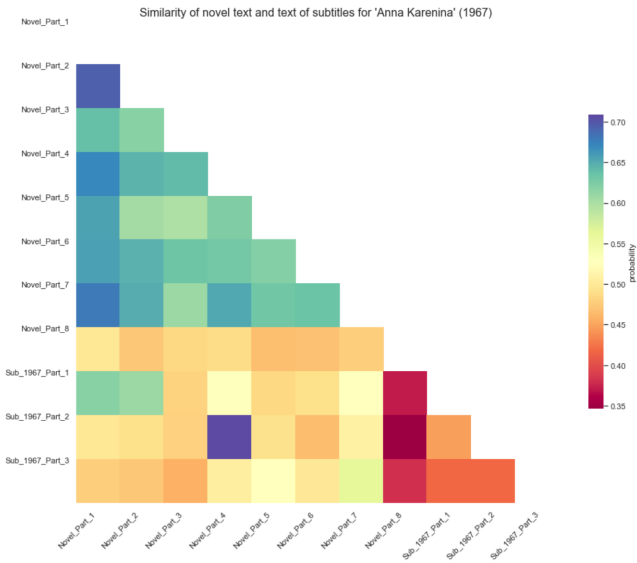


Рис. 4. Визуализация матрицы близости текстов

тели ранжируются от 0,6 до 0,7. Особенно схожими являются первая и вторая части, для них индекс близится к единице. Исключение составляет восьмая часть, коэффициенты в которой колеблются от 0,45 до 0,5. Связано это, судя по всему, с изменением проблематики романа: семейно-любовная линия отходит на второй план, в то время как в центре внимания оказывается философско-общественная. Как пишет Б. Эйхенбаум, последняя часть принимает «злободневный, публицистический характер» [Эйхенбаум 2009, с. 653], и в ней отражаются вопросы, беспокоящие Толстого: славянский вопрос, война в Сербии, вопрос веры и неверия в Бога. Все внимание обращено к фигуре Левина, которая постепенно становится все более автобиографичной. Как видно в правой нижней части тепловой карты, показатели близости последней части и субтитров, в свою очередь, близятся к нулю. Это можно мотивировать тем, что восьмая часть не отражена в экранизации, и фильм заканчивается гибелью Анны. Кроме того, показатели сходства субтитров друг с другом также довольно низкие. Вероятно, причиной этого является стремление режиссера отразить уникальную сюжетную линию романа, опустив размышления на более общие темы. Если же сравнивать

субтитры с текстом романа до седьмой части включительно, становится очевидной близость первой трети киносценария и первых двух частей романа (чуть больше, чем 0,6), а также практически полная идентичность второй трети субтитров и четвертой части (более 0,7). Последняя треть кинотекста не находит однозначного источника, хотя и близится к седьмой части романа. Повлиять на это мог слишком большой объем 6 и 7 части, свидетельствующий о расширении границ романа и затрагивании тем, отстраненных от сюжетной линии, о чем также упоминает Эйхенбаум. Несмотря на это, мы все равно в целом можем утверждать, что киносценарий написан довольно близко к тексту художественного источника, как с точки зрения сюжета, так и с точки зрения заимствованных из текста фраз.

Наконец, стоит рассмотреть, насколько тексты близки друг другу с точки зрения затрагиваемых в них тем. Для этого на основе чуть видоизмененного корпуса (теперь роман представлен не в восьми частях, а в трех, каждая из которых состоит из 8990 токенов) мы построили тематические модели, в основе которых лежит популярный алгоритм Латентного распределения Дирихле (LDA).

На вход алгоритм получает текст и количество топиков/тем/наборов из слов, которые с высокой долей вероятности появляются в тексте рядом, которые надо найти в корпусе. В качестве финального результата он выдает n списков самых характерных для тем слов. Чтобы определить оптимальное число топиков, мы провели следующие операции:

1. Для начала мы выбрали количество слов, описывающих топик (20) и интервал от 1 до 15, в котором предположительно находится оптимальное значение числа топиков. Для каждого из значений этого интервала была построена отдельная модель LDA, в результате чего для n -ной модели мы получили n списков токенов, определяющих темы (для первой – один список, для второй – два и т. д.).
2. Далее мы посчитали средний коэффициент Жаккара⁷ для каждой пары соседних моделей LDA (для 1-ой и 2-ой, для 2-ой и 3-ей и т. д.). Если, сравнивая n -тую модель с $n+1$ -ой моделью получилось значение больше, чем на предыдущем шаге (при сравнении n -той модели с $n-1$ -ой), то n является оптимальным значением топиков. Иными словами, изобразив на графике средний коэффициент Жаккара по всем парам

соседних тем, мы видим, что оптимальным числом топигов будет 5 – в точке минимума полученной кривой.

3. Вторая мера, подтверждающая правильность выбора числа топигов, называется согласованностью/когерентностью темы (Topic Coherence). В отличие от Жаккара, когерентность на вход берет только список токенов, соответствующих одной теме. Хорошей темой является та тема, в список токенов которой попали слова из одного контекста. Для подсчета когерентности каждой темы мы считаем сумму показателей контекстной близости всех пар слов из списка соответствующих ей токенов. Для каждой пары слов показатель будет высоким, если токены часто находятся рядом в наших текстах, и маленьким, если они чаще встречаются по отдельности. Далее мы усредняем показатели когерентности по всем темам данной модели (в 1-ой модели будет один показатель, поскольку только одна тема была выделена, во 2-ой – среднее от двух показателей, поскольку тем две, и т.д.). Выбор чрезмерного числа топигов будет приводить к появлению тем со случайным набором слов, а следовательно с низким показателем когерентности. Вследствие этого средняя когерентность будет снижаться, а значит интересующая нас точка оптимума будет достигаться в максимуме среднего.
4. Наконец, мы смотрим, где точка Жаккара достигает минимума, а одновременно с этим точка когерентности – максимума. В этом месте на графике (рис. 5) и будет обозначено идеальное число топигов для корпуса. В нашем случае число 6, полученное после подсчета меры Жаккара, подтвердилось.

После всех вышеописанных эвристик в корпусе было выделено шесть тем (рис. 7–12), а на основе полученной таблицы вероятностей встретить тему в том или ином документе была построена тепловая карта (рис. 6). Как можно было догадаться, последняя явно отражает тенденции, замеченные на графике схожести документов. Так, с вероятностью ~ 0.3 в первых третях романа и кинотекста можно обнаружить тему 3, среди ключевых слов которой есть «бал», «танцевать», «влюбленный», «увлечение», «гувернантка», «устрицы». Уже по этим леммам несложно предположить, что речь идет об измене Стивы Долии с гувернанткой, беседе Стивы с Левиным об этом, бале Китти и роковых скачках. В целом же настроение темы довольно светлое, в ней описывается далеко не безгрешная,

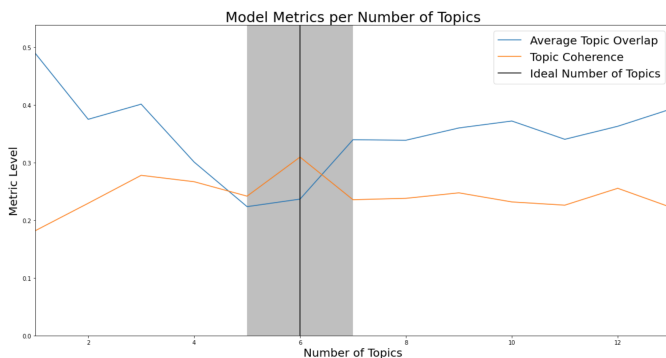


Рис. 5. Средние коэффициента Жаккара и когерентности темы

однако все же беспечная жизнь светского общества. Для начальных глав романа также свойственна тема 1, которая кажется амбивалентной: с одной стороны, в ней все еще присутствует возвышенность и легкость («наслаждение», «влюбленный», «слава», «петь»), а с другой появляются первые признаки неизбежной трагедии («дуэль», «преступный»). Лексема «погибать» используется пока что в ироничном ключе, хотя один раз Анна думает, что она «погибшая женщина». «Ужасный» же является усилением («ужасно устала»). Наконец, слово «здоровье» связано с линией Николая Левина, опущенной в фильме.

Для вторых третей общей является тема 4 (вероятность ее появления в текстах также равна ~0.3), которая по настроению кажется более серьезной и мрачной. «Влюбленность» и «увлечение» сменяются в ней словами «прелюбодеяние» и «любовник», а вместо развлечений речь идет о разводе и смерти («развод», «погибать», «умирать»). «Кормилица» же явно отсылает к рождению дочери Анны от Вронского. Кроме того, сюда попали такие слова, как «верить», «сомневаться», «раб», «рабочий» и «образование», что явно указывает на философские размышления Левина и его диспуты с Кознышевым, Песцовым и прочими представителя элиты на общественно-политические вопросы (например, есть ли в образовании польза для народа). «Раб» в данном случае выступает не в контексте крепостного права, которое тоже затрагивается в светских беседах, а в контексте религии, характерном для сцены венчания Левина и Китти.

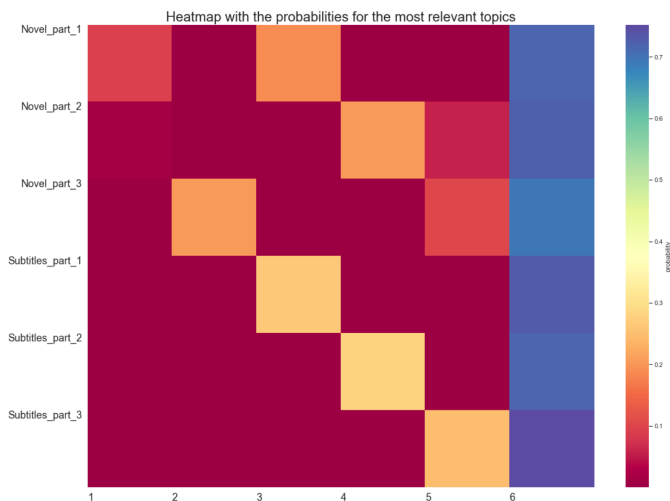


Рис. 6. Тепловая карта вероятностей топигов в текстах. Околонулевая вероятность обозначена красным цветом, вероятность, стремящаяся к единице – синим

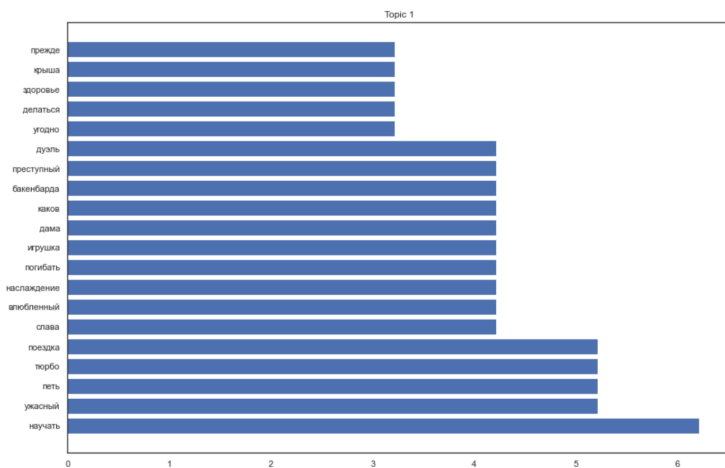


Рис. 7. 20 наиболее частотных слов в топиках

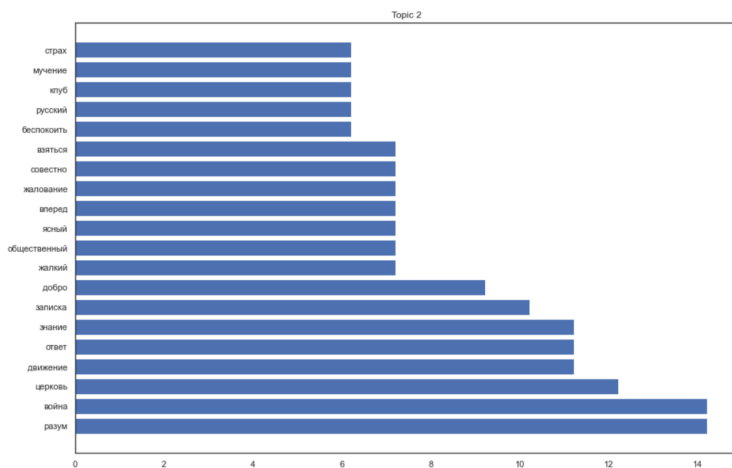


Рис. 8. 20 наиболее частотных слов в топиках

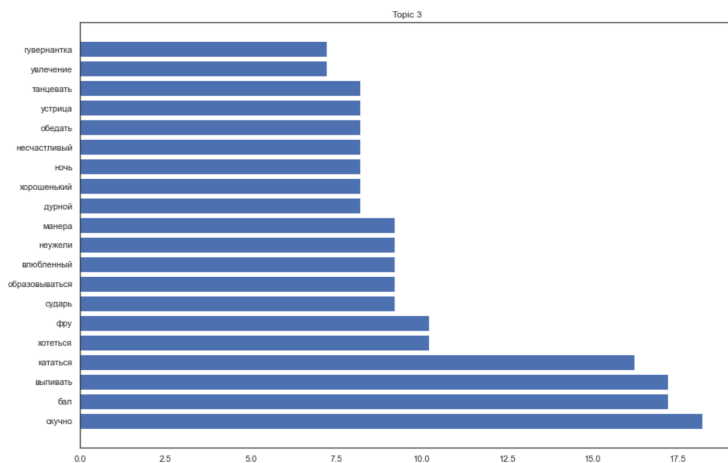


Рис. 9. 20 наиболее частотных слов в топиках

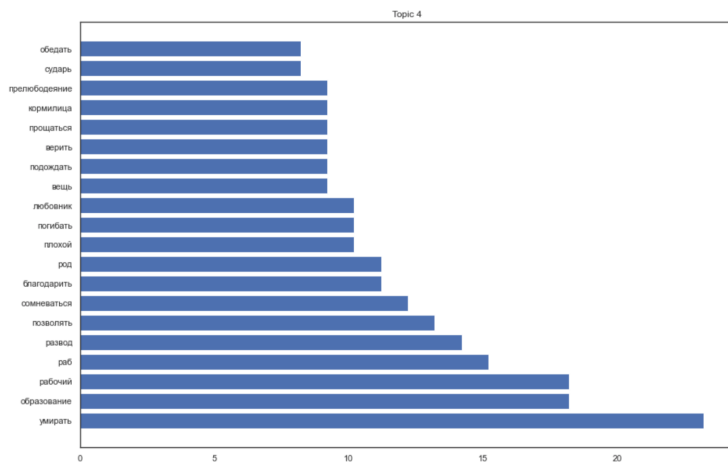


Рис. 10. 20 наиболее частотных слов в топиках

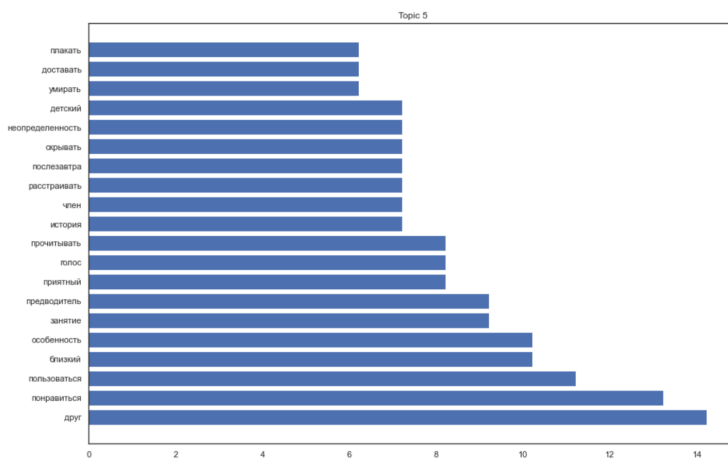


Рис. 11. 20 наиболее частотных слов в топиках

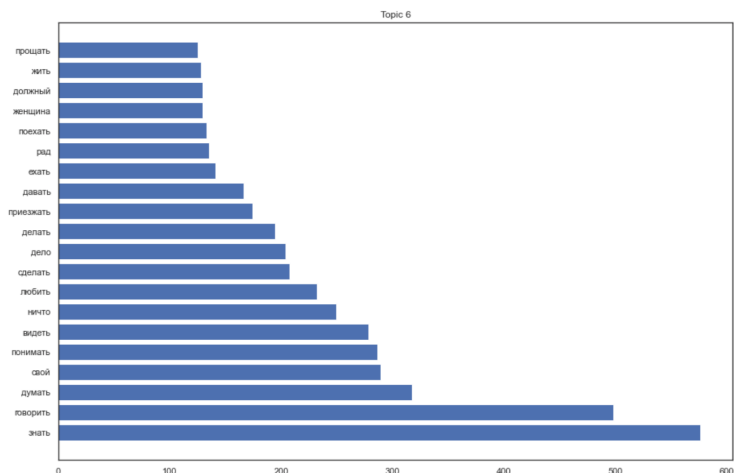


Рис. 12. 20 наиболее частотных слов в топиках

Наконец, в последней трети субтитров появляется тема 5, не слишком характерная для последней трети романа, хотя иступающая в ней. В списке наиболее характерных слов, описывающих этот топик, в целом можно выделить два семантических поля: военно-политическое, связанное с карьерой Каренина и Вронского («предводитель», «член», «занятие», «история»), и эмоционально негативное, описывающее складывающиеся между Анной и обществом, в частности Вронским, отношения («расстраивать», «неопределенность», «плакать», «умирать», «скрывать»). Даже на первый взгляд обладающие позитивной коннотацией слова «друг» и «приятный» выступают в негативном ключе: в контексте первого, которое на самом деле является взаимным местоимением, обычно возникает мотив ненависти («ненавидеть друг друга»), а в поле второго попадает частица «не» («не очень приятный»).

Финалу романа, в свою очередь, более свойственной является тема 2. Она довольно похожа на пятую, однако в ней религиозно-философское начало выражено ярче, а лексемы, связанные с эмоциями и отношениями, наоборот сведены к минимуму. Так, за испытываемые героями чувства отвечают слова «страх», «мучение», «совестно», «жалкий», в то время как общественно-политические и философско-религиозные направления мысли выражены токе-

нами «клуб», «русский» (в контексте народа), «жалование», «общественный», «знание», «ответ», «движение», «церковь», «война», «разум». Как мы говорили выше, все эти темы волновали Толстого и были отражены в восьмой части романа, которая не была адаптирована для кинозрителя.

Наконец, шестая тема, показатели которой являются высокими для всего корпуса текстов, состоит практически целиком из нейтральных глаголов, популярных вне зависимости от настроения текста и его тематики, и при этом не попавших в список стоп-слов.

Таким образом, анализ текстовой близости и тематических моделей позволил нам обнаружить сильную корреляцию между киноповествованием в экранизации Александра Зархи и текстом Толстого. Кинонарратив развивается в соответствии с сюжетом, заданным автором художественного произведения, а субтитры транслируют источник очень близко к тексту. При этом компаративный анализ позволил увидеть не только сходства, но и отличия субтитров от оригинала, которые заключаются в отсутствии восьмой части в киноадаптации, а также в упрощении затрагиваемых в произведении тем и следовании исключительно сюжетной линии практически без философских отступлений.

Примечания

- ¹ Выражаю признательность анонимному рецензенту за конструктивные и доброжелательные замечания и конкретные предложения по улучшению этого текста.
- ² Самые первые замечания по этому поводу были сделаны еще раньше, в 1911 году итальянским критиком и теоретиком Ричотто Канудо в «Манифесте семи искусств». Подробнее об этом: [Bordwell 1997, p. 29].
- ³ Судя по всему, в этом контексте используется термин из области лингвистики и переводоведения, где он означает максимально близкое соответствие переведенного текста и текста оригинала на определенном уровне, синтаксическом, прагматическом или других.
- ⁴ "Tolstoy Filmography." *Tolstoy Studies Journal. A Publication of the Tolstoy Society of North America*. URL: <https://web.archive.org/web/20240529190302/https://www.tolstoy-studies-journal.com/filmography>
- ⁵ "Анна Каренина (значения)". Википедия. Свободная Энциклопедия. URL: [https://ru.wikipedia.org/wiki/Анна_Каренина_\(значения\)#Фильмы](https://ru.wikipedia.org/wiki/Анна_Каренина_(значения)#Фильмы)
- ⁶ Лотман и Цивьян разделяют сценарии на литературный, похожий на рассказ или повесть, и на режиссерский, в котором литературный язык

переведен на язык кино – обозначена раскадровка, прописаны планы и ракурсы съемок, присутствуют режиссерские ремарки.

- ⁷ Мера Жаккара – бинарная мера сходства множеств. Например, у нас есть множества тем {любовь, смерть} и {любовь, жизнь}. С помощью данной меры во множествах обнаруживаются одинаковые темы (в нашем случае любовь), после чего их число делится на количество уникальных тем для всех множеств (здесь – любовь, смерть и жизнь). Таким образом мы получаем коэффициент Жаккара, который в этом примере равен 1/3.

Литература

Исследования

Буренина-Петрова 2017 — Буренина-Петрова О. Литература на экране. Wiener Slawistischer Almanach, 79. 2017. URL: <https://www.zora.uzh.ch/entities/publication/8de47806-85f2-4054-b143-3ab1ca69dea9>

Лейбов, Орехов 2022 — Лейбов Р. Орехов Б. В. Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике // Шаги/Steps. Т. 8. 2022. № 2. С. 205–232. DOI: 10.22394/2412-9410-2022-8-2-205-232.

Лотман, Цивьян 1994 — Лотман Ю., Цивьян Ю. Диалог с экраном. Таллин: Александра, 1994.

Тынянов 1977 — Тынянов Ю. Н. Поэтика. История литературы. Кино. М., 1977.

Тынянов 1973 — Тынянов Ю. Н. Либретто кинофильма «Шинель» // Из истории Ленфильма. Вып. 3. Л., 1973. С. 78–80.

Шкловский 1981 — Шкловский В. Энергия заблуждения. Книга о сюжете. М.: Советский писатель, 1981.

Шкловский 1927 — Шкловский В. Их настоящее. О советских кинорежиссерах. М., Л., 1927.

Эйхенбаум 2009 — Эйхенбаум Б. М. Лев Толстой: исследования. Статьи / Б. М. Эйхенбаум; сост., вступ. статья, общ. ред. проф. И. Н. Сухих; комментарий Л. Е. Кочешковой, И. Ю. Матвеевой. СПб.: Факультет филологии и искусств СПбГУ, 2009.

Эйхенбаум 1926 — Эйхенбаум Б. М. Литература и кино // Советский экран. 1926. № 42. С. 10.

Якобсон 2001 — Якобсон Р. «Поэзия грамматики и грамматика поэзии» / Семиотика: Антология, сост. Ю. С. Степанов; Изд. 2-е, испр. и доп. М.: Академический Проект; Екатеринбург: Деловая книга, 2001, 525–546.

Bordwell 1997 — Bordwell D. On the history of film style. Harvard: Harvard University Press, 1997.

Genette 1982 — Genette G. Palimpsestes: La littérature au second degré. Paris: Seuil, 1982.

Hutcheon 2006 — Hutcheon L. A Theory of Adaptation. New York, London: Routledge, 2006.

Kokobobo 2016 — Kokobobo A. Candid about the Camera: Tolstoy Scholars on Adapting Anna Karenina. Jordan Center, 2016. URL: <https://jordanrussiacenter.org/blog/candid-camera-tolstoy-scholars-adapting-anna-karenina>

ИНСТРУМЕНТЫ

Борис Орехов

ОТКРЫТЫЕ КОМПЬЮТЕРНЫЕ ИНСТРУМЕНТЫ ДЛЯ РЕШЕНИЯ ЗАДАЧ ОЦИФРОВКИ И АНАЛИЗА РУССКОЯЗЫЧНОГО ТЕКСТА В ОБЛАСТИ DIGITAL HUMANITIES

В статье дается обзор не очень известных модулей, которые можно использовать для решения задач Digital Humanities, связанных с текстовым анализом и оцифровкой. К таким модулям отнесены те, которые облегчают оцифровку текстов, напечатанных в дореформенной орфографии (OCR-модель и конвертер в новую орфографию), акцентуатор, расставляющий ударения, детектор прямой речи, код, позволяющий оценить формульность фольклорного текста, конвертер для формата TEI. В этом же ряду представлен модуль, облегчающий обработку текста для публикации в настоящем журнале.

Ключевые слова: python, digital humanities, old orthography, converter

Манифест

Даже¹ в эпоху почти всемогущих больших языковых моделей продолжают быть востребованы компьютерные инструменты, позволяющие решать частные задачи, потому что именно с таким классом задач LLM справляются хуже. Пока что современным версиям искусственного интеллекта недоступна тонкая нюансировка, завязанная на особенностях предметного поля.

Одновременно с этим информация о таких инструментах плохо распространяется и внутри сообщества, и среди внешних специалистов, потенциально заинтересованных во внедрении этих инструментов в свои продукты. Автору этих строк известно как минимум о двух попытках самостоятельно создать программный комплекс, дублирующий функционал одного из описанных ниже модулей для Python, несмотря на то, что модуль существует давно, показывает высокое качество работы и распространяется свободно. Проблема в информированности. Что тоже удивляет, учитывая, как много усилий вкладывается разработчиками поисковых систем в свои технологии, а через поисковые системы установить наличие обозреваемых ниже инструментов совсем не трудно. Возможно, тем, кто испытывает потребность в этих инструментах, кажется неправдоподобной мысль, что кто-то эту потребность уже удовлетворил, — настолько частной представляется задача, которую требуется решить — что они даже не пользуются поисковиками, чтобы на деле убедиться в том, в чем и так уверены.

Тем не менее, ситуация не остается статичной. В сфере Digital Humanities, в том числе и в России, постепенно происходит накопление полезной кодовой базы, и важной задачей сообщества становится информирование и самоинформирование о технических возможностях, которыми прирастает поле благодаря этой базе.

Окончательно решить проблему информированности не удастся: слишком насыщенным представляется современное информационное поле, а ресурсы сообщества на этом поле ограничены. Это, однако, не означает, что от решения поставленных задач следует отказаться. Мы с коллегами видим этот журнал как способ информирования всех заинтересованных, а этот текст — один из первых шагов в этом направлении.

Далее речь пойдет о разработанных отечественными компьютерными лингвистами и специалистами в области цифровых гуманитарных исследованиях инструментах, которые могут оказаться полезными для аналитической работы с текстами на русском языке.

Оговорка состоит в том, что в обзоре будут не все поддерживающие русский язык программные библиотеки, а только те, информация о которых нуждается в распространении. Во-первых, модули типа *natasha*, *deerpavlov* или *spacy*, с помощью которых также можно решать задачи анализа русскоязычного текста, и так достаточно известны, и даже входят в материалы соответствующих учебных курсов; узнать о них не представляет сложности. Информационной поддержки же требуют менее масштабные проекты, спектр возможностей которых не так широк, но инфраструктурная роль которых может оказаться важной в рамках предметного поля. Во-вторых, вышеназванные и другие подобные им библиотеки, предоставляющие функциональность для анализа текста, в достаточно сильной степени практически ориентированы на обеспечение потребностей компьютерной лингвистики, смежной, но не тождественной цифровым гуманитарным исследованиям области. В настоящем тексте я сосредоточу внимание именно на ДН-ориентированных модулях, которые могут оказаться полезными и компьютерным лингвистам, но по остаточному принципу.

Все модули, упомянутые в этом обзоре, представляют собой открытое, свободно и бесплатно распространяемое программное обеспечение, функционирующее как расширение для базовой сборки интерпретатора языка программирования Python. Все они были созданы либо под руководством либо усилиями преподавателей школы лингвистики московского кампуса НИУ ВШЭ.

Обозреваемые модули удобно разделить на три группы, основываясь на сфере их применения. В первой группе оказываются модули, рассчитанные на работу с русскоязычными текстами в старой (дореформенной) орфографии. Во второй — модули, помогающие в работе с преимущественно художественными текстами или текстами, которые ставятся в один ряд с художественными. В третьей группе, состоящей из одного модуля, находится код, помогающий в подготовке публикаций в настоящем журнале.

Старая орфография

Реформа русской орфографии произошла в 1917 году, разделив историю отечественного письма на «до» и «после». Из текстов исчезли некоторые буквы, отменено обязательное добавление буквы «ъ» к слову, оканчивающемуся на согласный, изменены правила написания некоторых приставок. Помимо всех культурных и семантических проекций этого шага, нужно учитывать и технические.

Для русского языка существует достаточно мощное оснащение инструментами автоматической обработки текста, включающими весь основной функционал такого рода — от морфологического анализа до извлечения именованных сущностей. Но все эти языкозависимые инструменты адаптированы для анализа текста в новой, пореформенной орфографии. Таким образом, дореволюционные тексты оказываются без технического обеспечения. При этом специалистам по цифровым гуманитарным исследованиям часто приходится иметь дело с культурно значимыми текстами, созданными в дореформенную эпоху и с тех пор не переиздававшимися. Уже сам процесс их оцифровки связан с известными сложностями, поскольку системы OCR для русского языка настроены на распознавание современных текстов. Такие инструменты не справляются ни с отдельными символами, свойственными старой, но не новой орфографии, ни с целыми словами, изменившими свой облик после реформы.

OCR-модель для распознавания текстов в старой орфографии

Созданию модели предшествовала публикация датасета распознанных и вручную исправленных текстов в старой орфографии [Orekhov 2024a]. На этих данных была обучена модель, использующая инфраструктуру `tesseract-ocr`.

Ссылка: <https://huggingface.co/Serovvans/trocr-prereform-orthography>

Достоинствами инструмента являются легкость установки и использования. Зависимости инструмента: `tesseract-ocr` и `poppler` и пакеты для Python: `pytesseract` (версии 0.3.10 и выше), `Pillow` (10.0.0), `transformers` (4.46.0), `torch` (2.5.0), `numpy` (1.23.0), `tqdm`, `pdf2image` (1.17.0, для распознавания книги целиком). Все они устанавливаются при помощи пакетных менеджеров в тех системах, где такие менеджеры доступны.

Код, вызывающий соответствующие функции, также не переусложнен:

```
from recognize_page import recognize_page
page_path = "page.png"
text = recognize_page(page_path,
text_output_path="output/file.txt")
print(f"Текст страницы:\n")
```

```
print(text)
```

Таким образом можно вывести на экран текст, размещенный на одной распознанной странице. Есть отдельная функциональность для распознавания текста целиком.

Качество распознавания, которое предоставляет модель, можно охарактеризовать как среднее:

- CER (Char Error Rate) = 0.095
- WER (Word Error Rate) = 0.298

Эти показатели означают, что если отдельные символы распознаются системой довольно уверенно, то с целыми словами возникают сложности, приводящие к тому, что правильно прочитывается только около 70 % лексических единиц. Таким образом, после применения этого инструмента требуется дополнительный контроль качества распознавания и корректировка результатов, с чем как раз довольно неплохо справляется большая языковая модель.

Вот результаты эксперимента с GPT-4o (лето 2025 года):

Запрос: Ниже приведен текст на русском языке в дореформенной орфографии, в нем содержатся ошибки распознавания после OCR. Исправь ошибки, сохрани старую орфографию.

дат-

скій король Христіанъ VII посѣтилъ академію наукъ.

Ответ: Ваш текст, исправленный с сохранением дореформенной орфографии:

датскій король Христіанъ VII посѣтилъ академію наукъ.

Ошибки распознавания исправлены, а орфография текста соответствует дореформенному порядку.

Еще одно уточнение следует ввести при разговоре о русскоязычной типографике XVIII века. Известно, что тексты этого времени отличаются от позднейших в том числе формой некоторых букв. Например, буква «т», приобретшая современный облик еще в XIX веке, за 100 лет до того напоминала скорее букву «m». Для таких текстов обозреваемая модель подходит плохо, нужно использовать альтернативные решения, некоторые из них разработаны на базе программы Transkribus, но существуют только в ее экосистеме. Перспективно выглядят модели «Russian print of the 18 c. (V. Okorokov's

Printing House)»² и «Russian print XVIII cent PyLaia»³. При их использовании, однако, приходится учитывать, что и сама программа Transkribus, и созданные для нее модели не являются свободным и бесплатно распространяемым продуктом, а значит, вся эта экосистема экзистенциально связана со стоящей за ней частной предпринимательской структурой, перспективы существования которой не всегда ясны.

Возвращаясь к основной обозреваемой разработке, следует добавить, что в ее описании присутствует абсурдная помета: «The model is trained to recognize printed texts in Old Russian language». Поскольку в основу модели легли данные текстов на современном русском, а не древнерусском языке, можно с уверенностью сказать, что этот тезис не соответствует действительности. Очевидно, создатели модели перепутали древнерусский язык и современный русский язык, фиксируемый на письме при помощи правил старой орфографии. У этих сущностей действительно есть некоторое внешнее сходство. Например, частично пересекается набор графем (буквы «ять», «фита», «ижица», «и десятеричное»), отсутствующих в современной системе письма. Однако даже сам набор графем в до-реформенной орфографии и древнерусской системе письма не тождествен. В старой орфографии отсутствуют древнерусские йотированные варианты букв, предназначенных для обозначения гласных фонем, юсы (большой и малый), написание под титлом, древнерусская пунктуация. Старая орфография наследует древнерусской письменности, но наследие не подразумевает тождества. Но главное — современный русский и древнерусский языки серьезным образом отличаются с точки зрения лингвистического описания. Они имеют разный лексический состав (среди бросающихся в глаза частотных лексико-грамматических отличий — в древнерусском языке иначе выглядят местоимения), разные грамматические системы (отличаются склонение имен, формы глагольного времени и т. д.). Отличить современный русский язык в старой орфографии и древнерусский язык с легкостью сможет любой носитель, даже не располагающий специальной лингвистической подготовкой.

Если не учитывать этого недостатка описания модели, она вполне пригодна к использованию как бесплатная альтернатива коммерческим решениям.

Модуль для транслитерации старой орфографии в новую

После того, как текст оцифрован, его можно обработать с помощью соответствующих программных средств. Но, как уже было сказано, большая часть из них языкозависимые и настроены на современное написание. Коммерческие компании, делающие существенный вклад в появление и распространение программных решений в области обработки текста, не заинтересованы в текстах в старой орфографии по очевидным причинам: не существует заказчиков, располагающих достаточным количеством подобных данных, и способных оплатить их обработку. Чтобы получить возможность беспрепятственного использования наличных инструментов, следует конвертировать текст, написанный в дореформенной орфографии, в новую орфографию. Для этого подходит модуль для Python *prereform2modern*.

Ссылка: <https://pypi.org/project/prereform2modern/>

Достоинства модуля в качественной работе и отсутствии сложных зависимостей, что облегчает установку. Все преобразования совершаются при помощи встроенных в базовую сборку Python средств работы со строками.

Несмотря на узкую специализацию модуля, он обладает достаточно широким спектром настроек. Тексты можно преобразовывать, сохраняя информацию о внесенных изменениях, а результат преобразования может быть представлен как в виде простого текста, так и в виде фрагмента XML, сформированного по правилам TEI:

```
<choice>  
<reg>пример</reg> <orig>примеръ</orig>  
</choice>
```

Функциональность модуля можно вызывать и из командной строки, и из интерфейса интерпретатора Python. Модуль несколько лет не обновлялся, но это не характеризует его с худшей стороны, так как качество его работы с самого начала было на достойном уровне, и обновления коду не требуются. Здесь будет уместно привести бурлескное стихотворение, иллюстрирующее случай обзвечаемого инструмента:

```
бог создал труд и обезьяну  
чтоб получился человек  
а вот пингвина он не трогал  
тот сразу вышел хорошо
```

*Художественные и парахудожественные тексты***Модуль для акцентуации русского поэтического текста**

Расстановка ударений в русскоязычном тексте — действие, которое необходимо для многих типов аналитических операций, но главным образом они связаны с областью интересов стиховедения. Ударения могут быть точкой отсчета для автоматизированного определения силлаботонического размера, установления конфигурации ритма строки, метр которой уже определен, для нахождения ритмизованных отрезков прозы. Разумеется, границы применения такого инструмента шире дисциплинарных рамок цифровых гуманитарных исследований, и включают интересы специалистов, работающих с учебными материалами для иностранцев, с программным обеспечением для синтеза речи и т. д. Трудности, которые встают перед создателями такого инструмента, заключаются в том, что основанный на словарях и правилах, он не сможет покрыть все лексическое разнообразие русского языка, а выстроенный на базе машинного обучения будет допускать ошибки в простых однозначных и при этом частотных случаях. Купировать эти проблемы способен программный модуль, заимствующий сильные стороны обоих подходов. Он был создан в 2022 году [Короткова 2022].

Ссылка: <https://rupi.org/project/ru-accent-poet/>

Модуль решает задачу постановки ударения, используя встроенный словарь и подключение модели рекуррентной нейросети, где это необходимо. Такой подход сразу повышает качество результата, метрики приведены в статье [Короткова 2022].

Единственным минусом обозреваемого программного обеспечения является скорость работы. Чтобы можно было сразу воспользоваться размеченными текстами, не дожидаясь, пока отработает программа, в 2024 году был опубликован датасет текстов классической русской прозы, размеченной при помощи *ru-accent-poet*. [Orekhov 2024b], кроме того, на этих данных построены уже опубликованные исследования [Орехов 2022].

Модуль для определения прямой речи персонажей в художественном тексте

Зона прямой речи является значимой и маркированной зоной в тексте художественной прозы. Ей было посвящено несколько работ в течение прошлого десятилетия (например, [Хисамова 2013]), некоторые из них были количественными: [Sobchuk 2016], но от-

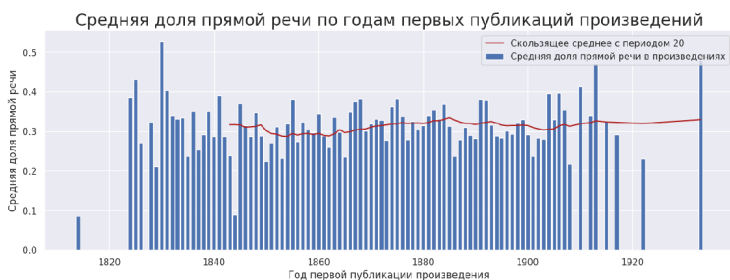


Рис. 1. Средняя доля прямой речи по годам первых публикаций произведений (рисунок автора модуля Д. Есяян)

существование возможности создавать прозрачные и репрезентативные выборки ограничивало свободу маневра исследователей. Обозреваемый модуль призван решить эти проблемы, одновременно сформулировав пользу для другого рода операций, связанных, например, с областью выравнивания двуязычных текстов для параллельных корпусов.

Ссылка: <https://pypi.org/project/direct-speech-extractor-ru/>

У модуля высокие значения метрик оценки: $accuracy = 0.989$, $precision = 0.998$, $recall = 0.96$ и $f\text{-measure} = 0.979$. При этом повторенное с помощью этого инструмента исследование [Sobchuk 2016] не подтвердило его выводов: на протяжении всего XIX века реплики литературных персонажей русской прозы составляли приблизительно одинаковую долю в тексте объемом в 30 % (см. рис. 1) в то время как О. Собчук пишет о ярко выраженном скачке этой доли.

Таким образом, можно констатировать, что сферу цифровых гуманитарных исследований догнал кризис воспроизводимости [Pashler, Wagenmakers 2012].

Модуль для оценки формульности фольклорного текста

Ссылка: <https://pypi.org/project/formularity-rfs/>

Формульность (повторяемость, стереотипность) текста в фольклоре — одна из системообразующих характеристик, она определяет облик текста, облегчает его устное бытование и передачу от одного носителя другому. Цифровой гуманитарный подход подразумевает

возможность перевести это понятие в термины исчисляемых категорий, что открыло бы возможность для сравнения разных текстов по параметру формульности.

Функциональность обозреваемого модуля построена на авторском алгоритме количественной оценки формульности текста, включающем разнообразные текстовые метрики: коэффициент вариативности словарного запаса (VocD), коэффициенты n-грамм, биномов, фразеологизмов и междометий.

Так выглядит размеченный программой текст (полужирным обозначены формульные в рамках анализируемого текста n-граммы и фразеологизмы из встроеного в модуль списка):

Под ракитою зеленой Русский раненый лежал, Ко груди, штыком пронзенной, Крест свой медный прижимал. Кровь лилась из свежей раны На истоптанный песок; Над ним вился **черный ворон**, Чуя лакомый кусок. «Ты не вейся, **черный ворон**, Над моею головою! Ты добычи не дожدهшься, Я солдат еще живой! Ты слетай в страну родную, Отнеси маменьке поклон. Передай платок кровавый Моей женке молодой. Ты скажи: она свободна, Я женился на другой. Я нашел себе невесту **В чистом поле**, под кустом; Моя сваха – востра сабля, И венчал граненый штык; Взял невесту тиху, скромну И приданно небольшо. Взял приданно небольшое – Много лесу и долин, Много сосен, много елок, Много, много вересин»

Найденные n-граммы:

- **черный ворон**

Коэффициент формульности: 1.207

В таблице 1 даны агрегированные результаты анализа нескольких текстовых произведений.

Название песни	Биномы	n-граммы	VocD	Итог
У зари-то, у зореньки	0	0,1	0,786	0,886
По улице мостовой	0,1	0,3	0,862	1,262
Выйду ль я на реченьку	0,1	0,6	0,844	1,544
Вниз по матушке по Волге	0	1,3	0,688	1,988
Не одна во поле дороженька	0	0	0,643	0,643

Название песни	п-	Биномы	граммы	VocD	Итог
----------------	----	--------	--------	------	------

Таблица 1. Результаты анализа песен из сборника «40 русских народных песен» (1988), сост. Юрий Зацарный (результаты получены автором модуля В. Сидненко)

Модуль зависит от нескольких сторонних пакетов: *spacy*, *pytorch2*, *numpy*, *sklearn*, что грозит потенциальными проблемами с установкой и обновлениями, но так как речь идет о достаточно известных и широко распространенных пакетах, есть надежда, что сообщество поможет с решением.

Модуль для конвертации текстов в формате TEI

Ссылка: <https://pypi.org/project/TEItransformer/>

TEI на основе XML — это конвенциональный формат хранения оцифрованных текстовых данных. С самого начала его появления подразумевалось, что сохраненные таким образом произведения будут легко конвертировать в любой другой необходимый формат, предназначенный для обработки или публикации. На деле универсальных решений скроссформатной конвертации долгое время не существовало. Модуль *TEItransformer* призван заполнить эту лакуну [Kostyanitsyna, Skorinkin 2024]. Было бы удобно, если бы он был построен на чистом Python или хотя бы использовал в качестве зависимостей только модули Python, но в реальности создать такой конвертер оказалось слишком сложно, поэтому *TEItransformer* представляет собой, по сути, обертку вокруг XSLT-преобразований, устаревающей и ненадежной технологии обработки XML.

Модуль позволяет конвертировать исходный TEI XML в docx (универсальный редактируемый формат), html (удобный для веб-публикации), json (универсальный формат для анализа данных). Для этого используются пользовательские сценарии, которые описывают, как именно следует преобразовывать исходный текст. При этом текст, преобразованный в html, включает не только статическое представление строк, но и инструменты поиска по разным полям исходной разметки. На рис. 2 показано, как можно искать все реплики персонажа Маша на веб-странице, сгенерированной из TEI-представления драматического текста.

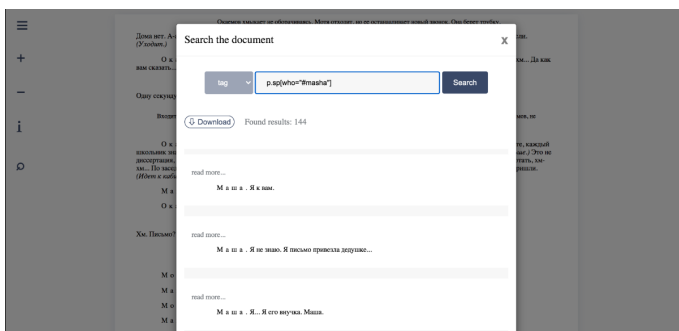


Рис. 2. Поисковый интерфейс на веб-странице, сгенерированной модулем *TEItransformer*

«Цифровые гуманитарные исследования»

Модуль предобработки текстовых материалов для журнала «Цифровые гуманитарные исследования»

Ссылка: <https://pupi.org/project/cgi-processor/>

Настоящий журнал готовится с помощью издательской системы *Latex*. Помимо очевидных редакторских действий подготовка включает и ряд технических преобразований, подразумевающих соответствие правилам отечественной типографики. Эти правила включают неразрывные пробелы определенной длины между фамилией и инициалами, короткое тире в числовых диапазонах и т. д. В *Latex* эти условия задаются при помощи специальных символов. К примеру, короткое тире в исходном коде *Latex* выглядит как два дефиса подряд. Ручная обработка текста, имеющая целью расстановку в тексте всех специальных символов, представляет собой трудоемкий процесс, в котором неизбежны множественные ошибки. Эту проблему призван решить модуль *cgi-processor* [Orekhov 2025]. В первую очередь он обрабатывает те случаи, которые специально учтены в системе стилей журнала «Цифровые гуманитарные исследования», хотя, вероятно, будет полезен и другим авторам и редакторам, работающим с русскоязычными текстами в *Latex*.

Обозреваемый модуль облегчает и ускоряет подготовку номера журнала, работающего на инфраструктурное единство и информационную проницаемость русскоязычного сообщества *Digital Humanities*.

Примечания

- ¹ Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ
- ² <https://app.transkribus.org/models/public/text/44358>
- ³ <https://app.transkribus.org/models/public/text/russian-print-xviii-cent>

Литература

Исследования

- Короткова 2022* — Короткова Ю. О. Комбинированный словарно-нейросетевой акцентуатор для разметки русского поэтического текста // Труды Института русского языка им. В. В. Виноградова. 2022. № 3 (33), С. 181–190. DOI: 10.31912/pvrl-2022.3.11
- Орехов 2022* — Орехов Б. В. Случайные метры в русской прозе XIX века // Вещество поэзии: К 70-летию Юрия Борисовича Орлицкого: Сборник научных статей. М.: РГГУ, 2022. С. 24–30.
- Хисамова 2013* — Хисамова Г. Г. Функции диалога в художественном тексте // Вестник Нижегородского университета им. Н. И. Лобачевского. 2013. № 6 (2). С. 245–247.
- Kostyanitsyna, Skorinkin 2024* — Kostyanitsyna A., Skorinkin D. A Python Library for TEI Conversion into Edition Formats // Texts, languages and communities: 24th Annual Conference and Members' Meeting of the Text Encoding Initiative Consortium. Ciudad Autónoma de Buenos Aires: Universidad del Salvador, 2024. P. 114–115.
- Orekhov 2025* — Orekhov B. Модуль предобработки текстовых материалов для журнала «Цифровые гуманитарные исследования» [Computer software]. 2025 URL: https://github.com/nevmenandr/cgi_processor
- Orekhov 2024a* — Orekhov B. Russian-Old-Orthography-Ocr (Revision 6f60636) // Hugging Face, 2024. DOI: 10.57967/hf/3280
- Orekhov 2024b* — Orekhov B. accentual-syllabic-verse-in-russian-prose (Revision 489fea1). // Hugging Face, 2024. DOI: 10.57967/hf/2438
- Pashler, Wagenmakers 2012* — Pashler H., Wagenmakers E. Editors' Introduction to the Special Section on Replicability in Psychological Science // Perspectives on Psychological Science. 2012. N 7. P. 528–530.
- Sobchuk 2016* — Sobchuk O. The Evolution of Dialogues: A Quantitative Study of Russian Novels (1830–1900) // Poetics Today. 2016. N 37 (1). P. 137–154.

ИСТОРИЯ ЦИФРОВЫХ МЕТОДОВ

Андрей Володин

ЦИФРОВЫЕ ГУМАНИТАРИИ: ОТ АКАДЕМИЧЕСКИХ ПЛЕМЕН К ЭПИСТЕМИЧЕСКОМУ СООБЩЕСТВУ

В статье предлагается периодизация развития цифровых гуманитарных исследований (Digital Humanities, DH) в первой четверти XXI века, основанная на трёх последовательных этапах, каждый из которых объясняется через соответствующую социологическую концепцию науки. Первый этап (2004–2012 гг.) — период самоопределения и формирования «академических племён и территорий», когда DH осознавали себя как междисциплинарное поле с уникальной методологией. Второй этап (2013–2022 гг.) — эпоха «зон обмена», характеризующаяся прагматичным взаимодействием гуманитариев и инженеров через общие артефакты, данные и локальные языки. Третий, современный этап (с 2023 г.) — становление DH как эпистемического сообщества, активно влияющего на академическую политику, стандарты и производство знания. Статья показывает, как DH прошли путь от фрагментированного движения до консолидированного экспертного сообщества, способного отвечать на современные вызовы цифровой эпохи.

Ключевые слова: цифровые гуманитарные исследования, цифровые гуманитарные науки, digital humanities, академические племена, зоны обмена, эпистемическое сообщество, периодизация эволюции научной дисциплины, междисциплинарность

Varietas delectat
Phaedrus

Андрей Юрьевич Володин
МГУ, СФУ
mailbox@volodin.info

DOI: 10.31860/cgi-2025-2-84-116

Мир¹ гуманитарного знания, веками хранимый в тишине архивов, библиотек и музеев, в начале XXI столетия услышал новый ритм — дискретный цифровой пульс. Под его воздействием начался не просто перевод в цифровые форматы старых текстов, изображений и артефактов, а рождение нового направления на карте науки — территории, где алгоритмы стали соавторами интерпретаций, а данные — новым языком диалога между эпохами. Эта история не о том, как компьютер пришёл в библиотеку, а о том, как сама мысль, встретившись с вычислительной мощью, отправилась в долгое, сложное и прекрасное путешествие — путь цифровых гуманитарных исследований, чьи этапы напоминают географические открытия: от описания первых академических «племён» и «территорий» до составления общего эпистемического атласа современного гуманитарного знания.

Периодизация и концептуализация развития цифровых гуманитарных исследований

Периодизация любого процесса всегда условна, но при этом полезна. Увидеть некоторые закономерности в многообразии изданий, дискуссий, манифестов и критических высказываний в Digital Humanities важно, чтобы хотя бы ненадолго понять, как складывается фарватер цифровых гуманитарных исследований².

В социологии науки разработано немало моделей жизненного цикла научных направлений, обычно они состоят из четырех этапов: нулевого — пионерского³, первого — самоидентификационного, второго — формационного и третьего — устойчивого (или завершающего) [Плотинский 2001; Mey 1992].

Сложность построения периодизации такого разнородного направления как Digital Humanities связана с тем, что есть слишком большое количество точек зрения на научный рост этого направления. Можно стараться смотреть свысока, наблюдая исключительно за глобальными изменениями (если они действительно существуют и имеют какой-то смысл за пределами англоязычного мира), можно смотреть за региональными тенденциями (чаще всего речь идет о европейской традиции, но в последнее время уделяется особое внимание и азиатским вариациям, прежде всего, в Индии, Китае, Японии и Южной Корее), можно сосредоточиться на страновых различиях (например, о ДН в России написано уже немало [Gritsenko et al. 2022; Kizhner et al. 2022; Пильщиков 2022; Skorinkin 2023; Колозариди & Беляк 2024; Орехов & Володин 2024]). В этой ста-

тье предпринимается попытка увидеть общие тенденции, которые характеризуют направления развития Digital Humanities в первой четверти XXI века, с точки зрения тех, кто к этому направлению себя причисляет или хотя бы тяготеет. Причем понятно и логично, что ни один из читателей не согласится с представленной в статье трактовкой развития ДН, но весьма вероятно увидит что-то знакомое и похожее на тот уникальный путь цифровых гуманитарных исследований, в котором участвовал сам.

Первый этап развития цифровых гуманитарных исследований можно начать 2004 годом, когда увидел свет «Компаньон по цифровым гуманитарным наукам» [Companion 2004], когда был создан Альянс ДН-организаций (2005) и состоялся первый конгресс “Digital Humanities – 2006”, и завершить 2011–2012 годами, когда появились «Манифест Digital Humanities» [Манифест 2011], вышел первый сборник из серии “Debates in the Digital Humanities” [Debates 2012], а также был опубликован учебник “Digital Humanities” [Burdick et al. 2012].

Для описания этого этапа развития ДН удачно подходит *концепция «академических племен и территорий»*, разработанная Тони Бекером и позже развитая в соавторстве с Полом Троулером [Becher & Trowler 2001; Trowler et al. 2012]. Суть концепции заключается в том, что академическое сообщество — это не единая монолитная структура, а совокупность разнообразных «племен» (сообществ ученых), каждое из которых обладает своей уникальной «территорией» (дисциплиной или областью знаний) со специфической культурой, ценностями, языком и способами познания. Смысл такой концепции состоит в том, чтобы посмотреть на академический мир через призму единства когнитивных (знание) и социальных (сообщество) аспектов, где каждая дисциплина представляет собой уникальную культурную и интеллектуальную экосистему со своими правилами игры. Дисциплинарные ландшафты знаний отличаются предметами изучения (что именно исследуют ученые в конкретной области?), методами и методиками (как именно они это исследуют?), структурой знания (как организована сама дисциплина?), темпами развития (быстро или медленно меняется область знаний?). Сообщества ученых, которые населяют эти дисциплинарные территории, характеризуются академической культурой (неформальными правилами, традициями, нормами поведения), ценностями и идеалами (прежде всего, что именно считается в дисциплине «хорошей работой», как определяется научная репутация, что важнее — публикация в престижном журнале или приклад-

ное значение исследования), языком и жаргоном (специфический язык, который одновременно является инструментом работы и системой опознавания «свой-чужой»), процессом социализации (как новые члены сообщества (например, магистранты, аспиранты) обучаются не только методам, но и негласным правилам «племени» через общение с руководителем, коллегами, участие в конференциях), идентичностью (ученый часто идентифицирует себя в первую очередь именно со своей дисциплиной («я – историк», «я – филолог» и т. д.), а не просто с профессией «ученый»). Главный вывод Бекера и Троулера заключается в том, что «территории» и «племена» неразрывно связаны. Когнитивная структура дисциплины (территория) формирует социальные практики сообщества (племени), и наоборот — социальные нормы и культура сообщества влияют на то, как развивается знание на этой территории. Говоря о гуманитарных науках, они отмечают их размытые общие парадигмы и низкий уровень консенсуса (сравнительно с сообществами ученых в естественных науках), а гуманитарные «племена» более индивидуалистичны, ценят личную интерпретацию и работу в одиночку, а репутация строится на авторитете отдельного ученого.

Второй этап развития цифровых гуманитарных исследований можно условно начать 2013 годом (выход статьи Алана Лю «Значение Digital Humanities» [Liu 2013] обозначил момент, когда внимание сообщества сместилось с технических вопросов об инструментах и методах на фундаментальные философские, культурные и политические вопросы («почему это нужно?», «каковы ценности ДН?», «как ДН соотносится с историей гуманитарной мысли?»), а в конце 2015 года уже появляется сигнальный номер «Нового компаньона по цифровым гуманитарным наукам» [A New Companion 2016], сформулировавший новую «целостную» и «самостоятельную» модель развития направления. Закончить этот период можно постепенным завершением пандемии коронавирусной инфекции COVID-19 в 2022 году.

Этот этап развития ДН можно описать разработанной Питером Галисоном *концепцией «зон обмена»*, которая описывает пространство (как физическое, так и интеллектуальное), где разные научные субкультуры с несовместимыми системами убеждений, языками и практиками могут успешно взаимодействовать для достижения общих целей через создание локальных средств коммуникации [Galison 1997; Галисон 2004]. Традиционная модель науки (например, у Т. Куна) предполагала, что ученые работают в рамках единой «парадигмы», но Галисон, изучая большие научные коллабора-

ции, показал и доказал, что академическое взаимодействие многослойно, и включает различные традиции — теоретизирование, эксперимент, изготовление инструментов и инженерная деятельность — которые зачастую пересекаются и преобразуют друг друга, но при этом не теряют своей самостоятельности. Суть концепции «зон обмена» состоит в том, что для успешного сотрудничества не требуется, чтобы все участники разделяли одни и те же фундаментальные убеждения или полностью понимали друг друга. Как на настоящем рынке, можно успешно торговать с человеком другой культуры, не зная его языка и верований, если у вас есть локальные правила обмена. В науке «валютой» обмена становятся данные, материалы, алгоритмы. Для этих целей в науке появляется «пиджин» — упрощенный язык, необходимый для базового общения между группами с разными родными языками (это могут быть и термины, и протоколы, и форматы данных). Со временем у нового поколения исследователей появляется креольский язык — это пиджин, который стал родным для нового поколения, то есть развился в полноценный язык. Часто появление такого языка связывается с новой гибридной дисциплиной (например, биоинформатикой, когда новый язык понятен и биологам, и программистам). Часто основой для зоны обмена служат не абстрактные идеи, а материальные объекты и стандартизированные практики. Самый простой пример материального объекта — база данных, которая становится платформой для обмена между учеными в конкретной дисциплине и специалистами по информационным технологиям. Фокус внимания концепции смещается с чистого «знания» на материальные артефакты, инструменты и практики, которые являются реальной основой научной кооперации. Галисон показывает, что наука развивается не только через революции и смену парадигм, но и через «горизонтальную» интеграцию разных, иногда даже кажущихся несовместимыми, научных культур. Концепция объясняет, что разнородность и непонимание в науке не являются препятствием, а, будучи грамотно организованными через локальные языки и обмен артефактами, становятся источником инноваций и двигателем научного прогресса.

Третий этап развития цифровых гуманитарных исследований происходит на наших глазах. Начался этот период, с одной стороны, с возобновления очных ежегодных ДН-конгрессов (в 2023 году состоялся в австрийском Граце [Володин 2023a], в 2024 году — в Вашингтоне [Володин 2024], в 2025 году — в Лиссабоне [Володин 2025a]) и нового витка активного международного сотрудничества.

А с другой стороны, цифровые гуманитарии оказались в центре дискуссий о смысле гуманитарного знания в свете замечательных успехов машинного обучения и генеративных предобученных трансформеров как раз в 2023 году. Современный этап развития будет продолжаться, до условной середины 2030-х годов, когда, как уже было предсказано в предисловии к «Новому компаньону по цифровым гуманитарным наукам», весьма вероятно, эпитет «цифровой» может оказаться плеоназмом [A New Companion 2013 P. xvii].

Возможным способом понимания текущего этапа развития ДН может послужить разработанная Питером Хаасом *концепция «эпистемического сообщества»*, которая заключается в том, что группы экспертов, объединенные общими научными убеждениями и пониманием актуальной проблематики, могут оказывать решающее влияние на формирование академической политики, особенно в сложных технических областях [Haas 2015]. Эпистемическое сообщество отличается общими принципиальными убеждениями, ценностями и представлениями о том, что является правильным или желательным. Ключевые ценности эпистемического сообщества: ценность действенного метода, ценность нового знания, ценность экспертной дискуссии. Члены сообщества используют взаимопонятные методы и стандарты для оценки достоверности информации, которые позволяют им приходиться к консенсусу по поводу новых данных. При этом, эпистемическое сообщество не просто изучает проблему, но и стремится к реализации конкретных мер для ее решения, участников сообщества объединяет общее видение желаемого результата, обращая первостепенное внимание на роль качественных знаний и комплексной экспертизы. Суть концепции Хааса состоит в том, что в современном сложном мире власть основана не только на экономических или политических рычагах, но и на знании. Эпистемические сообщества могут играть роль «переводчиков» сложной реальности для лиц, принимающих решения, которые помогают определить проблему, ее причины и предлагают конкретные решения, тем самым формируя повестку дня.

Первое десятилетие цифровых гуманитарных исследований: племена и территории (само)определяются

Академические территории представляются как дисциплинарные ландшафты знаний. Речь идет о самом содержании дисциплин, с определенным предметом изучения, принятыми методами, с до-

статочной строгой структурой знания и привычным темпом развития. Племенами мы называем сообщества ученых, которые населяют эти территории. Такие сообщества характеризуются своей академической культурой, ценностями и идеалами, профессиональным языком и жаргоном. Концепция академических племен и территорий удачно подходит для анализа фазы становления цифровых гуманитарных исследований как динамичной и гибридной научной области. Ведь цифровые гуманитарные исследования изначально не были устоявшейся «территорией» с понятными методологическими границами, а представляли собой постоянное «пограничье», где происходит столкновение и смешение нескольких разных «племен».

Период с 2004 года стал временем самоопределения цифровых гуманитарных наук. В первое десятилетие с момента провозглашения понятия “digital humanities” в первом компаньоне по цифровым гуманитарным наукам [A Companion 2004], исследовательское поле цифровых гуманитариев перестало быть просто набором технических экспериментов в рамках традиционных гуманитарных дисциплин, а начало осознаваться как самостоятельное движение с собственными методологическими спорами, теоретическими парадигмами и внутренними конфликтами. Центральным процессом этого периода стало поиск самоопределения, которое выразилось в переходе от термина “Humanities Computing” (условно «гуманитарная информатика») к более широкому и интеллектуально заряженному “Digital Humanities” («цифровые гуманитарные науки» или «цифровые гуманитарные исследования»). Такой терминологический сдвиг символизировал не просто технологическое обновление, но и методологический поворот. Произошло смещение акцента с простого применения компьютеров как инструмента при решении гуманитарных научных задач на исследование в широком смысле взаимодействия человека и машины, а также на то, как именно цифровые технологии изменяют саму природу человеческого знания. Такой переход требовал от участников движения не только технических навыков, но и способности к рефлексии, теоретизации и критическому анализу самих цифровых практик.

Одним из самых значительных методологических событий этого периода стала публикация в 2004 году коллективной монографии “A Companion to Digital Humanities” под редакцией Сьюзан Шрайбман, Рэя Сименса и Джона Ансворта. Книга стала своего рода программным документом для нового поколения исследователей, потому что не просто собрала различные проекты и подходы, но и попыталась

систематизировать новое поле, объединив его вокруг общих методологических проблем и междисциплинарных интересов. В книге прозвучали голоса самых разных специалистов в области гуманитарных исследований: от лингвистов до археологов, объединенных идеей, что за пределами традиционных гуманитарных вопросов существует общее ядро методологических задач, связанных с созданием, анализом и сохранением гуманитарных объектов исследований в цифре. Таким образом, «Компаньон» установил новый стандарт сложности, изменив отношение к ДН как простой технической поддержке традиционной науки, утверждая новый уровень качества интеллектуальных результатов. «Компаньон» заложил основы для дальнейших дебатов, которые вскоре начали наполнять поле содержанием.

Понятие “digital humanities” можно назвать изобретенной традицией. Как показали историки Э. Хобсбаум и Т. Рейнджер, многие традиции, кажущиеся давними или претендующие на то, что они являются таковыми, нередко оказываются изобретёнными недавно [Hobsbawm & Ranger 1992]. Если в предыстории цифровых гуманитарных исследований можно встретить не только Р. Бузу, но и Ч. Бэббиджа, А. Лавлейс, а иногда и отсылки к Абу Абдуллаху Мухаммаду ибн Мусе аль-Хорезми, то сам термин имеет вполне понятную датировку — 2004 год. Именно благодаря «Компаньону по цифровым гуманитарным наукам» понятие “digital humanities” разлетелось по научному миру. А уже в «Новом компаньоне по цифровым гуманитарным наукам» редакторы С. Шрибман, Р. Сименс и Дж. Ансворт вспоминают: «Оглядываясь назад, становится ясно, что решение группы редакторов, подсказанное издателем, назвать оригинальный „Компаньон“ (2004 года – А.В.) изменило наш подход к этой области: мы перестали говорить о „гуманитарной информатике“ (“humanities computing” – А.В.) и стали говорить о „цифровых гуманитарных науках“ (“digital humanities” – А.В.). Редакторы этого и предыдущего томов, в разговоре с издателем, выбрали именно такое название для деятельности, представленной в нашем коллективном труде, чтобы сместить акцент с „вычислительной техники“ (“computing” – А.В.) на „гуманитарные науки“ (“humanities” – А.В.)» [A New Companion P. xvii].

На эту перемену самоназвания и был расчет в 2004 году, когда те же редакторы отмечали: «Эта коллекция („Компаньон по цифровым гуманитарным наукам“ – А.В.) знаменует собой поворотный момент в области цифровых гуманитарных наук: впервые широкий круг теоретиков и практиков, как тех, кто работает в этой

области уже десятилетиями, так и тех, кто недавно занялся ею, экспертов по дисциплинам, специалистов по информатике, библиотековедению и информационным исследованиям, собрались вместе, чтобы рассмотреть цифровые гуманитарные науки как самостоятельную дисциплину, а также поразмышлять о том, как они соотносятся с областями традиционных гуманитарных исследований» [A Companion P. xiii]. На страницах первого компаньона можно встретить размышления и о классической филологии, и об истории искусства, и об изменениях в археологии, и в лексикографии, и в литературоведении, и в исследованиях музыки и мультимедиа. Разнообразие сфер применения цифровых подходов наталкивает на мысль о периоде, когда цифровые гуманитарные исследования находились как раз в ситуации академических племен и территорий.

Центральным методологическим конфликтом, который развернулся в середине рассматриваемого периода, стала дихотомия между «медленным чтением» (*close reading*) и «дальним чтением» (*distant reading*). Концепция дальнего чтения, популяризированная и развитая литературоведом Франко Моретти и его коллегами из *Stanford Literary Lab*, представляла собой важный методологический вызов [Moretti 2013, Моретти 2016]. Вместо глубокого герменевтического анализа одного или нескольких текстов, дальнейшее чтение предлагало анализировать огромные корпуса литературных произведений с помощью алгоритмов для выявления масштабных паттернов, тем и структур, которые невозможно заметить при традиционном прочтении. Моретти в своих работах, таких как «Графы, карты, деревья» (Moretti 2007), продемонстрировал, как сетевая теория, тематическое моделирование и другие количественные методы могут использоваться для исследования литературной истории в ее целостности. Идея вызвала бурную реакцию. С одной стороны, она была воспринята как мощный инструмент для преодоления ограничений канона и изучения того, что Моретти назвал «великим непрочитанным». С другой стороны, сторонники традиционного герменевтического подхода возражали, что такой количественный анализ лишает текст контекста и не может заменить искусство глубокого понимания. Дискуссия вышла далеко за рамки литературоведения и затронула фундаментальные вопросы методологии в гуманитарных науках, заставив исследователей задуматься о роли интерпретации в эпоху больших данных.

На фоне этих методологических дебатов все большее значение приобретало понятие «смешанных методов» (*mixed methods*). Стало

очевидно, что ни один подход не является универсальным решением. Исследователи начали осознавать, что количественные данные должны быть осмыслены в качественном контексте, а качественные данные могут быть проанализированы с помощью количественных инструментов для выявления скрытых закономерностей. Этот подход, заимствованный из социальных наук, стал одним из ответов на критику, направленную против чрезмерной зависимости цифровых гуманитариев от алгоритмов. Концепция смешанных методов в гуманитарных науках явно стремилась найти баланс между человеческой интуицией и технологическими возможностями. Более того, методологические дискуссии стали выходить за рамки конкретных техник анализа текстов, началось осознание того, что сами данные, используемые для анализа, могут быть источником как возможностей, так и проблем. Например, Д. Гадд в своей работе «Использование и злоупотребление Early English Books Online (EEBO)» выступил с жесткой критикой популярного цифрового корпуса, указав на его ограниченную выборку и потенциальную предвзятость, которая может исказить результаты исследования [Gadd 2009]. Аналогичные проблемы стали возникать всё чаще, например, голландский проект *Pidemehs*, изучающий политическую культуру XX века, был вынужден ограничить свой анализ материалами до 1945 года из-за правовых ограничений на доступ к более свежим материалам, что создавало серьезные проблемы с репрезентативностью данных.

Другой серьезной методологической проблемой была высокая частота ошибок при оптическом распознавании символов (OCR), особенно при работе со старинными текстами. В некоторых случаях процент ошибок OCR превышал 80 %, что делало автоматизированный анализ практически невозможным без последующей доработки и верификации данных человеком. Например, в одном из проектов по анализу голландских газет XVIII века неверное чтение слова “*verzuijing*” привело к появлению ложных совпадений и искажению исторических выводов. Такие проблемы подчеркивали, что даже самые передовые цифровые методы требуют тщательной проверки и критического отношения к первоначальному материалу. Они также демонстрировали, что цифровые гуманитарные исследования не могут стать «автоматической» наукой, а требуют от исследователя тех же навыков критики источников, что и традиционные гуманитарные дисциплины, но в новой, цифровой среде.

Параллельно с развитием новых методов анализа текстов, происходила активная работа над методологией визуализации данных. Здесь главным теоретиком и одновременно критиком цифровых гуманитарных наук выступила Джоанна Дрюкер. В статье «Человеческие подходы к графическому отображению» она призвала отказаться от механистических и чисто статистических моделей визуализации в пользу подходов, которые бы сохраняли многозначность, неопределенность и эстетику оригинального текста [Drucker 2011]. Она считала, что цифровые гуманитарные науки должны интегрировать в свои методы гуманистические ценности, такие как вероятностный характер и исполнительская природа знания, которые часто игнорируются в строгих количественных подходах. В этой статье Дрюкер обратилась к понятию «капта» (*capta*), описывающему самостоятельно собранные исследователем данные, противопоставляя капту данным, которые кем-то даны, где-то загружены, получены «готовыми». Критика была частью более широкого движения по интеграции гуманистического мышления в саму технологию, что стало одной из ключевых идей Аллана Лю. Лю утверждал, что цифровые гуманитарные науки должны не просто использовать технологии, но и критически анализировать их, занимаясь культурной критикой самих технологических платформ, корпоративного контроля и неолиберальных тенденций в образовании [Liu 2012]. Он призывал цифровых гуманитариев обращаться к таким областям, как наука и технологии (STS) и медиа-археология, чтобы развивать более глубокую методологическую базу, основанную на понимании того, как технологии формируют нашу реальность. Стивен Рэмзи ввел понятие «алгоритмической критики», предполагающее, что компьютерные трансформации текста должны служить основой для интерпретации, а не для подтверждения уже сформулированных истин [Ramsay 2011]. Таким образом, в первое десятилетие развития ДН методология цифровых гуманитарных наук представляла собой сложный, фрагментированный и постоянно развивающийся ландшафт, полный как вдохновляющих инноваций, так и серьезных этических и методологических вызовов.

С одной стороны, возникала борьба за легитимность. Территория ДН — это область, оспариваемая у традиционных гуманитарных наук, где ключевым аргументом становится, что ДН не просто использует инструменты, а задает новые исследовательские вопросы. С другой стороны, возникают пограничные конфликты, которые связаны с вопросами, кто может называться «цифровым гума-

нитарием»? Достаточно ли использовать цифровые инструменты, или нужно иметь соответствующее образование и публикации в соответствующих журналах? При этом начинает формироваться язык как маркер освоения новой территории. Возникает специфический устойчивый жаргон (TEI, OCR, GIS, sentiment analysis, distant reading), маркирующий принадлежность к научному направлению.

Главным глобальным институциональным решением этого этапа стало создание Альянса ДН-организаций (Alliance of Digital Humanities Organizations, ADHO) в 2005 году. Первая ежегодная конференция под названием “Digital Humanities” состоялась в 2006 году в Университете Сорбонна (Париж, Франция). (До этого момента с 1989 года конференции проходили под эгидой двух ассоциаций — АСН (Association for Computers and the Humanities) & ALLC (Association for Literary and Linguistic Computing). С этого момента конференции ADHO становятся местом встречи цифровых племен, где они ищут общий язык. Постепенно складывается и общая инфраструктура, возникают первые ДН-репозитории на GitHub (сам сервис появляется в 2008 году), появляются специализированные проекты вроде Programming Historian, и другие общие территории. Стоит отметить, что статус участников сообщества определяется не только публикациями, но и созданием работающих инструментов, цифровых архивов и успешных цифровых проектов.

Однако разногласия относительно определения границ Digital Humanities закрепились в связи с проектом Мэтью Голда, профессора английского языка и цифровых гуманитарных наук, из Нью-Йоркского городского колледжа (CUNY), открывшего сайт “What Is Digital Humanities?” в 2010 году. Более восьмисот определений ДН были собраны краудсорсинговым методом накануне ежегодной конференции Альянса цифровых гуманитарных наук (ADHO). В 2009–2010 годах в ДН-сообществе шли активные дебаты о том, как именно определить быстро развивающуюся область. Вместо того чтобы искать одно единственное определение, М. Голд решил продемонстрировать его множественность и разнообразие. Он и его коллеги использовали открытую таблицу, доступную для редактирования любым желающим. Ссылка на эту таблицу была распространена через рассылку по почтовым спискам сообщества (в первую очередь, по списку рассылки Humanist; сама рассылка была создана Уиллардом Маккарти в 1987 году) и в социальных сетях. Участников конференции ADHO и всех заинтересованных лиц попросили добавить в таблицу свои краткие определения Digital Humanities. Сообщество откликнулось очень активно, и в корот-

кие сроки были собраны сотни формулировок. Сайт устроен очень просто: при каждом обновлении страницы случайным образом выбирается одна из этих дефиниций. Некоторые из них серьёзные, некоторые ироничные, некоторые технические, а некоторые — поэтические. Вместе они создают «портрет» дисциплины этапа самоопределения через призму мнений самих её представителей.

Ответом на такую разногласицу стал опубликованный на многих языках в 2011 году «Манифест Digital Humanities»: «“Цифровая революция” современного общества видоизменяет и ставит под вопрос традиционные формы создания и распространения знаний. По нашему мнению, цифровые методы исследований имеют значение для всех гуманитарных наук. Digital Humanities развиваются не с „чистого листа“. Напротив, они опираются на все научные парадигмы, знания и умения, накопленные каждой из соответствующих научных дисциплин, используя инструменты и перспективы, открывшиеся благодаря цифровым технологиям. Цифровые гуманитарные науки по определению междисциплинарны и несут в себе все методы, средства и перспективы познания, связанные с цифровыми технологиями в области гуманитарных наук» [Манифест 2011]. Для разных академических племен «Манифест» был выражением надежды для объединения «под зонтиком» Digital Humanities (много образов возникло из определения ДН как зонтичного термина или *umbrella term*). «Манифест» достаточно удачно сформулировал повестку дня на следующее десятилетие, закрепив представление, что ДН — это не про компьютеры в гуманитарных науках, а про новые формы мышления и познания, что практика создания может быть методом исследования, что критика должна быть неотъемлемой частью ДН, а сами ДН должны быть публичными, открытыми и вовлекающими сообщества за пределами академии.

В концепции «племен и территорий» ДН видится не как монолит, а как конгломерат племен, борющихся за определение своей территории и легитимность внутри академии. Причем внутреннее размежевание (условно кодеры vs. интерпретаторы) столь же важно, как и внешние противопоставления (ДН vs. традиционные гуманитарии) [Бурдье 2018].

В 2012 году публикуется “Digital_Humanities” — учебник и своего рода дорожная карта для цифровых гуманитарных наук в момент их трансформации. Авторы (А. Бердик, Дж. Дрюкер, П. Лунефельд, Т. Пресснер, Дж. Шнапп) сравнивают переживаемую эпоху с изобретением печатного станка, когда гуманитарные

знания и методы получают беспрецедентное расширение благодаря цифровым технологиям. Текущую ситуацию авторы называют концом «нормальности». Время, когда о ДН можно было говорить как о новости, подходит к концу. Работа с цифровыми инструментами и средами становится повседневной нормой для гуманитариев. Однако это изменение таит риск утраты критического и экспериментального духа, который был движущей силой ранних ДН. Поле ДН расширяется, потому что речь идет не просто про применение компьютеров в традиционных дисциплинах, а качественное расширение самого предмета и методов: акцент смещается с индивидуального, текстоцентричного анализа на коллаборативное создание, дизайн и визуализацию знаний в мультимедийных форматах. Ближайшим будущим видятся «генеративные гуманитарные науки», ядром современной практики ДН становится создание прототипов, итеративность, принятие продуктивных неудач, работа с версиями. Знание создается в процессе дизайна и производства (например, платформ, баз данных, визуализаций), а не просто описывается. Нельзя не обращать внимания, что ДН радикально меняют социальную жизнь гуманитарного знания, появляется распределённое авторство, открытые сетевые экономики знаний, стирание границ между академией и публичной сферой, вовлечение сообществ. Более того, любая публикация результатов гуманитарного исследования может становиться социальным, интерактивным актом. Ближайшие перспективы развития цифровых гуманитарных исследований виделись авторам в создании новых методов и жанров. Причиной тому необходимость развивать не только количественные методы (анализ больших данных), но и качественные, гуманистические вычислительные инструменты, способные работать с нюансами, субъективностью, многозначностью, что является сутью гуманитарного знания. Высказывалось требование пересмотреть системы академического признания (продумать критерии оценки проектов, а не только статей и монографий) в ситуации, когда будущее — за гибридными специалистами. Тем не менее, авторы видели в ДН большие перспективы, потому что именно цифровые гуманитарные исследования должны стать основой обновлённого гуманитарного ядра в образовании XXI века, формируя критически мыслящих, медиаграмотных граждан, способных к синтезу и творческому решению проблем в сетевом мире. При этом ДН-сообщество находится на передовой борьбы за открытый доступ к историко-культурному наследию, требующий реформирования авторского права и проти-

водействия коммерческой изоляции знаний, потому что будущее сообщества связано со свободой исследований современной культуры. Главный посыл книги, что Digital Humanities не техническая дисциплина, а масштабный гуманитарный проект по переосмыслению производства и социальной роли знания в цифровую эпоху, и необходимо сохранять в нём дух критики, экспериментирования и социальной ответственности.

Также в 2012 году выходят “Debates in the Digital Humanities” под редакцией Мэтью К. Голда в гибридном формате — как печатная книга и как открытый онлайн-ресурс, что само по себе было символическим жестом, воплощающим принципы обсуждаемой дисциплины [Gold 2012]. Первый том (будущей серии, которая ещё не планировалась) не просто документировал, а активно формировал дисциплину. «Дебаты» перевели разговор с вопроса «Как мы это делаем?» на вопрос «Почему мы это делаем и каковы последствия?». Благодаря сборнику внутренние споры в(округ) ДН стали достоянием широкой академической общественности. В итоге полифония дебатов стала точкой отсчета для любого серьезного разговора о цифровых гуманитарных науках. По сути, этот сборник был моментом взросления ДН, когда сообщество осознало себя достаточно зрелым, чтобы начать критиковать и исследовать само себя. До 2012 года ДН существовали как динамичное, но достаточно разрозненное поле: были «пионеры» с техническими проектами, теоретики, скептики и множество экспериментаторов. Сборник стал важной попыткой собрать воедино, каталогизировать и осмыслить основные направления, методы и, что самое главное, споры внутри формирующегося поля.

Сборник, стал знаковой публикацией для своего времени, потому что представил не перечень определений о цифровых гуманитарных науках, а живую дискуссионную площадку, зафиксировавшую ключевые споры и точки роста ДН. Книга завершает период самоопределения и легитимации ДН в академическом сообществе. Её центральная задача — не дать окончательные ответы, а структурировать дебаты между так называемыми «дигитализирующими» гуманитариями (использующими цифровые инструменты для традиционных задач) и «дигитальными» гуманитариями (стремящимися к методологическому переосмыслению исследований под влиянием вычислительных методов). Структурно сборник делится на три ключевых раздела: «Дебаты», включающие теоретические и идеологические вызовы, критику ДН со стороны традиционных гуманитариев, вопросах нового канона, институционального при-

знания, проблеме «двойного следа» (публикация в цифре и печати); «Практики», где демонстрируются конкретные исследовательские проекты и методы, от анализа культурных сетей до создания цифровых архивов, показывая ДН в действии; «Профессии», рассматривающие влияние ДН на академические карьеры, педагогику, библиотечное дело и вопросы открытого доступа. Главная ценность издания 2012 года — в его полемичности и диалогичности. В нём сталкиваются позиции скептиков (Стивен Рэмси, Дэниел Оллуэй) и апологетов (Джоанна Дрюкер, Тара Макферсон), обсуждаются риски инструментализации гуманитарного знания и его новые возможности. Книга зафиксировала момент превращения ДН в модный и влиятельный тренд, сопровождающийся как энтузиазмом, так и напряжённой рефлексией.

Название «Дебаты» стало программным. Сборник отошел от чисто технического или апологетического взгляда на ДН сфокусировался на критическом осмыслении [Liu 2012]. Высказывались разные точки зрения. Звучала критика «больших данных» (являются ли ДН просто сервисом по оцифровке для традиционных гуманитариев?), обсуждались вопросы институциональной власти (кто получает гранты и чьи проекты оказываются в центре внимания, не воспроизводят ли ДН традиционные иерархии?), как теория соотносится с практикой (есть ли в ДН место для критической теории? является ли «кодирование» новой формой «чтения?»), может ли сообщество стать инклюзивным (насколько ДН открыто для женщин, расовых и других меньшинств?). Таким образом, сборник стал манифестом (хотя и не назывался так, как программный документ 2011 года) «критических цифровых гуманитарных наук», которые не просто используют инструменты, но и рефлексиируют над их влиянием на знание и общество. Серия продолжается до сих пор (тома выходили в 2016 [Gold & Klein 2016], 2019 [Gold & Klein 2019], 2023 годах [Gold & Klein 2023]). Каждый новый том фиксирует состояние дисциплины на новый момент времени, но именно пионерская роль первого издания 2012 года остается непревзойденной.

Таким образом, первое десятилетие самопровозглашенных Digital Humanities напоминало не планомерное освоение тегга nullius, а бурный процесс собирания племен вокруг нового концептуального костра. Этот костер, разожженный «Компаньоном» 2004 года, осветил общую территорию, но одновременно отбросил тени, обнажив границы между «кодерами» и «интерпретаторами», между количественным «дальним чтением» и герменевтическим

«медленным». Рождающееся направление проходило достаточно болезненный путь роста и самоопределения от первых сборок цифровых инструментов в руках разных племен к становлению собственного методологического ландшафта.

Второе десятилетие цифровых гуманитарных исследований: зоны обмена

В междисциплинарных проектах «племена» с разными исследовательскими культурами, языками и практиками могут сотрудничать, создавая локальные языки («пиджины» или «креольские языки»), а также понятные протоколы для обмена ресурсами и решения конкретных задач без необходимости глубокого понимания мировоззрения друг друга.

В этом смысле ДН является типичной «зоной обмена», когда гуманитарий предлагает исследовательские вопросы и знание контекста, а компьютерный специалист поставляет инструменты, методы, инфраструктуру. Предмет обмена также очевиден: гуманитарий получает данные, визуализации, инструменты для анализа, а инженер получает интересную проблему для решения, данные для работы, публикации, а иногда и финансирование. Прогресс направления происходит тогда, когда гуманитарий и инженер вдруг прорастают в одном человеке, а со временем таких людей становится всё больше и больше [Berry 2013].

Взаимодействия в зоне обмена порождают простой, ситуативный язык, например, для конкретного проекта («нужно извлечь все имена собственные из этого корпуса текстов»), а могут стремиться и к более развитому «креольскому языку», на котором говорят, например, цифровые историки, которые в равной мере владеют и историческим методом, и основами работы с шейп-файлами в ГИС [Kemman 2021]. Институционализация «зон обмена» происходит достаточно быстро, например, на рынке труда появляются смешенные позиции вроде «куратора данных в гуманитарных науках».

При этом успешный обмен в одном проекте (например, создание цифрового архива) не гарантирует взаимопонимания в другом (например, при попытке применить машинное обучение для интерпретации образов). Часто высказываются опасения о риске «неравного обмена», когда интересы одной дисциплины (например, компьютерной) доминируют над исследовательскими вопросами другой (например, гуманитарной) [Dobson 2019]. Концепция П. Галисона

объясняет, как возможно практическое сотрудничество в таком разнородном поле, как ДН, показывая, что ДН функционирует не через унификацию, а через создание множества локальных, прагматичных альянсов, где важен не консенсус, а успешность взаимных обменов.

Программным высказыванием для этого периода становится статья Алана Лю «Значение Digital Humanities» [Liu 2013], в которой подводится итог периоду самоидентификации и ставятся задачи развития направления. В работе утверждается, что ДН находятся на этапе активного формирования идентичности, с дискуссиями о том, кто входит в поле, обязательно ли нужно уметь программировать, как сочетать техническую работу и глубокую интерпретацию. При этом цифровые гуманитарные исследования отличаются от близких областей (например, новых медиа) фокусом на текстах, архивах, исторических материалах и «старом» знании в широком смысле этого слова. Ключевая проблема развития обозначена как проблема значения (*meaning problem*), когда необходим переход от количественных данных к качественной интерпретации («от чисел к значению»), в качестве примера приводятся исследования Райна Хьюзера и Лонг Ле-Хака, когда машинные методы требуют подключения семантических ресурсов (исторического тезауруса) для генерации смысла. Важно учитывать, что ДН изначально построен на гибридной методологии, предполагающей диалог между машинными методами и гуманитарной интерпретацией, а значит необходимо совмещать *tabula-rasa*-подход (объективного машинного анализа) с человеческим семантическим знанием. Важной сложностью момента Лю называет теоретические рамки ДН и предлагает включить в цифровые гуманитарные исследования дизайн как метод генерации знаний, а не просто способ визуализации данных. А такой подход, как исследования науки и технологий (STS) может помочь понять, как технологические и человеческие практики взаимно формируют знания в ДН. Хотя традиционные гуманитарии склоны критиковать цифровых, нельзя не заметить, что ДН — это зеркало кризиса смыслов в гуманитарных науках, когда традиционные ценности (дух, субъективность, идентичность) сталкиваются с системными требованиями «цифрового знания» (данные, эффективность, влияние). Следствием этого оказываются вопросы преподавания гуманитарного знания, трудоустройства гуманитариев, а также институциональной поддержки в ДН как части общего вызова для гуманитарных наук в цифровую эпоху. И правильно использовать ДН не только как инструмент регистрации

кризиса, но и как силу для защиты и переосмысления гуманитарных ценностей. В частности, предлагая инициативу *4Humanities.org* как пример того, как ДН могут выступать в роли адвокатов гуманитарного знания. Таким образом, своим визионерским взглядом Лю смог перенести акцент с технических аспектов ДН на эпистемические и социальные вопросы, определить проблему значения как центральную для ДН и гуманитарного знания, предложить новые теоретические рамки (дизайн, STS) для осмысления ДН. И такой подход позволил позиционировать ДН как поле, которое не просто использует технологии, но и ставит под вопрос будущее гуманитарного знания в цифровую эпоху. А значит ДН должны не только анализировать, но и защищать гуманитарные ценности в современном мире.

В 2016 году увидел свет «Новый компаньон по цифровым гуманитарным наукам» под редакцией тех же троих благовестников ДН Сьюзан Шрайбман, Рэя Сименса и Джона Ансворта, которые уже в предисловии заявили, что хотя и «остается спорным, следует ли рассматривать цифровые гуманитарные науки в качестве самостоятельной области знаний, а не всего лишь набора взаимосвязанных методов, но без сомнения, в 2015 году, цифровые гуманитарные исследования являются динамичной и быстро развивающейся областью научной деятельности» [A New Companion 2016 P. xvii]. Редакторы «Нового компаньона» вспоминают, что сознательно решили отказаться от термина “humanities computing” и начали использовать название “digital humanities” с целью перенести ударение с компьютеринга на гуманитарные науки. «Может быть, через десять или двадцать лет определение “цифровой” будет казаться излишним применительно к гуманитарным наукам. Возможно, по мере того, как все большая доля нашего культурного наследия будет оцифрована или уже создана в цифре (born digital), станет ничем не примечательным тот факт, что цифровые методы используются для изучения человеческого творчества, а мы будем думать об исследованиях, описанных в этой книге, просто как о “гуманитарных”. Между тем, редакторы этого „Нового компаньона по цифровым гуманитарным наукам“ рады представить тщательно обновленный отчет о предметной области, как она существует сегодня» (P. xvii).

Новый компаньон включает 5 частей: инфраструктуры, создание, анализ, распространение и прошлое, настоящее, будущее цифровых гуманитарных наук. (Для сравнения в компаньоне 2004 года было 4 части: история, принципы, приложения и производство, рас-

пространение, архивирование.) В разделе «Инфраструктуры» обсуждаются такие вопросы, как интернет вещей, принципы коллективного использования данных и средств хранения оцифрованного культурного наследия. Раздел «Создание» посвящен особенностям междисциплинарных связей в цифровом контексте, новым медиа и вопросам моделирования, конструированию виртуальных миров и электронных библиотек. Раздел «Анализ» освещает моделирование данных, картографирование наблюдений, использование графических и мультимедийных форматов в цифровых исследованиях, анализ текстов и семантическую разметку. Раздел «Распространение» включает дискуссии о возможностях и ограничениях интерфейсов в цифровых проектах, о перспективах использования краудсорсинга и надеждах на разработку профессионального программного обеспечения для нужд цифровой гуманитаристики. Раздел о положении в современных гуманитарных науках описывает существующий научный ландшафт в этой области, изучает влияние глобализации и интернетификации, обращает внимание на характерные черты цифровых исследовательских практик, прогнозирует ближайшие перемены в цифровой науке.

«Новый компаньон» показывает, что цифровые гуманитарные исследования переходят от теоретического самоопределения к научной практике – академическим открытиям и новым интерпретациям, памятуя об опасности оказаться «во власти программного обеспечения». Как справедливо замечает К. Уорвик, не стоит буквально понимать поговорку «больше вкалывай – меньше болтай» (“more hack less yack”) в отношении таких создающихся областей, как цифровая гуманитаристика [A New Companion P. 538]. У. Томас поддерживает эту идею следующим размышлением: «Историки, литературные критики, философы, филологи, ученые, открывшие для себя цифровую гуманитарные науки, начинают перестройку научной деятельности и ее организационных форм для нового цифрового мира. Ученые стали открытыми для самых различных исследовательских методик, для обмена источниками и материалами (данными), и признали крупномасштабные распределенные модели научных проектов. Ученые пришли к важному признанию, что мы сейчас живем в эпоху огромной емкости, вездесущего хранения, связанной сетевой информации и беспрецедентного доступа. Вместо привычной манеры исследований, ориентированных на редкие материалы, к которым ограничен доступ, а эксперты самостоятельно проводят его отбор, цифровые гуманитарные науки в своих наиболее ярких проявлениях основываются на расширении сферы

гуманитарных исследований, открывая доступ к источникам, а также обогащая понятие научной деятельности» [A New Companion P. 524].

Период с 2013 по 2022 год стал для цифровых гуманитарных исследований стал временем ускорения роста, интенсивной интеграции в академическую систему и значительного усложнения методологического арсенала. Если предыдущее десятилетие было озабочено вопросами о своей самобытности и роли в университетской среде, то рассматриваемый этап знаменует собой переход к стадии становления, где ДН перестает быть нишевым направлением и начинает формировать собственные научные сообщества, инфраструктуру и линии внутренней критики.

Библиометрические исследования предоставляют убедительные данные о быстром росте ДН. Анализ публикаций в различных базах данных показывает медленный, но стабильный рост до 2017 года, после чего следует резкий скачок в 2019 году, что указывает на ускорение исследовательской активности именно в рассматриваемый период. Одним из ключевых шагов в этом процессе стали создание и популяризация ДН-журналов. Например, была проведена работа по составлению списка из 143 журналов, связанных с ДН, путем комбинации экспертной оценки и анализа цитирований [Spinaci et al. 2022]. Список был классифицирован по степени содержания ДН-контента: 34 журнала были отнесены к категории «исключительно» ДН, 18 — «значительно» ДН, 87 — «с признаками» ДН, а также были указаны четыре «мега-журнала», где ДН-публикации могут встречаться. Такие издания, как “Digital Scholarship in the Humanities” (до 2015 года назывался “Literary and Linguistic Computing”) и “Digital Humanities Quarterly”, стали центральными площадками для дискуссий и распространения исследований, формируя собственную экосистему научной коммуникации. Профессиональные журналы говорят о зрелости поля, способном поддерживать и развивать свои собственные каналы публикации, что является признаком институциональной зрелости.

В 2011 году появился ещё и журнал “The Journal of Digital Humanities”, который был объявлен экспериментом в области научной коммуникации, проводимый в рамках проекта электронного издательства PressForward «Центра истории и новых медиа имени Роя Розенцвейга» (RRCHNM). Журнал рассматривался как агрегатор быстро растущего поля ДН-исследований и проектов. Журнал, финансируемый за счёт первого гранта PressForward (2011–2014) Фонда Альфреда П. Слоуна, был тесно связан с изданием “Digital

Humanities Now”. Метрики публикаций в блоге и отклики на презентации на конференциях использовались для отбора контента, который затем дорабатывался, расширялся и в итоге публиковался в ежеквартальных выпусках. “JDH”, будучи экспериментальным изданием, выполнил свою миссию по предоставлению необходимой информации для запуска аналогичных публикаций в широком спектре исследовательских организаций, что было основным направлением деятельности PressForward на втором (2014–2015) и третьем (2015–2018) этапах. После завершения этой роли “JDH” был заморожен, а предыдущие выпуски сохранены в архиве и доступны для скачивания по сей день. Среди весьма влиятельных статей этого экспериментального журнала была статья Кристофа Шёха «Большие? Умные? Чистые? Грязные? Данные в гуманитарных науках» [Schöch 2013].

Внимание к данным — важная характеристика периода развития «зон обмена» в Digital humanities, потому что если можно спорить о пользе и успехах обмена методиками между разными научными подходами под зонтиком ДН, то обмен данными или даже в более материальной формулировке обмен базами данных — насыщенный и практически легко реализуемый способ цифрового обмена. Данные и базы данных в терминологии Галисона являются артефактами обмена. Как отмечал К. Шёх, гуманитарные данные — это цифровая, выборочно сконструированная, управляемая компьютером абстракция, представляющая некоторые аспекты определенного объекта гуманитарного исследования. «Независимо от того, являемся ли мы историками, использующими тексты или другие культурные артефакты как окна в другое время или другую культуру, или же мы литературоведы, использующие знания других времен и культур для построения смысла текстов, — утверждает К. Шёх, — цифровые данные добавляют еще один уровень посредничества в уравнение. Данные (а также инструменты, с помощью которых мы ими манипулируем) добавляют сложности в отношения между исследователями и их объектами изучения» [Schöch 2013 P. 3–4].

Методически данные получаются путем абстрагирования, т.е. отвлечения от несущественных (для исследователя и текущего исследования) характеристик объекта с целью определить его основные свойства и признаки, что позволяет сформулировать абстрактные понятия и модели, которые должны помочь понять суть изучаемого явления. При этом наблюдения собираются в определенную коллекцию (датасет), собранные сведения можно назвать каптой

(capta), или исследовательским «уловом» из изученных материалов, полнота и представительность которого требуют дополнительного доказательства и научного контекста [Володин 2019; Drucker 2011; Lavin 2021].

Постепенно происходил и методологический сдвиг в сторону использования все более сложных вычислительных техник: всё больше исследований было посвящено обработке естественного языка (natural language processing), тематическому моделированию (topic modeling), сетевому анализу (social network analysis), а также набирающему популярность машинному обучению (machine learning). Если ранний этап цифровых гуманитарных исследований был преимущественно сосредоточен на текстах, то период с 2013 по 2022 год характеризуется значительным расширением методологических рамок, включением новых типов данных и объектов исследования, выходящих за пределы текстов. Помимо анализа текстов, появился и начал активно развиваться визуальный анализ, включающий исследования потребления и производства изображений и пространственной информации для ответа на гуманитарные вопросы. Большое внимание таким сюжетам стала уделять «культурная аналитика». Такая реакция стала ответом на доминирование текстовых данных в раннем ДИ, возникла необходимость адаптировать вычислительные методы к работе с визуальным и пространственным знанием. Активно развивались методы компьютерного зрения для анализа произведений искусства, исторических фотографий, карт и других визуальных материалов, позволяющие выявлять скрытые паттерны, стилистические особенности и семантические связи, часто недоступные для непосредственного человеческого восприятия. Стали создаваться трехмерные модели архитектурных ансамблей, археологических находок или целых городских ландшафтов с помощью фотограмметрии, лазерного сканирования и компьютерной графики, позволяющие не только визуализировать, но и исследовать пространственные отношения, исторические изменения и функциональные характеристики культурных объектов. Стали широко применяться геоинформационные системы для исследования истории городов, культурных ландшафтов, распространения информации, миграции населения и других явлений, имеющих пространственные характеристики.

Важно отметить, что доминирующей формой публикаций на этом этапе остались рецензируемые статьи, что подчеркивает стремление ДИ к интеграции в традиционные академические стандарты и каноны. Хотя и происходит методологическое со-

зрвание направления и стремление к его признанию в рамках существующей системы оценки научных достижений. Однако публикационная активность была далеко не равномерной. Географический анализ показывает явное доминирование англоязычных стран. США являются лидером по количеству публикаций, за которыми следуют Австралия, Индия и Великобритания. Ранние сборники “Debates in the Digital Humanities” фокусируются преимущественно на опыте исследователей из США, Канады и Великобритании, хотя к концу десятилетия наблюдается некоторое увеличение видимости исследователей из Азии, Латинской Америки и Африки.

Концепция «зон обмена» показывает, что наука развивается не только через революции и смену парадигм, но и через горизонтальную интеграцию разных, часто несовместимых, культур. Цифровые гуманитарные исследования яркий пример такого развития. С одной стороны, такая концепция объясняет успехи междисциплинарности: от биоинформатики, нейроэкономики к культуромике, цифровым гуманитарным исследованиям, где объединяются специалисты с абсолютно разным научным видением мира. П. Галисон обращает внимание на материальные артефакты, инструменты и практики, которые являются реальной основой научной кооперации. И в этом свете яснее можно увидеть дискуссии цифровых гуманитариев о репозиториях, аналитических сервисах и устойчивости методических решений. В итоге концепция «зон обмена» Питера Галисона — это мощный инструмент для понимания того, как разнородность и непонимание в науке не являются препятствием, а будучи грамотно организованными через локальные языки и обмен артефактами, становятся источником инноваций и двигателем научного прогресса [Láng & Megyesi 2024].

Третье десятилетие цифровых гуманитарных исследований: эпистемическое сообщество

Эпистемическое сообщество — это сеть профессионалов с признанной экспертизой в определенной области и авторитетом в формулировании политики и стандартов. Профессионалов объединяют общие эпистемические убеждения, общие критерии валидации знания, а также общая политическая программа. Если мы переносим эту концепцию на ДН нельзя не заметить убежденность сообщества в том, что вычислительные методы могут привести к новым, неочевидным и масштабируемым выводам в гуманитарной сфере

(простым примером являются дискуссии вокруг «дальнего чтения» Франко Моретти). При этом сообщество провозглашает общими ценности открытости и коллаборативности, интероперабельности данных и повторяемости исследований. Процедуры валидации знания в ДН вполне устоялись. Статьи и проекты могут оцениваться как на академическую строгость, так и на техническую состоятельность. При этом критерии успеха расширились, так как не только цитируемость, но и использование сообществом созданного инструментария или датасета становится важным. Причем принятие сообществом стандартов кодирования (например, TEI) является актом выработки консенсуса о том, что считать «правильными» данными.

Члены эпистемического сообщества ДН становятся экспертами в грантодающих организациях, в том числе, по возможности, продвигая определенные типы проектов. Заметно и влияние на университетскую политику, включая лоббирование создания новых ДН-центров и департаментов, разработку критериев найма, учитывающих цифровой портфель.

Концепция эпистемического сообщества Питера Хааса может оказаться очень продуктивной для анализа и понимания области современных цифровых гуманитарных исследований. Фактически, ДН можно рассматривать как складывающееся эпистемическое сообщество в академической среде. Digital Humanities имеют ясные общие принципиальные убеждения. К ним можно отнести веру в потенциал цифровых технологий для трансформации гуманитарных исследований, преподавания и сохранения культурного наследия. Устойчивым можно назвать и убеждение в том, что количественные методы, визуализация данных и алгоритмический анализ могут раскрыть новые паттерны и задать новые вопросы в традиционно качественных дисциплинах. Цифровые гуманитарии высказываются в пользу открытости, сотрудничества и междисциплинарности, часто в противовес традиционной работе гуманитария-«одиночки». Принцип публичности гуманитарного знания отражает весьма распространенное стремление сделать результаты исследований доступными и полезными за пределами научных кругов.

Цифровые гуманитарные исследования опираются и на общие научные принципы, причем эти принципы можно считать общими несмотря на то, из каких дисциплин в цифру гуманитарии пришли. Цифровые гуманитарии уверены, что культуру и историю можно моделировать как данные, а эти данные можно структури-

ровать, анализировать и визуализировать. Цифровые гуманитарии разделяют точку зрения, что цифровые инструменты (от текстового и сетевого анализа до ГИС и 3D-моделирования) являются не просто вспомогательными, а эпистемическими, то есть меняют саму природу задаваемых вопросов и производимого знания. При этом, достигнут консенсус о важности стандартов и интероперабельности (например, использование TEI для разметки текстов, принципов Linked Open Data при разработке баз данных) как необходимых условий для прогресса направления [Цифровые 2023].

ДН-сообщество активно вырабатывает свои стандарты, внедряя практики рецензирования не только для статей, но и для цифровых проектов, инструментов и наборов данных. (Про открытое рецензирование заявок на ежегодные ДН-конгрессы с помощью системы ConfTool см. [Володин 2023а]). Важным шагом в развитии направления стало признание цифровых проектов в качестве полноценных академических результатов наравне с монографиями. Многими цифровыми гуманитариями была воспринята из компьютерных наук культура открытого кода и данных (публикуемых в репозиториях вроде GitHub или Zenodo), а также подробного документирования исследований, гарантирующего повторяемость результатов. Всё чаще звучат призывы к критическому рефлексивному подходу к самим инструментам и методам (в частности речь о цифровой герменевтике [Володин 2025]).

При этом, вполне заметной является и формирующаяся политическая программа ДН-сообщества. Если в начале пути аббревиатура ДН была красивым брендом, то сейчас она становится и определенной силой. Сообщество ДН активно лоббирует конкретные изменения в академии и за ее пределами. Например, ясно звучат настоятельные требования реформировать систему оценки научного труда, которая должна признать цифровые публикации, датасеты, цифровые гуманитарные проекты [Парад 2025]. В образовательной сфере ДН-сообщество наиболее успешно — внедрение курсов по программированию, работе с данными для студентов-гуманитариев уже распространено повсеместно, хотя часто и без достаточной рефлексивной компоненты [Будь 2024]. Цифровые гуманитарии в разных странах борются за устойчивое финансирование не только исследований, но и цифровой инфраструктуры (проекты, лаборатории, хранилища, платформы), одновременно отстаивая политику открытого доступа к научным и культурным материалам.

Существование такой метаорганизации, как ADHO (Alliance of Digital Humanities Organizations), со своими ежегодными ДН-конгрессами, ключевым международным журналом (Digital Scholarship in the Humanities) и цифровыми стандартами — институциональные признаки зрелого эпистемического сообщества.

ДН как сообщество сформировалось в ситуации неопределенности, когда традиционные гуманитарные науки столкнулись с «цифровым поворотом». Цифровые гуманитарии выступили в инициативной роли переводчиков, объясняющих коллегам-гуманитариям, что такое данные, и коллегам-инженерам — что такое культурный контекст [Drucker & Albrezzi 2026]. Внутренние конфликты оказываются уверенными признаками жизни. Дискуссии между «строителями» (tool-builders) и «интерпретаторами» (theorists) или споры о «дальнем» (distant) и «медленном» (close) чтении — это не слабость, а признак творческого подхода сообщества к выработке общих убеждений. В последнее время в некоторых регионах заметно влияние ДН-сообщества на академическую политику, когда сообщество напрямую влияет на политику финансирующих организаций, продвигая гранты именно на цифровые инфраструктурные проекты для гуманитарных исследований.

Тем не менее, нельзя не сделать несколько оговорок. Во-первых, несмотря на ясные общие принципы цифровых гуманитариев, нельзя не заметить тематическую гетерогенность сообщества: филолог, работающий с текстовыми корпусами, и археолог, создающий 3D-реконструкцию, наблюдают разный материал и могут говорить на совсем разных языках. ДН отличается высокой степенью динамичности: исследовательское поле меняется слишком быстро, и общие принципы могут устаревать с появлением новых технологий (особенно, конечно, такие вопросы обострились с очередным витком развития технологий искусственного интеллекта [Орехов 2023]). Как и в любой экспертной группе, в ДН есть свои центры власти, доминирующие институции и «звезды», что может противоречить идеализированному образу сообщества равных, объединенных только поиском истины. Причем звезды часто склонны больше других сомневаются в единстве сообщества или правильности выбранного пути [см. Moretti 2022].

Концепция эпистемического сообщества помогает объяснить, как междисциплинарная область смогла консолидироваться, выработать общие нормы, повлиять на академическую политику и создать узнаваемую идентичность. Изучение ДН через призму сообщества позволяет увидеть не просто набор проектов, а соци-

альный механизм производства нового типа гуманитарного знания. Цифровые гуманитарии — интересный пример эпистемического сообщества, которое сформировалось в ответ на вызовы цифровой эпохи и активно трансформирует ландшафт гуманитарного знания. Главная эпистемическая роль цифровых гуманитарных наук — быть мостом между двумя культурами: точными компьютерными науками и герменевтическими гуманитарными науками, предлагая гибридный методологический аппарат. Сообщество цифровых гуманитариев становится лабораторией новых форм научной коммуникации, коллаборации и производства знания, опыт которой важен для науки в целом. Перспективы развития цифровых гуманитарных исследований связаны с преодолением внутренних методологических разрывов, ответом на этическую критику и углубленной теоретической рефлексией собственных практик в ситуации развития агентного искусственного интеллекта.

Таким образом, от первых робких шагов самоопределяющихся «академических племён» через плодотворные, хотя и непростые «зоны обмена» цифровые гуманитарные исследования пришли к состоянию зрелого эпистемического сообщества — сообщества, которое не только отражает цифровой поворот в гуманитарном знании, но и активно формирует его повестку, стандарты и будущее. Этот путь — от фрагментированного движения к консолидированной экспертной силе — демонстрирует, как диалог между гуманитарной рефлексией и вычислительной мощью способен порождать не только новые методы, но и новые формы знания, новые способы его производства и социального бытования. Сегодня цифровые гуманитарные исследования — это уже не просто инструментарий или междисциплинарная гибридизация, а эпистемический императив, живая лаборатория мысли, где создаются мосты между смыслом и алгоритмом, традицией и инновацией, человеком и машиной. Как писал, комментируя послание Павла к Ефессянам, Фома Аквинский, вдохновивший Роберто Бузу на первый цифровой гуманитарный проект: “Non progredi est regredi”.

Примечания

- ¹ Хочу выразить благодарность *О. В. Алиевой* и *Б. В. Орехову*, организаторам круглого стола «Digitalia Humaniora: память о будущем» (28.11.2024 г.) на конференции «Гуманитарные науки в XXI веке: между текстом и цифрой», побудивших к этому рассуждению. Отдельная признательность коллегам — *А. А. Бонч-Осмоловской*,

Л. И. Бородину, Д. А. Гагариной, И. А. Кижнер, П. В. Колозариди, Р. Б. Кончакову, М. А. Лаптевой, М. С. Мироненко, Е. М. Севериной — за возможность бесед, кратких и долгих, очных и онлайн, с обсуждением отдельных аспектов развития эпистемического сообщества цифровых гуманитариев, рассмотренных в этой статье. Всё ценное в статье — результат общения с коллегами, все ошибки — мои собственные.

² Digital Humanities, ДН, цифровые гуманитарные науки, цифровые гуманитарные исследования используются в статье как синонимы. На русском языке в 2010-е годы для перевода Digital Humanities использовались преимущественно названия «цифровые гуманитарные науки» и «цифровая гуманитаристика», которые в 2020-е годы постепенно вытесняются определением «цифровые гуманитарные исследования», особенно после выхода одноименной коллективной монографии [Цифровые 2023].

³ Про пионерский этап в этой статье ничего не будет сказано, с одной стороны, потому что уже достаточно на эту тему написано, с другой стороны, авторская версия пионерского этапа изложена в первой главе коллективной монографии «Парад цифровых гуманитарных проектов» [Парад 2025 С. 6–33].

Литература

Исследования

Будь 2024 — Будь в курсе цифровых гуманитарных исследований: коллективная монография / Отв. ред. А. Ю. Володин. Красноярск: СФУ, 2024. 204 с. URL: [https://bik.sfu-kras.ru/elib/view?id=BOOK1"=\T2A\CYRB\T2A\CYRB\T2A\CYRK71%2F\T2A\CYRB+903--289731](https://bik.sfu-kras.ru/elib/view?id=BOOK1)

Бурдьё 2018 — Бурдьё П. Homo academicus. М.: Изд-во Ин-та Гайдара, 2018. 464 с.

Володин 2014 — Володин А. Ю. Digital humanities (цифровые гуманитарные науки): в поисках самоопределения // Вестник Пермского университета. Серия История. 2014. Т. 26, № 3. С. 5–12.

Володин 2019 — Володин А. Ю. Между data и capta: проблемы датафикации исторических исследований // Вестник Перм. ун-та. История. 2019. № 3 (46). С. 137–145.

Володин 2023 — Володин А. “Digital Humanities-2023” в Граце живьём: идеи, методы и тыквенное масло // Историческая информатика. 2023. № 4. С. 167–175. DOI: 10.7256/2585-7797.2023.4.69431

Володин 2023a — Володин А. “Digital Humanities-2023” в Граце живьём: идеи, методы и тыквенное масло // Историческая информатика. 2023. № 4. С. 167–175. DOI: 10.7256/2585-7797.2023.4.69431

- Володин 2024* — Володин А. “Digital Humanities-2024” в Вашингтоне: переосмысление, ответственность и гибрид как lifestyle // Историческая информатика. 2024. № 3. С. 130–143. DOI: 10.7256/2585–7797.2024.3.71479
- Володин 2025* — Володин А. Ю. Цифровая герменевтика исторического источника: формализация как толкование // Вестник Пермского университета. История. 2025. № 2 (69). С. 87–100. DOI: 10.17072/2219–3111–2025–2-87–100
- Володин 2025a* — Володин А. “Digital Humanities-2025” в Лиссабоне: доступность, гражданственность и *pavo cristatus* // Историческая информатика. 2025. № 3. С. 241–256. DOI: 10.7256/2585–7797.2025.4.75507
- Галисон 2004* — Галисон П. Зона обмена: координация убеждений и действий // Вопросы истории естествознания и техники. М., 2004. № 1. С. 64–91.
- Колозариди & Беляк 2024* — Колозариди П., Беляк Г. Н. Цифровая гуманитаристика как стадия научного знания: четыре метафоры // Логос (После алгоритмов). 2024. Т. 34. № 6(163). С. 179–202.
- Манифест 2011* — Манифест Digital Humanities (26.03.2011). URL: <https://tcp.hypotheses.org/501>
- Моретти 2016* — Моретти Ф. Дальнее чтение / пер. с англ. А. Вдовина, О. Собчука, А. Шели. Науч. ред. пер. И. Кушнарева. М.: Изд-во Ин-та Гайдара, 2016. 352 с.
- Орехов & Володин 2024* — Орехов Б. Володин А. Digital Humanities в России и конец истории // Цифровые гуманитарные исследования. 2024. № 1. С. 63–85. 10.31860/cgi-2024-1-63–85
- Орехов 2023* — Орехов Б. В. Текст и знание в аспекте больших языковых моделей // Историческая информатика. 2023. № 4. С. 104–113. DOI: 10.7256/2585–7797.2023.4.44180
- Парад 2025* — Парад цифровых гуманитарных проектов: коллективная монография / Отв. ред. А. Ю. Володин. Красноярск: СФУ, 2025. 286 с. URL: [https://bik.sfu-kras.ru/elib/view?id=BOOK1"=\T2A\CYRB\T2A\CYRB\T2A\CYRK71%2F\T2A\CYRP+180--445824](https://bik.sfu-kras.ru/elib/view?id=BOOK1)
- Пильщиков 2022* — Пильщиков И. А. Семь бесед о филологии и Digital Humanities: Интервью и дискуссии (2015–2021) / общая редакция и составление В. С. Полиловой. М.: Изд-во Моск.ун-та, 2022. 190 с.
- Плотинский 2001* — Плотинский Ю. М. Модели социальных процессов. М.: Логос, 2001. С. 123–138.
- Цифровые 2023* — Цифровые гуманитарные исследования: коллективная монография. Красноярск: СФУ, 2023. 272 с. URL: [https://bik.sfu-kras.ru/elib/view?id=BOOK1"=\T2A\CYRB\T2A\CYRB\T2A\CYRK71%2F\T2A\CYRC+752--494468](https://bik.sfu-kras.ru/elib/view?id=BOOK1)

- A Companion 2004* — Schreibman S., Siemens R., & Unsworth J. (Eds.). A Companion to Digital Humanities. Blackwell Publishing, 2004. 612 p.
- A New Companion 2016* — Schreibman S., Siemens R., & Unsworth J. (Eds.) A New Companion to Digital Humanities. Wiley Blackwell. 2016. 568 p.
- Becher & Trowler 2001* — Becher T., Trowler P.R. Academic Tribes and Territories: Intellectual Enquiry and the Cultures of Discipline. Open University Press, 2001. 238 p.
- Berry 2013* — Berry, D.M. (ed.). Understanding digital humanities. Palgrave Macmillan, 2013. 318 p.
- Burdick et al. 2012* — Burdick A., Drucker J., Lunenfeld P., Presner T., & Schnapp J. Digital_humanities. MIT Press, 2012. 152 p.
- Dobson 2019* — Dobson J.E. Critical Digital Humanities: The Search for a Methodology, University of Illinois Press, 2019. 200 p.
- Drucker & Albrezzi 2026* — Drucker Johanna, Albrezzi Francesca. The Digital Humanities Coursebook. Applied Concepts and Critical Approaches. Routledge, 2026. 154 p.
- Drucker 2011* — Drucker J. Humanities Approaches to Graphical Display // DHQ: Digital Humanities Quarterly. 2011. Vol. 5. № 1. URL: <https://dhq.digitalhumanities.org/vol/5/1/000091/000091.html>
- Drucker 2012* — Drucker J. Humanistic Theory and Digital Scholarship // Debates in the Digital Humanities / Matthew K. Gold (ed.). Minneapolis, MN, 2012. DOI: 10.5749/minnesota/9780816677948.003.0011
- Gadd 2009* — Gadd I. The Use and Misuse of Early English Books Online // Literature Compass. 2009. № 3 (6). P. 680–692.
- Galison 1997* — Galison P. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, 1997. 982 p.
- Gold & Klein 2016* — Gold, M. K., & Klein, L. F. (eds.). Debates in the Digital Humanities 2016. University of Minnesota Press, 2016. 632 p.
- Gold & Klein 2019* — Gold, M. K., & Klein, L. F. (eds.). Debates in the Digital Humanities 2016. University of Minnesota Press, 2019. 560 p.
- Gold & Klein 2023* — Gold, M. K., & Klein, L. F. (eds.). (2023). Debates in the digital humanities 2023. University of Minnesota Press. 520 p.
- Gold 2012* — Gold, M. K. (ed.). Debates in the Digital Humanities. University of Minnesota Press, 2012. 504 p.
- Gritsenko 2021* — Gritsenko, D., Wijermars, M., & Kopotev, M. (eds.). The Palgrave handbook of digital Russia studies. Palgrave Macmillan. 2021. 640 p.
- Haas 2015* — Haas P. Epistemic Communities, Constructivism, and International Environmental Politics. Routledge, 2015. 420 p.

- Hobsbawm & Ranger 1992* — Hobsbawm E., Ranger T. The Invention of Tradition. Cambridge University Press, 1992. 324 p.
- Kemman 2021* — Kemman M. Trading Zones of Digital History (Studies in Digital History and Hermeneutics). Walter de Gruyter, 2021. 188 p.
- Kizhner et al. 2022* — Kizhner I., Terras M., Orekhov B., Manovich L., Kim I., Rumyantsev M., Bonch-Osmolovskaya A. The History and Context of the Digital Humanities in Russia // Global Debates in the Digital Humanities / ed. Domenico Fiormonte, Paola Ricaurte, Sukanta Chaudhuri. University of Minnesota Press, 2022. P. 55–70.
- Láng & Megyesi 2024* — Láng B., Megyesi B. An STS analysis of a digital humanities collaboration: trading zones, boundary objects, and interactional expertise in the DECRYPT project // Palgrave Communications. 2024. Vol. 11(1) December. Pp. 1–17.
- Lavin 2021* — Lavin M. Why Digital Humanists Should Emphasize Situated Data over Capta // DHQ: Digital Humanities Quarterly. 2021. Vol. 15, no. 2. URL: <https://www.digitalhumanities.org/dhq/vol/15/2/000556/000556.html>
- Liu 2012* — Liu A. Where Is Cultural Criticism in the Digital Humanities? // Gold, M. K. (ed.). Debates in the digital humanities. University of Minnesota Press, 2012. P. 490–509.
- Liu 2013* — Liu A. The Meaning of the Digital Humanities // PMLA (Publications of the Modern Language Association of America). Vol. 128, No. 2 (March 2013), pp. 409–423.
- Mey 1992* — Mey M. de. The cognitive paradigm. University of Chicago Press, 1992. 346 p.
- Moretti 2007* — Moretti F. Graphs, Maps, Trees: Abstract Models for Literary History. New-York, London: Verso Books, 2007. 128 p.
- Moretti 2013* — Moretti F. Distant Reading. New-York, London: Verso Books, 2013. 256 p.
- Moretti 2022* — Moretti F. Falso Movimento. La svolta quantitativa nello studio della letteratura. Milano: Nottetempo, 2022. 160 p.
- O'Sullivan 2023* — O'Sullivan, J. (ed.). The Bloomsbury handbook to the digital humanities. Bloomsbury Academic, 2023. 512 p.
- Ramsay 2011* — Ramsay S. Reading Machines: Toward an Algorithmic Criticism (Topics in the Digital Humanities). University of Illinois Press, 2011. 112 p.
- Schöch 2013* — Schöch C. Big? Smart? Clean? Messy? Data in the Humanities // Journal of Digital Humanities. 2013. Vol. 2 №. 3 (Summer).
- Skorinkin 2023* — Skorinkin D. Digital Humanities in Russia Was Forever, Until It Was No More: The Story of Russian Digital Humanities in 2011–2022 // Canadian-American Slavic Studies 57, 1–2 (2023): 209–221.

Spinaci et al. 2022 — Spinaci G., Colavizza G., Peroni S. A map of Digital Humanities research across bibliographic data sources // *Digital Scholarship in the Humanities*. 2022. № 4 (37). С. 1254–1268.

Trowler et al. 2012 — Paul Trowler, Murray Saunders, Veronica Bamber (eds.) *Tribes and Territories in the 21st Century Rethinking the significance of disciplines in higher education*. Routledge, 2012 312 p.

ХРОНИКА

Мария Кешишян

КОНФЕРЕНЦИЯ «АКТУАЛЬНЫЕ ОШИБКИ ГУМАНИТАРНЫХ НАУК»

Хроника конференции «Актуальные ошибки гуманитарных наук» (27–29 марта 2025, ИТМО).

27–29 марта 2025 г. в Университете ИТМО состоялась конференция «Актуальные ошибки гуманитарных наук», организованная ДН-центром ИТМО. Мероприятие продолжило линию, заложенную в 2024 году, и закрепило традицию регулярного обсуждения проблем цифровых и междисциплинарных наук.

Предшествующая конференция «Гуманитарные проблемы актуальных наук: цифровые дисциплины и проекты» (2024 г.) была посвящена противоречиям между современными междисциплинарными подходами и традиционными гуманитарными практиками в условиях цифровизации исследований.

В нынешнем году фокус сместился: предметом обсуждения стала ошибка как особый объект научной рефлексии. Цель конференции состояла в том, чтобы вывести разговор об ошибках из рамок частных случаев на полях и сделать его предметом систематического академического обсуждения, охватывающего разные уровни и формы заблуждений в науке.

Вопрос о научных ошибках приобретает особую остроту в контексте междисциплинарных и цифровых наук. Автоматизация исследовательских методов и активное использование цифрового инструментария формируют новые режимы производства знания. Эти процессы не только несут риск частных сбоев, галлюцинаций, но и могут содержать систематические искажения в своих методологических научных основаниях.

Мария Ованесовна Кешишян
Университет ИТМО
m.keshishian@itmo.ru

Для структурирования столь разнородной проблематики организаторы предложили художественный приём — интерпретацию системы «идолов» Фрэнсиса Бэкона, превратив её в рабочий инструмент для картографирования ошибок и искажений:

- идолы рода — ограничения, связанные с происхождением знания, его дисциплинарными рамками и каноническими формами;
- идолы пещеры — заблуждения «здравого смысла», основанные на доверии к обыденному знанию;
- идолы рынка — ошибочные смещения понятий, методов и подходов;
- идолы театра — искажения, возникающие в моделях и процедурах представления знания

Участникам конференции предлагалось обсудить широкий спектр исследовательских проблем: от дисциплинарных ошибок и скрытых изъянов теоретических моделей до ошибок внедрения и интерпретации. В центре внимания оказались также различия между человеческими и машинными ошибками, вопрос о «праве на ошибку» и возможности её признания, а также размышления о том, всегда ли ошибки ведут к отрицательным результатам или же иногда способны открыть новые исследовательские перспективы.

Конференция длилась три дня, каждый из которых заканчивался рефлексией и краткими сводками от модераторов секций.

Вступительное слово прозвучало от программного директора конференции Полины Колозариди. Полина назвала тему ошибок объединяющей для разных гуманитарных, социальных, и технических дисциплин. Ведь источники ошибок — не имеют ясного происхождения, и зачастую неясность с природой ошибки и создаёт основную сложность. Секции были собраны не по дисциплинам, напротив, чаще всего одна секция сочетала в себе представителей разных традиций. В названиях, темах и докладах секций отражено, в каких ситуациях может быть расположена или обнаружена ошибка.

Секционную часть конференции открыла секция «**теория <—> поле**», на которой обсуждались ошибки в теориях и предметных областях: от критики теории Ньютона Гёте (Глеб Куприянов, НИУ ВШЭ) и истории советских компьютерных сетей (Илья Бузлуков, ИТМО) до «альтернативно-исторической науки» (Даниил Коськов, ЕУ СПб), культуры университетов (Татьяна Акунеева, НИУ

ВШЭ) и производства истины (Андрей Валов, независимый исследователь) с перспективы исследований науки и технологий. Глеб Куприянов, модератор секции, на обсуждении итогов первого дня конференции заметил, что общей темой секции стал вопрос о том, где расположена ошибка. При этом в разных докладах способ выявления ошибок был разным. В докладе Татьяны способом обнаружения ошибки стало сомнение в применимости теории к конкретному полю, Илья предложил последовать теории, оттолкнуться от самого поля и сделать вывод о наличии там противоречий. Сам Глеб рассматривал, как ошибки теории и метода могут влиять на ошибки поля, и как их можно выявить. В докладе Андрея ошибка и сомнение — симптом состояния науки, и отношение к ошибке позволяет различать разные типы науки и объяснить кризисы, в которых эта наука оказывается.

В секции «метод \longleftrightarrow теория» обсуждались ошибки, возникающие на стыке подходов: пересмотр исследования атомизации общества (Юлия Красюкова, МВШСЭН), пределы эмпиризма (Илья Ляшко, СПбГУ), методологические ловушки в гуманитарных науках (Артём Гозбенко, НИУ ВШЭ) и эффекты управленческих команд в университетах (Владимир Крестинин, НИУ ВШЭ).

По словам Артёма Гозбенко: «На нашей секции мы пытались рассказать о своих исследованиях исходя из ошибки, показать, в чём мы были не правы, и это позволило нам понять, связана ли наша теория с методом, какой метод мы используем и какой метод использовать лучше».

Секция «гуманитарные \longleftrightarrow данные» была посвящена проблемам ошибок в работе с данными: от метафор «эрозии и гниения» (Дарья Радченко, РАНХиГС) и семантических искажений (Андрей Володин, МГУ) до неточностей геовизуализации (Марина Шилова, ТГУ), картографических проекций (Евгений Гришин, Константин Кунавин, «История в пространстве») и ошибок фотографии (Екатерина Юшкевич, Мария Гусева, РОСФОТО).

Модератор секции Андрей Володин подытожил обсуждение: «Мы обсуждали сегодня гуманитарные данные — вопрос и простой, и сложный одновременно. Всем хочется утверждать, что гуманитарные данные не такие, как обычные данные, что они обладают особой спецификой. И нам, кажется, удалось это показать на примере конкретных ошибок, которые возникают при работе с ними».

Полина Колозариди заострила вопрос: «Являются ли гуманитарные данные чем-то специфическим сами по себе, или речь идёт

скорее о специфике гуманитарного субъекта, который их собирает и интерпретирует? В докладах секции акцент смещался именно на исследователя и его практики работы с данными».

Андрей Володин уточнил свою позицию: специфика гуманитарных данных определяется прежде всего целью их создания и использования. «Вопрос не в том, чтобы просто собрать данные, а в том, как их построить так, чтобы увидеть в них не только количественные параметры, но и сложное внутреннее содержание, — пояснил он. — Данные — это всегда абстракция. Ключевой вопрос: что именно из этой абстракции нам необходимо извлечь для наших исследовательских наблюдений?»

Отвечая на собственный вопрос о существовании «действительно гуманитарных данных», Володин предложил следующий критерий: «Гуманитарные данные экстремально нюансированы. Эта нюансированность ломает привычное представление о традиционной базе данных, где в каждом поле должна быть однозначная, недвусмысленная запись. Гуманитарные данные требуют сложной, многослойной организации. Со стороны может показаться избыточным: "а кому это надо?" — но именно в этой сложности и заключается возможность адекватного представления гуманитарного знания».

Обсуждение специфики гуманитарных наук в исследованиях продолжились на секции **«метод ↔ алгоритм»**. Обсуждались ошибки, возникающие при алгоритмизации методов: от разметки эмоциональной лексики (Анастасия Кочкурова, НИУ ВШЭ) и OCR (Лев Шадрин, LAGOOS) до автоматической аннотации (Милана Ходжаметова, НИУ ВШЭ), распознавания эмоций (Яна Сосновская, ЕУ СПб) и выявления ошибок качественного анализа стиля Толстого машинным анализом (Борис Орехов, НИУ ВШЭ).

По словам Яны Сосновской, центральной темой секции стали трудности адаптации алгоритмов к специфике гуманитарного материала: «Гуманитарные исследования полны нюансов, которые сложно формализовать. Когда готовые инструменты не дают нужного результата, исследователи начинают разрабатывать собственные решения — и часто это приводит к ещё большим проблемам. Возникает парадокс: стремление учесть специфику материала обрачивается усложнением инструментария и накоплением новых ошибок».

Модератор секции Борис Орехов дал критическую оценку дискуссии: «Мы не смогли выйти на обсуждение ключевого вопроса —

как именно алгоритмизация влияет на характер ошибок и порождает ли она новые типы искажений».

Постерная сессия, по словам модератора Марии Могилевич, объединила интересные вопросы, которые не попали в тематический план конференции. При этом формат постерной сессии открыл возможность задавать вопросы напрямую: где ошибка? Кто её совершил? Что с этим делать?

На подведении итогов дня Борис Орехов поделился наблюдением: «Гуманитарий — царь Мидас: всё, к чему он прикасается, становится гуманитарным». Он высказал предположение, почему гуманитарные данные такие нюансированные: «У гуманитария есть страх редукции. В отличие от других наук, у гуманитариев не сложилось практик на этот счёт. Гуманитарии не привыкли к идее, что что-то нужно отбросить, потому что не знают, что важно, а что не важно. Поэтому происходят усложнения этих данных, что иногда хорошо, а иногда плохо, но это системный процесс».

Доклад **«ошибки как они есть»** задал тон последнему дню, обсуждая культуру сообщений об ошибках и то, как интерфейсы формируют их восприятие (Александр Королев, ЕУ СПб; Полина Колозариди, ИТМО).

В докладе была представлена цифровая археология сообщений об ошибках: анализ исторической эволюции способов, которыми машины коммуницируют с пользователями о сбоях. Исследование показало, что за формулировками ошибок стоят различные концептуальные модели понимания сбоев, их классификации и предполагаемых способов решения. Культура сообщений об ошибках оказывается не нейтральной технической практикой, а пространством, где воплощаются представления о природе ошибки и о взаимодействии человека и машины.

Участники обсуждения подчеркнули ответственность исследователей-гуманитариев в формировании языка описания систем: выбор метафор, способов классификации и стратегий коммуникации об ошибках — это не только техническая, но и культурная работа, которая определяет, как пользователи понимают и переживают сбои технологий.

В секции **«исследование <—> данные»** речь шла о сложностях поля: вовлечённости этнографа (Аркадий Моргун, КубГУ), необходимости отказа от избыточных данных (Владислава Боброва, ЕУ СПб) и преждевременности кейс-стади (Лидия Рыжова, ЕУ СПб). В ходе обсуждения выявились две ключевые проблемы. Первая касается исследовательской идентичности: выбор и обос-

нование методов работы с данными оказывается неотделим от того, как исследователь определяет свою дисциплинарную принадлежность и роль в поле. Вторая проблема связана со сложностью самого эмпирического материала: поле перенасыщено потенциальными объектами исследования и источниками ошибок, что требует от исследователя стратегического выбора — фокусироваться на тех проблемах, которые поддаются решению имеющимися средствами.

Секция «**интерфейс <—> инструмент**» обратилась к ошибкам интерфейсов: критика направления их развития (Ирина Антонова, ИТМО), роль критического голоса (Алексей Евстифеев, «Собака Павлова») и исследование библиотечных интерфейсов (Татьяна Полежаева, ТГУ).

Модератор секции Ирина Антонова сформулировала объединяющий вопрос дискуссии: «Является ли ошибкой то, что мы называем ошибкой?». Парадокс секции заключался в том, что докладчики критиковали интерфейсные решения, которые при этом функционируют и используются. Это поставило под сомнение сам статус «ошибки» в контексте интерфейсов: возможно, речь идёт не о технических сбоях, а о противоречии между разными представлениями о том, каким должен быть интерфейс. Второй проблемой стал вопрос об исследовательской позиции: кто выступает субъектом критики интерфейсов и какое пространство для действия открывает такая критическая рефлексия?

В секции «**вывод <—> внедрение**» обсуждались ошибки на пути из науки в практику: языковая политика (Ксения Викторова, ЕУ СПб), ошибки машинного анализа в медицине (Елена Введенская, ИНИОН РАН), когнитивные искажения в законах Азимова (Илья Суров, ИТМО) и достоверность источников в переводе (Виталий Щербина, СПбУТУиЭ).

Модератор секции Азиз Аширов (ИТМО) отметил, что дискуссия оказалась особенно продуктивной с точки зрения проблематизации: обсуждение высветило фундаментальный вопрос о том, на каких основаниях научные результаты могут и должны переноситься в практическую плоскость.

Завершила программу секция «**кино и глитч**», где ошибки и сбои рассматривались как художественный метод: от глитч-картографии (Алина Латыпова, ИТМО/СПбГУ/ЛИКИ) до инструментов glitch-арта (Фёдор Ерофеев, НИТУ МИСИС). Эта часть программы вместе с последующим показом экспериментального кинотеатра «Курок» перевела проблематику ошибок в плоскость художественного осмысления. Докладчики продемонстрировали стратегии не устра-

нения ошибок, а сосуществования с ними — превращения сбоя в выразительное средство и источник новой оптики восприятия.

Полина Колозариди в заключительном слове выделила несколько уровней осмысления ошибок, которые проявились в ходе конференции.

На методологическом уровне стало очевидно, что исследователи по-разному присваивают и локализуют ошибки. Одни обнаруживают их в своём теоретическом поле, другие — на стыке метода и материала, третьи — в процессе внедрения результатов.

Важным практическим выводом стало понимание того, что исследователь всегда берёт на себя риск ошибки: публикуя статью, выходя в поле, внедряя новое правило или метод. Этот риск неустрашим и составляет неотъемлемую часть исследовательской практики. Метод при этом становится не столько гарантией от ошибок, сколько инструментом их выявления и основанием для критики — в том числе для постановки вопросов к другим дисциплинарным полям.

«Совершать ошибки — это наша работа, и не стоит этого бояться», — резюмировала Колозариди.

Конференция показала необходимость перехода от замалчивания ошибок к их систематическому изучению и осмыслению как источника развития науки. Обсуждённые ошибки должны ложиться в разработку правил, исследовательских практик, инструкций, методологических принципов.

РЕЦЕНЗИЯ

Софья Порфирьева

РЕЦЕНЗИЯ НА КНИГУ «ГЕРМЕНЕВТИКА: КОМПЬЮТЕРНАЯ ИНТЕРПРЕТАЦИЯ В ГУМАНИТАРНЫХ НАУКАХ» СТЕФАНА СИНКЛЕРА И ДЖЕФФРИ РОКВЕЛЛА

Рец. на кн.: *G. Rockwell and S. Sinclair. Hermeneutica: Computer-Assisted Interpretation in the Humanities. Cambridge: The MIT Press, 2016.*

Монография «Герменевтика: компьютерная интерпретация в гуманитарных науках» (*Hermeneutica: Computer-Assisted Interpretation in the Humanities*) канадских исследователей Стефана Синклера и Джеффи Роквелла была опубликована в 2016 году издательством Массачусетского технологического института. Одним из преимуществ этой работы является то, что книга написана не просто философами, далекими от программирования, или программистами, не разбирающимися в философии, но исследователями, которые стремятся объединить эти две области. Хотя их компетенции в программировании явно преобладают над гуманитарной составляющей, это не обедняет теоретического размышления. Перед нами — значимая попытка интеграции философского осмысления в развивающуюся сферу цифровых гуманитарных исследований.

Как очевидно из названия, основная задача книги — понять, какие возможности открывают нам цифровые методы интерпретации текста. Или, как изящно ставят вопрос сами авторы: являются ли цифровые технологии троянским конем, в котором сидит вирус количественных методов, или же позволяют иначе взглянуть на процесс интерпретации? [Sinclair & Rockwell 2016, P. 103–104].

Софья Игоревна Порфирьева
University of Ottawa
sporfir@uottawa.ca

Синклер и Роквелл не просто размышляют о том, как цифровые технологии внедряются в гуманитарные науки, но делают свою работу частью этого процесса. Монография «Герменевтика» существует как минимум в двух ипостасях: как аналоговая книга, привычная любому гуманитарною, и как вебсайт, близкий сердцу цифровым энтузиастам. Бумажная версия книги содержит одиннадцать глав, которые можно поделить на два тематических блока: теоретические размышления (главы 1, 2, 3, 5, 7, 9 и 11) и практическая работа, или — как ее называют сами авторы — интерлюдии (главы 4, 6, 8, 10). Практические главы опубликованы в открытом доступе на сайте книги, где можно обнаружить еще две дополнительные главы-инструкции: как подготовить текст для цифрового анализа и список полезных ресурсов. Но и это еще не всё: главным практическим инструментом для анализа текстов является веб-инструмент для чтения и анализа цифровых текстов *Voyant Tools*, авторами которого также являются Синклер и Роквелл.

Диджитализация гуманитарного знания позволяет отказаться от классического, или линейного, прочтения книги, создавая из текста своеобразный конструктор. Учитывая необычный формат данной монографии, я позволю себе в некотором смысле «пересобрать» ее: сначала я предлагаю проанализировать теоретические главы, а затем перейти к краткому обзору практических глав и оригинального инструмента *Voyant*. Действительно ли авторам удалось взглянуть на цифровой метод анализа текстов сквозь философскую призму, или же термин «герменевтика» в названии работы остается лишь вежливым кивком в сторону гуманитариев?

Качественным отличием «Герменевтики» от других работ в области цифровых гуманитарных исследований является проработанная и детализированная методология, речь о которой идет в первой главе «Введение: корректировка метода». В традиционном смысле понимая герменевтику как метатеорию, Синклер и Роквелл рассматривают цифровые методы как инструменты, расширяющие классические практики чтения и осмысления текстов:

«Несмотря на избыток интерпретаций, нам необходимо продолжать читать и интерпретировать наш мир, сверять свои интерпретации с чужими. Более того, мы должны задуматься об использовании компьютеров в процессе интерпретации [. . .], что открывает перед нами как новые возможности, так и потенциальные опасности» [Sinclair & Rockwell 2016, P. 18–19].

В разговоре о методологии авторы начинают с «Размышлений» Рене Декарта, который выступал в защиту уединенных размышлений. Перефразируя Картезия, Синклер и Роквелл пишут:

«Мы провели целый день, закрывшись в перегретой компьютерами лаборатории, имея полный досуг разговаривать, обдумывая инструменты и экспериментируя с текстами» [Sinclair & Rockwell 2016, P. 6].

Акцент на совместной — коллаборативной — деятельности действительно оказывается отличительной чертой цифровых методов, противопоставленных картезианской модели уединенной интеллектуальной рефлексии. В этом смысле авторы, скорее, подчеркивают постоянную необходимость диалога, поскольку цифровые исследования неизбежно междисциплинарные: они требуют высокого уровня экспертизы в различных областях, от программирования до лингвистики и литературного анализа. Даже если исследователь работает с цифровыми инструментами самостоятельно (что сегодня возможно благодаря zero-coding³), он неизбежно вступает в невидимый диалог с их создателями. Эти программы разработаны другими специалистами, а значит, сам процесс работы становится как минимум диалогичным. Такой подход к интерпретации текстов Синклер и Роквелл предлагают назвать «гибкой герменевтикой» (*agile hermeneutics*) [Sinclair & Rockwell 2016, P. 6] по аналогии с гибкой методологией разработки. Гибкость этого подхода заключается в том, что исследователи предлагают обращать внимание не только на то, *что* мы интерпретируем, но и на то, *посредством чего* мы это делаем. Другими словами, сами инструменты могут стать объектом интерпретации.

Однако авторы не остаются слепо оптимистичными. Отчасти в традиции хайдеггеровской философии техники они говорят о трех опасностях цифровых методов: во-первых, цифровые инструменты зачастую воспринимаются как нейтральные, будто бы существующие сами по себе; во-вторых, программирование грозит сделать гуманитарные науки служанкой технологий; в-третьих, любая технологическая эволюция несет в себе определенные риски, поскольку невозможно заранее сказать, насколько успешной будет изначальная задумка. И если первые две опасности действительно являются Сциллой и Харибдой цифровых гуманитарных исследований, то третья характерна для любой интеллектуальной работы. А потому, вторя Августину, Синклер и Роквелл предлагают отправиться в исследование этой гибкой герменевтики:

«Мы все... пытаемся плыть дальше, признавая вероятную невозможность когда-либо найти окончательные основания или цели» [Sinclair & Rockwell 2016, P. 21].

Следующие две главы «Слова, которые посчитали: как компьютеры анализируют текст» (2) и «От конкорданса к вездесущей аналитике» (3) прослеживают развитие компьютерного анализа в жанре очерка. Во второй главе авторы предлагают базовый технический экскурс в детали работы компьютера, подчеркивая, что последние могут гораздо больше, чем просто «перебирать цифры» [Sinclair & Rockwell 2016, P. 41]. Поскольку компьютер требует формализации, он помогает нам с большей строгостью относиться к тексту, который мы пытаемся проинтерпретировать: «[компьютеры] заставляют нас формализовать то, что мы знаем о тексте, и то, что мы хотим о нем узнать» [Sinclair & Rockwell 2016, P. 41]. В некотором смысле авторы пытаются показать, что компьютерный анализ текста — это не столько про сухие подсчеты, сколько про построение (гибких) моделей:

«Текстовый анализ позволяет вам создавать модели, проводить с ними различные манипуляции, разрушать их, а уже затем — рассуждать о них [. . .] Моделирование также является частью герменевтического круга» [Sinclair & Rockwell 2016, P. 41].

Еще одной отличительной чертой формализации является то, что она всегда открыта для критики в широком смысле слова. В гуманитарных дисциплинах большинство тезисов подтверждается цитатами из других источников, проверить или опровергнуть которые можно только через другие отсылки и цитаты. Код же — или любое другое формализованное описание метода — может быть проверен и протестирован на других примерах [Sinclair & Rockwell 2016, P. 42]. Синклер и Роквелл подчеркивают, что это не значит, что всякий аргумент или всякая интерпретация нуждается в формализации. Это лишь один из способов иначе взглянуть на текст и доверить компьютерам то, в чем они действительно хороши — в анализе длинных и сложных конструкций. В отличие от человека, ни один компьютер (по крайней мере пока!) не является воплощенным умом (*embodied mind*):

«То, что мы анализируем, может влиять на нас. Современные компьютеры не так *пластичны*, как мы» [Sinclair & Rockwell 2016, P. 43; курсив мой — С. П.].

Третья глава посвящена, с одной стороны, обзору того, как технический прогресс привел нас от конкордансов к облаку слов на новостных сайтах (или сайтах академических журналов), визу-

ализации текстов и инфографике. С другой стороны, авторы вновь обращают внимание на то, что именно цифровые инструменты делают с текстом — они позволяют реорганизовать текст. Хотя Синклер и Роквелл ссылаются на интерпретацию Уилларда Маккарти¹, вопрос о пересборке текста и разрушении изначальной линейности вполне можно рассматривать через призму ризоматичности и (де)территориализации, предложенных Делезом и Гваттари:

«Сборка и есть такое пересечение измерений в множестве, которое с необходимостью меняет природу в той мере, в какой наращивает свои соединения. В ризоме нет точек или позиций, какие мы находим в структуре — дереве или корне. Есть только линии» [Делёз и Гваттари 2010, Р. 14]

Хорошим примером такой пересборки может стать алгоритм LDA, способный привести мысль-дерево к мысли-ризоме². В этом случае тематическое моделирование делает из текста карту, которая состоит не из статичных точек, но из линий и направлений. Да и сами цифровые методы не остаются статичными. Они интерактивны в том смысле, в каком «объясняют себя, позволяя вам использовать их» [Sinclair & Rockwell 2016, Р. 66].

Чем же являются эти цифровые инструменты? Каковы их эпистемологический и онтологический статусы? Об этом Синклер и Роквелл размышляют в пятой главе «В моем тексте какая-то игрушка: проблемы риторики текстового анализа». Здесь авторы вновь сталкиваются две «стихии» — аналитику и герменевтику. С одной стороны, длинные рассуждения о методе отвлекают от интерпретации; с другой стороны, методология редко бывает в центре внимания гуманитарных исследований, «потому что подтверждение находится в интерпретации, а не в том, как мы пришли к этой интерпретации» [Sinclair & Rockwell 2016, Р. 96]³.

Цифровые методы, пишут авторы, подобны лестнице Витгенштейна — они отбрасываются, как только исследователь поднимается выше в поисках нового решения проблемы. Более того:

«. . . метод стал *невидимым* для гуманитарных наук, и в риторике последних есть нечто фундаментальное, что противостоит методу, но не размышлению о методе [. . .] Возможно, гуманитарные науки стыдятся методов, потому что они выглядят как *игрушки* (toys)» [Sinclair & Rockwell 2016, Р. 97; курсив мой, — С. П.].

Синклер и Роквелл утверждают, что методы (эти игрушки, или даже «безделушки») относятся не к дискурсу, но к *вещам* (things), а потому ими интересуются не гуманитарные дисциплины, а более практические — инженерия, дизайн, или компьютерные науки

[Sinclair & Rockwell 2016, P. 98]. Тезис, с которым на первый взгляд хочется не согласиться: во-первых, постгуманистический подход в философии не только обращает внимание на «вещи», но и в целом преодолевает бинарность субъекта и объекта⁴. Во-вторых, сами авторы продолжают свое рассуждение о «вещественности» методов с опорой на философские работы Рене Декарта, Дэвида Юма и Мартина Хайдеггера. С другой стороны, очевидно, что такая в определенной степени отчаянная демаркация связана с реальной проблемой невидимости методов в гуманитарных исследованиях. Кажется, что корень этой проблемы авторы «Герменевтики» видят в господстве *cogito* в западной философии как универсального и главного метода познания, которое, однако, «становится инструментом для познания одних вещей, неспособным познать другие» [Sinclair & Rockwell 2016, P. 99]. Хотя философский экскурс от Декарта к Юму и Хайдеггеру местами выглядит сбивчиво и не вполне убедительно [Sinclair & Rockwell 2016, P. 99–100], итоговый вывод авторов открывает любопытное и перспективное направление размышлений: цифровые методы являются средствами, или — используя хайдеггеровскую терминологию — *утварью* (*das Zeug*), которые должны быть одновременно подручными (*zuhandene*) и наличными (*vorhandene*). Другими словами, они должны быть *незаметными* (подручными), когда используются как инструмент для интерпретации текста; и *видимыми* (наличными), чтобы самим стать предметом анализа.

Как любая вещь, цифровые методы не находятся в изоляции от других инструментов, объектов и, главное, субъектов. Они все находятся в некоторой конфигурации, изменения которой меняют *статус* этих нечеловеческих и человеческих акторов: в одной сборке цифровые методы являются инструментами, а другой — предметами анализа. И, возможно, понимание того, *как и для чего* эти методы работают, поможет избежать превращения цифровых инструментов в троянского коня, пытающегося разрушить гуманитарную дисциплину количественными методами.

Тем не менее, компьютеры обладают эпистемологическим преимуществом — они действительно способны обрабатывать большие количества информации или *большие данные*. Как остроумно замечает Этьен Брюне в статье, посвященной анализу животных во французском литературном 19–20 веков:

«Достаточно знать, что эти кривые получаются в результате перекрестного умножения, квадратных корней и многих других ингредиентов,

одни названия которых могут испортить вам аппетит, хотя компьютер переварит их без проблем» [Brunet 1989, P. 303].

Очевидно, что просто переварить данные недостаточно: их необходимо проинтерпретировать. В седьмой главе «Ложно-положительные результаты: возможности и опасности анализа больших текстовых данных» Синклер и Роквелл задаются вопросом о том, что значит «заниматься интерпретацией в гуманитарных науках, области, в которой мы традиционно предполагаем, что чем больше информации, тем лучше» [Sinclair & Rockwell 2016, P. 114].

Хотя человеческое познание и отличается пластичностью (которой лишены компьютеры), эта особенность проявляется не только в наших положительных способностях — воображении и умении интерпретировать, — но и в биологической ограниченности. Здесь на помощь приходят цифровые методы. Авторы «Герменевтики» отсылают в первую очередь к Франко Моретти, чья методология дальнего чтения, «метафорически предполагает, что... нам необходимо сделать шаг назад, чтобы увидеть целое» [Sinclair & Rockwell 2016, P. 116]. Как отмечает сам Моретти в эссе «Конец начала: ответ Кристоферу Прендергасту»:

«... нам не нужно больше интерпретаций не потому, что им нечего сообщить, но потому что они *уже более или менее сообщили то, что могли* [...] настоящий вызов и надежда на истинный прорыв заключены в области причинного обусловливания и крупномасштабных объяснений» [Моретти 2016, P. 215–216].

Синклер и Роквелл, однако, с такой позицией не согласны: суть объяснения состоит не в том, чтобы обратиться к смыслу текста, но в том, чтобы выявить в нём определенные закономерности: «Это эпидемиология; она ищет симптомы, но не причины» (Ibid. p. 116). Интерпретация же, по их мнению, это добродетель, связанная с тем, что Аристотель называл началом философии — с удивлением. Кстати, сам Моретти признается, что не занимается программированием и обработкой данных, а работает уже с готовым материалом [Distant Reading 2016, P. 7].

С практической точки зрения большие данные — это «игровая площадка» для цифровых гуманитарных исследователей, которая позволяет им (i) фильтровать и группировать данные; (ii) улучшать данные, делая их более удобными и полезными в использовании; (iii) проследживать одинаковые паттерны в различных текстах⁵; (iv) проводить диахронический анализ изменения языка (такие проекты как TLF, TLG, Google Ngram Viewer); (v) классифицировать

и кластеризовать информацию; (vi) анализировать социальные связи; (vii) наконец, использовать данные для анализа собственной жизни (например, использовать фитнес-трекеры, умные часы или любые другие приложения, собирающие повседневную информацию о пользователе) [Sinclair & Rockwell 2016, P. 129–130]. Однако с большими данными приходит и большая ответственность, поэтому Синклер и Роквелл подчеркивают, что работа с ними в рамках цифровых исследований — это прежде всего ключ к пониманию того, как большое количество информации изменяет само знание. А понять это можно только посредством «мышления через» (*thinking through*) большие данные [Sinclair & Rockwell 2016, P. 125].

Девятая глава «Модельная теория: мышление через герменевтические вещи» и заключительная одиннадцатая глава «Гибкая герменевтика и разговор о гуманитарных дисциплинах» подытоживают размышления авторов: гибкая герменевтика подразумевает не только интерпретацию текста, но и создание методов, посредством которой эта интерпретация возможна [Sinclair & Rockwell 2016, P. 199]. Цифровые методы, в свою очередь, выступают не только как вспомогательные инструменты для размышления о текстах, но и сами становятся предметом философского осмысления. Можно сказать, что такой масштаб задач выглядит весьма амбициозным для одного цифрового гуманитарного исследования. Тем не менее, Синклер и Роквелл подчеркивают, что их задача — «излечить метод (в целом) от его солипсизма» [Sinclair & Rockwell 2016, P. 200], вновь приглашая исследователей к реальному диалогу и сотрудничеству в цифровых лабораториях.

Что может стать результатом такой работы? Авторы предлагают взглянуть на четыре интерлюдии, вплетенные в теоретическое полотно работы. Однако прежде чем перейти к их краткому обзору, стоит сказать несколько слов непосредственно о цифровом инструменте Voyant, разработанном авторами книги. Этот инструмент доступен в онлайн и оффлайн форматах и предназначен для создания корпуса, анализа и визуализации цифровых текстов. Voyant может проанализировать частоту и распределение терминов в тексте, а затем представить это в виде облака слов или «пузырей». Коллокации и корреляции Voyant может показать в виде графика или таблицы. В онлайн версии книги все четыре интерлюдии сопровождаются визуализациями, построенными на платформе Voyant.

Первая интерлюдия «Воробей стремительно летит: анализ рассылки *Humanist*»⁶ прослеживает эволюцию понятия *digital humanities* в кругах самих исследователей. Авторы приходят к вы-

воду, что утверждение *digital humanities* именно как дисциплины связано не только с административными изменениями (поддержка университетов и финансирование), но и с трансформацией самих методов исследования, где важную роль играют новые технологии. Во второй интерлюдии «Теперь анализируй это! Сравнение дискурсов о расовой политике» авторы сравнивают выступления Барака Обамы и пастора Джеремайи Райта по вопросам расовой политики. Синклер и Роквелл показывают, что цифровые методы позволяют выявить важные риторические различия — так Обама в своем обращении стремился объединить аудиторию вокруг общих проблем, в то время как Райт акцентирует внимание на необходимости признать различия — предлагая новые возможные интерпретации политических дискуссий. Третья интерлюдия «Снежный ком: анализ исследования видеоигр» анализирует становление и развитие дисциплины *game studies* (исследования видеоигр) с момента ее официального учреждения в 2001 году. Авторы пытаются понять, насколько успешно эта область избежала «колонизации» со стороны других дисциплин (кино- и литературоведения). Синклер и Роквелл исследовали архив журнала *Game Studies* за одиннадцать лет: определили самые часто употребляемые слова в статьях, идентифицировали важных авторов и ключевые организации в этой области, а также проанализировали, какие работы и авторы чаще всего цитируются в журнале. Финальная интерлюдия «Искусность диалога: мышление через скептицизм в „Диалогах“ Юма» сосредотачивается на скептицизме как методе познания. Цифровые методы позволяют определить ключевые термины и формализовать структуру аргументации участников диалогов в тексте Юма. И хотя вряд ли цифровая интерпретация добавляет что-то радикально новое к философии скептицизма, она, тем не менее, предлагает иначе взглянуть на привычный текст.

В статье «Где будут цифровые гуманитарные исследования через сто лет?» Хосе Кальво Телло утверждает, что цифровые гуманисты должны совмещать свой энтузиазм к цифровым технологиям с критикой этой сферы [Calvo Tello, 2024, P. 7]. Кажется, что Стефану Синклеру и Джеффи Роквеллу удалось это сделать. В книге (или даже цифровом проекте) «Герменевтика» они не только предлагают новый подход к осмыслению компьютерных технологий, но и переосмысливают сам формат работы в гуманитарных дисциплинах, возвращая в него живой диалог и коллаборацию. Хотя работа была опубликована еще до пандемии COVID-19, авторы указывают на наш естественный страх оказаться в чужом пространстве

и пустить других в свое. Но именно к этому призывает гибкая герменевтика. И даже если завтра *digital humanities* выйдут из моды, создавшиеся социальные связи — конфигурации и ассамбляжи — останутся продуктивной средой для новых методов и новых интерпретаций.

Примечания

- ¹ «Разбирая текст, а затем вновь собирая его в виде серии конкордансов — мы получаем новый взгляд на этот текст; возможно, даже новый текст. Реконструкцию можно представить как перестановку или трансформацию в соответствии с нелинейным, прерывистым принципом организации, будь то тематический, алфавитный или любой другой порядок» [Sinclair & Rockwell 2016, P. 48].
- ² Историко-философское осмысление этого метода можно найти в статье Ольги Алиевой [Алиева, 2024].
- ³ Совершенно случайным подтверждением этого тезиса может служить философская статья Антона Вавилова «Загадка „всегда уже“ (утраченного следа феноменологии) в деконструкции Деррида» [Вавилов 2023]. В статье автор анализирует оборот «всегда уже» (*toujours déjà, immer schon*) в работах Хайдеггера, Гуссерля и Деррида. И действительно автор приводит внушительный список работ, где используются эти обороты у этих философов [Вавилов 2023, С. 131–133], однако совсем неясно, какой *метод* использовал автор для обнаружения всех вхождений. Я благодарю Марию Стенину за то, что она обратила мое внимание на эту работу.
- ⁴ Хотя авторы упоминают Бруно Латура, возможно, полезной перспективой для их анализа могли бы стать работы Донны Харауэй, Розы Брайдотти и Карен Барад.
- ⁵ Эта возможность основывается на биоинформатическом методе выравнивания последовательностей, который используют для сравнения генов, но его также можно применять к текстам. Можно сравнить этот метод с тем, как программы для выявления плагиата ищут похожие фразы в студенческих работах.
- ⁶ Метафора воробья отсылает к «Церковной истории народа Англов» святого Беды Достопочтенного. Эту аналогию впервые использовал Уиллард Маккарти в 2001 году в тексте, посвященном четырнадцатилетию группы *Humanist*, подчеркивая изменчивость дисциплины цифровых гуманитарных исследований.

*Литература**Исследования*

- Алиева 2024* — Алиева О. В. Танцы, эрос и зачатие: о чем писали «Платоновские исследования» за последние 10 лет // Системный блок, 2024. <https://sysblok.ru/metascience/tancy-jeros-i-zachatie-o-chem-pisali-platonovskie-issledovaniya-za-poslednie-10-let>
- Вавилов 2023* — Вавилов А. Загадка «всегда уже» (утраченного следа феноменологии) в деконструкции Деррида // Horizon. Феноменологические исследования. Т. 12. №. 1. С. 115–140.
- Делёз и Гваттари 2010* — Делёз Ж., Гваттари Ф. Тысяча плато: Капитализм и шизофрения. М.: Астрель, 2010.
- Моретти 2016* — Моретти Ф. Дальнее чтение. М.: Изд-во Института Гайдара, 2016.
- Brunet 1989* — Brunet E. 1989. L'exploitation des grands corpus: Le bestiaire de la littérature française. Literary and Linguistic Computing 4 (2): 121–134.
- Calvo Tello 2024* — Calvo Tello J. (2024). Where will the digital humanities be in 100 years? The humanities as a hope for the digital. *Metode Science Studies Journal*, (15). <https://turia.uv.es/index.php/Metode/article/view/27672/31252>
- Distant Reading 2016* — Distant Reading, Computational Criticism, and Social Critique: an Interview with Franco Moretti // Zurich Open Repository and Archive, 2016. https://www.zora.uzh.ch/id/eprint/135683/1/Franco_Moretti_Interview.pdf
- Sinclair & Rockwell 2016* — Rockwell G., Sinclair S. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge: The MIT Press, 2016.

Федеральное государственное бюджетное учреждение науки
Институт русской литературы (Пушкинский Дом)
Российской академии наук
Издание зарегистрировано Министерством печати и информации
Российской Федерации
Свидетельство о регистрации ЭЛ № ФС 77 - 86683
от 22 января 2024 г.

Адрес редакции «Цифровые гуманитарные исследования»:
tsifrovye.issledovaniya@gmail.com
199034 Санкт-Петербург, наб. Макарова, д. 4
Оформление обложки И. Гурьянов