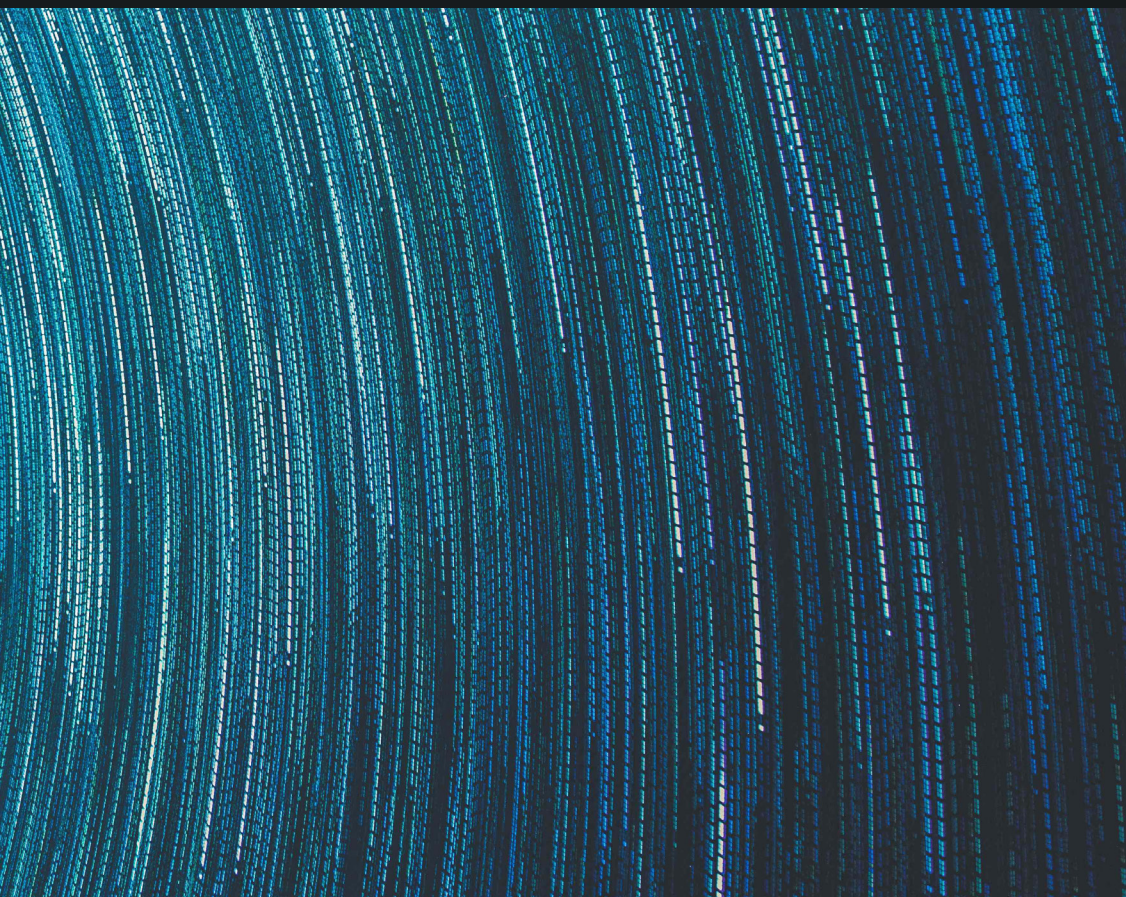
СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ



Министерство науки и высшего образования Российской Федерации
Сибирский федеральный университет

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ

Монография

Красноярск
СФУ
2023

УДК 009:004.0
ББК 71.034+32.97
Ц752

Авторы:
А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин, А.Ю. Володин, Д.А. Гагарина, Е.С. Гришин, И.А. Кижнер, Б.В. Орехов, М.В. Румянцев, А.В. Сметанин

Рецензенты:
С. А. Баканов, доктор исторических наук, заведующий кафедрой истории России и зарубежных стран историко-филологического факультета Челябинского государственного университета;
В. Н. Владимиров, доктор исторических наук, профессор кафедры отечественной истории Института истории и международных отношений Алтайского государственного университета, вице-президент Ассоциации «История и компьютер»;
Е. А. Пастернак, кандидат филологических наук, научный сотрудник Института мировой культуры Московского государственного университета им. М. В. Ломоносова

Ц752 **Цифровые гуманитарные исследования** : монография / А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин [и др.]. — Красноярск : Сиб. федер. ун-т, 2023. — 272 с.
ISBN 978-5-7638-4876-2

Впервые на русском языке комплексно рассмотрено актуальное междисциплинарное направление — цифровые гуманитарные исследования, или digital humanities. Приведены примеры (само)определения направления, дан их обзор. «Цифровой поворот» в гуманитарных исследованиях и масштабные проекты оцифровки историко-культурного наследия описаны в контексте датафикации и вызовов больших данных и машинного обучения. Особое внимание уделено современным подходам к компьютерному анализу текстов и культуромике, направлению исследований культуры и языка с помощью больших текстовых данных. Представлена широкая палитра цифровых подходов, призванных находить решения насущных гуманитарных исследовательских задач: от базы данных к сетевому анализу, от геоинформационных систем к виртуальным реконструкциям и дополненной реальности. Происходящие процессы рассмотрены в связи со становлением сложной и противоречивой информационной инфраструктуры цифровых гуманитарных исследований.

Будет интересна широкому кругу гуманитариев — историкам, филологам, философам, культурологам — и всем сочувствующим и сопереживающим цифровой трансформации современной культуры.

Электронный вариант издания см.:
http://catalog.sfu-kras.ru

УДК 009:004.0
ББК 71.034+32.97

ISBN 978-5-7638-4876-2

© Сибирский федеральный университет, 2023

Оглавление

Предисловие

Идея коллективного труда, который вы держите в руках, появилась во время подготовки онлайн-курса «Введение в цифровые гуманитарные исследования» (https://openedu.ru/course/sfu/IDH/) на базе производственно-продюсерского центра Сибирского федерального университета.

Цифровые гуманитарные науки (или digital humanities) стали важным направлением развития современных методических подходов к решению исторических, лингвистических, культурологических, философских проблем.

Однако все еще ощущается недостаток в литературе на русском языке, которая могла бы поспособствовать вхождению желающим в такую сложную, но захватывающую проблематику.

В издательстве Сибирского федерального университета увидела свет хрестоматия «Цифровые гуманитарные науки», переводное издание, отразившее сложные перипетии становления и самоопределения направления. Хрестоматия вошла в программы большинства смежных дисциплин и получила признание у преподавателей. Но исследования не стоят на месте, в России активно развиваются разные методологические направления цифровых гуманитарных исследований, настал удачный момент соединить исследовательские наработки с опытом преподавания дисциплин цифрового цикла.

Под одной обложкой вы найдете главы о данных и базах данных, о культуромике и анализе текстов, о географических информационных системах и сетевом анализе, о трехмерном моделировании и об инфраструктурах современной цифровой гуманитарной науки.

Не все аспекты богатого исследовательского поля цифровой гуманитаристики удалось охватить в этом первом коллективном опыте, но есть надежда, что коллективная монография станет живой — обновляющейся и дополняющейся от издания к изданию. И наши труды будут способствовать развитию и процветанию цифровой гуманитарной науки в России.

Максим Румянцев,
ректор СФУ

Глава 1

Digital humanities: (само)определение, обзор направлений

(*А. Ю. Володин, Б. В. Орехов*)

Digital humanities — направление исследований, которое за два десятилетия завоевало свое место в научной повестке гуманитарных междисциплинарных компьютеризированных изысканий. Компьютеризация в гуманитарных науках, в принципе, имеет давние традиции: к помощи компьютерной техники ученые-гуманитарии начали обращаться еще в эпоху больших вычислительных машин. Но настоящий «цифровой поворот» в гуманитарных науках начался после микрокомпьютерной революции, с развитием вычислительных мощностей и персонализацией компьютерных систем, позволяющих производить сложные подсчеты в домашних условиях, не только создавать сложные виртуальные реконструкции, но и представлять их в электронной среде с помощью многообразных средств Всемирной паутины. За последние десятилетия существенно снизился порог входа в тематику цифровых методов, все больше ученых вовлекается в орбиту количественных исследований, что приводит к значительным институциональным и концептуальным изменениям в смежных гуманитарных дисциплинах.

Термину digital humanities не повезло с переводом на русский язык. Проблема кроется в отсутствии в русскоязычной традиции удобного эквивалента для слова humanities, которое трудно перевести кратко. Наиболее адекватным является длинный перевод из двух слов «гуманитарные науки», в реальном употреблении он как будто требует сокращения, и это сокращение находится в неудачном слове «гуманитаристика».

Неудачность этого слова в том, что стилистически оно не нейтрально, то есть создает ненужный сниженный эффект, давно

замеченный лингвистами для слов с формантом -истика (шагистика, ерундистика)[1]. В более выгодном положении оказываются давно заимствованные слова «журналистика», «логистика», «лингвистика», но, чтобы оказаться в этом ряду, все еще имеющей ореол новизны «гуманитаристике» придется проделать длинный и сложный путь адаптации в русском языке. Отдельно от определения «цифровой» это слово в русском научном языке почти не употреблялось, поэтому и не воспринимается как привычный нейтральный термин.

Новизна здесь не случайна: digital humanities появились в научном поле как самостоятельная сфера всего несколько десятилетий назад, поэтому их границы, объект, методы и ценности не успели приобрести отчетливых очертаний. Процесс оформления в дисциплину продолжается на наших глазах, поэтому любые обобщения, сделанные сейчас, будут носить предварительный характер.

Если максимально упростить схему, то digital humanities являются продолжением соответствующих областей гуманитарного знания. Цифровые гуманитарные науки распадаются на цифровое литературоведение, электронные публикации, цифровое искусствоведение (включая музееведение), цифровое киноведение. Но важным направлением внутри digital humanities является и цифровая история, несмотря на то, что статус истории как гуманитарной, а не социальной науки остается дискуссионным[2]. В этом смысле компонент «гуманитарный» в составе термина «цифровые гуманитарные науки» оказывается и обманчивым, и точным.

Точен он потому, что в отечественной традиции история обычно локализуется внутри гуманитарного поля. Если смотреть на это через призму институционализации, то исторические дисциплины изучаются внутри гуманитарного блока учебных предметов, кафедры истории размещены на гуманитарных факультетах, а исторические факультеты являются частью гуманитарных вузов (например, Российский государственный гуманитарный университет возник на базе Историко-архивного института). Еще в середине XX века привычным для советского вуза было наличие общего историко-филологического факультета.

---

[1] Пацюкова О. А. Переразложение и закономерности развития протяженных аффиксов в русском языке: дис. … д-ра филол. наук. Нижний Новгород, 2014. С. 186.

[2] Валлерстайн И. Миросистемный анализ. Введение. М.: УРСС: ЛЕНАНД, 2018. С. 60–65.

Обманчив он потому, что не дает ясного понимания специфики digital humanities. Как выглядела бы эта область без истории? Если бы digital humanities состояли только из цифрового литературоведения (computational literary studies), цифрового искусствоведения и других подобных дисциплин, то можно было бы сравнительно четко очертить круг исследовательских интересов, составляющих ядро цифровых гуманитарных наук. Этот круг включал бы прежде всего семиотические объекты второго порядка или целые вторичные семиотические системы.

Семиотика — это наука о знаковых системах. Эти системы позволяют людям обмениваться информацией. Например, знаковую систему составляют сигналы светофоров, дорожная разметка, азбука Морзе, математические формулы, музыкальная нотация. Красный свет светофора — это знак, у него есть значение: команда остановиться. Буква Σ в математической формуле — это знак, у него есть значение: сумма. Но самой сложной и развитой знаковой системой в распоряжении человека является естественный язык: русский, английский, амхарский, рутульский и т.д. Описанием того, как устроен и как функционирует язык, занимается наука лингвистика (стереотипное представление, будто бы лингвистика — это изучение иностранных языков, следует считать распространенным заблуждением; кстати, заблуждение, по всей видимости, и то, что лингвистика относится к гуманитарным наукам, она гораздо ближе естественным, вроде биологии). Лингвистика создала сложный терминологический аппарат для описания слов, их грамматических характеристик и семантики, способов сочетания этих слов между собой в предложении и тексте. Выше мы использовали как раз лингвистические термины: «формант», «стилистическая нейтральность». Пользуясь лингвистическим знанием, можно не только описать, какие слова и как расставлены в тексте, но и объяснить, почему высказывание устроено именно так.

Базовым основанием лингвистики как науки является представление о языке как о всеобщем и самостоятельном инструменте общения. Иными словами, язык знают все его носители, он не принадлежит кому-то одному (поэтому всеобщий), а еще то, что мы знаем о языке, невозможно вывести из наших знаний о человеческой биологии или психологии (поэтому самостоятельный). Это довольно важная причина считать языкознание самостоятельной наукой: ни одна другая научная дисциплина не способна описать согласовательные классы слов, актантную деривацию или временной дейксис.

Язык — это семиотическая система первого порядка. С помощью содержащихся в языке знаков мы передаем друг другу информацию. Но, используя язык, люди стали создавать такие объекты, описать и объяснить которые лингвистика уже не может и не стремится. Речь о художественной литературе. В литературных произведениях появились свои знаки, которые так или иначе имеют языковое выражение, но которые невозможно свести к языку и ограничиться лингвистическим описанием.

К таким знакам, например, можно отнести кольцевую композицию, лирические отступления в нарративном тексте, систему персонажей. Лирическое отступление в повествовательном по своей сути романе в стихах «Евгений Онегин» — это знак, он указывает читателю на игру Пушкина с традицией лиро-эпических поэм Байрона. Кольцевая композиция рассказа Набокова «Круг» — это знак, он актуализирует форму организации текста, выдвигает ее на первый план в восприятии произведения по отношению к содержанию. Появление в детективном сюжете опытного разгадывателя загадок и его наивного, но верного и преданного спутника, как в романе «Имя Розы» Умберто Эко, — это знак, он вызывает в памяти ассоциации с классическими рассказами о Шерлоке Холмсе.

Знаковая природа таких элементов текста очевидна, она использует язык как материал, но недоступна для описания с помощью терминологического аппарата лингвистики, потому что названные (а также множество не упомянутых здесь) элементы созданы в иной семиотической системе. Это уже не система языка, а надстроенная над ней система литературы, имеющая единицы, правила и закономерности.

Такими семиотическими системами второго порядка, надстроенными над первичными, но не сводимыми к последним, и занимается гуманитарная наука. Сигналы светофора — это семиотика первого порядка, с ее помощью передаются сигналы участникам дорожного движения. Светофор, изображенный на картине или на художественной фотографии, встраивается в систему второго порядка, участвует в композиции кадра, определяет цветность и мотивное наполнение изображения. К регулированию действий пешеходов и автомобилистов все это отношения уже не имеет.

Но история в этом смысле не похожа на гуманитарные науки, ее предметом не являются семиотические системы ни первого, ни второго порядка. Разумеется, историк может привлекать к своим

изысканиям текстовые источники — так же, как это делает лингвист или литературовед, но делает он это с другими целями.

Историческое знание принципиально опосредовано объекту своего исследования и основывается на реконструкции. Исследователь восстанавливает прошлое на основе исторических источников, «остатков» и «преданий» старины. Можно назвать такую реконструкцию информационным моделированием прошлого. Как верно подметил Ю. М. Лотман, «историк с самого начала попадает в странное положение: в других науках исследователь начинает с фактов, историк получает факты как итог определенного анализа, а не в качестве его исходной точки»[1]. Историки традиционно чутки к различиям в средствах передачи информации (как в знаменитой формуле М. Маклюэна — «The medium is the message» ‘то, что передает сообщение, само по себе является сообщением’). Доказательством тому служит долгая дискуссия о классификации исторических источников и разнообразие специальных исторических дисциплин, ориентированных на конкретные виды исторических источников.

То общее, что интересует и историков, и представителей типичных гуманитарных наук, лежит в особой плоскости, а именно в плоскости противопоставления номотетических и идиографических дисциплин.

Слово «номотетический» восходит к греческому *nomos* ‘закон’, этим термином объединяются науки, которые работают с повторяющимися явлениями, а их конечной целью является описание законов и закономерностей. Хрестоматийные примеры — из классической механики вроде ускорения свободного падения. Нахождение таких законов и их математическое описание является сверхзадачей естественных наук: физики, химии, физиологии и подобных.

Им противостоят идиографические науки, которые сосредоточены на уникальных явлениях, к которым относятся исторические события и события истории литературы и искусства. Приход к власти Елизаветы Петровны, роман «Обломов» или стихотворение «Письмо матери» рассматриваются как единственные в своем роде, и в оптике этих дисциплин не могут стать частным случаем проявления какого-то закона. К идиографическим относятся и история, и гуманитарные науки.

[1] Лотман Ю. М. Изъявление Господне или азартная игра? (Закономерное и случайное в историческом процессе) // Ю. М. Лотман и тартуско-московская семиотическая школа. М.: Гнозис, 1994. С. 353–363.

Именно в этом кроется нетипичность цифрового подхода к гуманитарному материалу. Количественные методы хорошо зарекомендовали себя там, где анализ повторяющихся фактов помогает открывать и описывать закономерности, подтверждаемые новыми наблюдениями. То, что является ценным для идиографических дисциплин, не повторяется и не может быть подтверждено.

Чтобы перейти от традиционных идиографических описаний к применению количественных методов, необходимо перестроить сам взгляд на материал таким образом, чтобы он позволял видеть общее в уникальном. Например, не ставя под сомнение единственность в своем роде каждого отдельного стихотворения, ученый может сосредоточиться на некотором формальном приеме, истории его применения в рамках определенной художественной традиции. В таком случае попавший в фокус внимания исследователя прием уже может иметь количественное измерение, которое может быть описано как закономерность. Так действуют применяющие квантитативные методы стиховеды, подсчитывающие частотность использования поэтических размеров: «Обследуются не типы, а массы, не единичные явления, а общие состояния. Рассматриваются многие тысячи стихов данного размера, рассматриваются в различных разрезах, и — “большие числа” торжествуют над случайностью, выстраивая закон»[1].

Такой подход, естественно, вызывает закономерный скепсис в глазах традиционных филологов и искусствоведов. Подсчет частотности вступает для них в противоречие с привычными формами работы с материалом. Поэтому неверно было бы разделять историю гуманитарных наук на доцифровой и цифровой периоды: одновременно с активным использованием компьютерных методик существуют и более привычные подходы.

Тем не менее вряд ли можно найти гуманитарную специальность, которую бы так или иначе не затронул «цифровой поворот». Любое гуманитарное исследование сегодня основано на спонтанной или систематической, выборочной или сплошной оцифровке текстов, документов, изображений или каких-то объектов историко-культурного наследия, что делает эти объекты более доступными для исследователя, то есть снижает порог входа для ученого. Если в доцифровую эру на пути у филолога, историка, искусствоведа могла

[1] Шенгели Г. А. Об исследовании узбекского стиха // Научная мысль. 1930. № 1. С. 29.

стоять недоступность текста, живописного полотна или документа, то в современной ситуации наличие цифровой копии вовлекает в исследовательский процесс все большее число специалистов. Так, академик Е. Э. Бертельс, составляя историю персидско-таджикской литературы, обозначал масштаб проблемы: «ссылаться на книгу, которая имеется только, скажем, в библиотеке Института востоковедения в Ленинграде и которую нельзя найти даже и в Москве, или на рукопись, находящуюся в частной библиотеке, было бы просто недобросовестно»[1]. Оцифровка стала одной из важных ежедневных практик ремесла гуманитария. В этой связи возникает широкий спектр вопросов, в чем преимущества и недостатки наступления цифровой эры в гуманитарных исследованиях. Такие вопросы оказываются главными в весьма обширной литературе, посвященной проблемам определения, самоопределения и развития междисциплинарного направления digital humanities.

В каком-то смысле возможности электронной публикации и сетевого доступа начинают играть роль «дополненной реальности», когда классические формы научного творчества (статьи, монографии) дополняются электронными ресурсами, содержащими цифровые приложения, часто имеющие самостоятельное научное значение. Следует отметить, что в среде специалистов наблюдается относительный консенсус о том, что цифровые гуманитарные науки предполагают не только использование компьютера как исследовательского инструмента, но и расширение цифрового историко-культурного наследия путем публикации электронных ресурсов, реконструкций и визуализаций. Без таких публикаций исследование может быть компьютеризированным, но не может относиться к направлению DH.

По той же причине для истории использование компьютерных методов оказывается особенно актуальным. Сегодня можно по-новому взглянуть на знаменитую максиму Э. Ле Руа Ладюри — «историк будущего будет программистом или его не будет вовсе» (“l’historien de demain sera programmeur ou il ne sera plus”; Le nouvel observateur, 1968). Историки, бесспорно, стали пользователями (а иногда и создателями) весьма разнообразного программного обеспечения. Начали реализовываться давние мечты об историко-ориентированном программном обеспечении. К таким разработкам

[1] Бертельс Е. Э. Избранные труды. Т. 1. История персидско-таджикской литературы. М.: Изд-во восточной литературы, 1960. С. 28.

можно отнести уже приобретшие мировую известность продукты Центра истории и новых медиа имени Р. Розенцвейга: программа Zotero позволяет сохранять и управлять найденными онлайн научными материалами; Omeka предназначена для создания специализированных электронных ресурсов — электронных коллекций и онлайн-выставок; Scripto создан для облегчения совместной работы по расшифровке и установлению текстов по электронным копиям архивных документов. Именно по этой причине рассмотрение новых возможностей работы с источниками информации определяет профессиональные аспекты исторического исследования в цифровую эпоху. Историки сосредоточились на изучении исторических источников, представлении исторических сведений в формате баз данных, оцифровке и электронной публикации свидетельств прошлого[1], а вслед за оцифровкой — на моделировании исторических процессов и объектов в самом широком смысле этого понятия: от математических моделей поведения до трехмерных моделей объектов прошлого[2].

Цифровая история (хотя такой перевод digital history и вызывает критику) сегодня часто рассматривается как область активной разработки инструментов обработки и исследования исторических источников для их адекватного представления в современных медиаформатах (в последние годы преимущественно онлайн). Термин digital history приобрел права гражданства в 1997 году, когда американские исследователи Э. Айерс и У. Томас основали Вирджинский центр цифровой истории (Virginia Center for Digital History, VCDH) при университете Вирджинии, хотя один из пионеров разработок в этой области Р. Розенцвейг еще в 1994 году открыл Центр истории и новых медиа (Center for History and New Media, CHNM) в университете Дж. Мейсона. Первые работы, посвященные осмыслению цифровой истории, были написаны на рубеже XX–XXI веков, в частности, можно отметить полемическую статью Э. Айерса «Прошлое и будущее цифровой истории» и фундаментальную монографию Д. Коэна и Р. Розенцвейга «Цифровая история: руководство по сбору, сохранению и представлению прошлого во Всемирной паутине»[3].

---

[1] Rosenzweig R., Grafton A. Clio Wired: The Future of the Past in the Digital Age. Columbia University Press, 2011.

[2] Бородкин Л. И. Моделирование исторических процессов: от реконструкции реальности к анализу альтернатив. СПб.: Алетейя, 2016.

[3] Ayers Edward. L. 1999. The Pasts and Futures of Digital History (online essay). URL: http://www.vcdh.virginia.edu/PastsFutures.html (дата обращения: 26.08.2023);

В коллективной монографии «История в цифровую эпоху» под редакцией Т. Веллер сформулирован ключевой для историков подход к цифровой эпохе — «цифра» затрагивает всех, кто профессионально изучает историю, при этом из этого не следует, что историк обязательно должен становиться компьютерным гуру или разбираться в языках программирования. Главное, чтобы информационные технологии и цифровые достижения, напрямую касающиеся профессиональных нужд историков, не оставались закрытым самодостаточным полем исследований отдельных специалистов (как частично получилось в случае с «количественной историей»), не становились маргинальными вопросами хотя бы по той причине, что по большей части цифровая эпоха касается основ методологии и методики исторического исследования[1].

Кроме того, сегодня заметен определенный историографический переход, состоявшийся с развитием средств компьютерной визуализации и сетевых технологий. Данный переход можно условно датировать серединой 2000-х годов, когда постепенно стало происходить терминологическое изменение: от исторического или гуманитарного компьютинга (humanities computing, history and computing) к цифровым гуманитарным наукам или цифровой истории (digital humanities, digital history). Перемена названия означала постепенное изменение статуса — от технической поддержки к интеллектуальному прорыву со своими профессиональными практиками, научными стандартами и теоретическими построениями. Во многом переход от «измерительных» возможностей компьютерных технологий к реконструкционным и презентационным связан с освоением интернет-технологий в исторической науке[2].

Сегодня и в зарубежной литературе все чаще встречается точка зрения, что, например, «цифровая история» (digital history) имеет

Cohen D., Rosenzweig R. Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. University of Pennsylvania Press, 2002.

[1] History in the digital age. London; New York: Routledge, 2013. В книге также делается занятное различение между «цифровыми историками» (специалистами, целенаправленно изучающими и внедряющими информационно-технологические решения в исторические исследования) и «историками в цифровую эпоху» (всеми профессиональными исследователями прошлого).

[2] Владимиров В. Н. Интернет для историка: и все-таки новая парадигма! // Круг идей: историческая информатика в информационном обществе: Труды VII конференции АИК. М., 2001; Володин А. Ю. «Цифровая история»: ремесло историка в цифровую эпоху // ЭНОЖ «История». 2015. Т. 6. № 8; Гарскова И. М. Историческая информатика. Эволюция междисциплинарного направления. СПб.: Алетейя, 2018.

важные отличительные черты, которые нежелательно смешивать с общими для гуманитарных наук трендами «дигитализации». Ярко эту мысль выразил С. Робертсон (директор Центра истории и новых медиа имени Р. Розенцвейга): рассматривая объединенные общей методологической платформой «цифровые гуманитарные науки» нельзя не заметить, что «источники, исследовательские вопросы и подходы, которые они используют в своих проектах, дисциплинарны, равно как дисциплинами определяется выбор цифровых инструментов»[1]. Такие мысли вполне созвучны российским дискуссиям на эту тему[2]. Например, М. Таллер разделил сообщество DH на четыре условные группы: 1) исследователи «текста как такового», 2) исследователи-собиратели «фактов» в электронных (иногда весьма обширных) коллекциях, 3) исследователи «нетекстов» (в том числе виртуальных реконструкций), 4) исследователи влияния цифровой среды на гуманитарные науки в целом[3].

В 2004 году был опубликован доклад о будущем historical information science (условно можно перевести как «об исторической информатике»), в котором основная перспектива развития виделась в сотрудничестве историков с учреждениями — хранителями историко-культурного наследия[4], так как именно в рамках такого сотрудничества могут сложиться основы для оцифровки и последующего компьютерного исследования исторических источников в самом широком смысле этого слова[5]. В тот момент историки лишь подходили к поиску «цифровой перспективы» исследований, о которой уже громко заявляли в работах филологи. Важной отправной точкой для дискуссии о развитии исследований в русле цифровых исследований в этой области стало издание «Компаньона по цифровым гуманитарным наукам», в котором были собраны многочисленные статьи, посвященные вопросам междисциплинарного синтеза в истории, филологии, археологии, антропологии и социальных науках. Фактически именно с выходом «Компаньона» в 2004 году

[1] Робертсон С. Различия между цифровыми гуманитарными науками и цифровой историей // Электронный научно-образовательный журнал «История». 2016. Т. 7. Вып. 7 (51). DOI: 10.18254/S0001648–1–1.

[2] Бородкин Л. И. Digital history: применение цифровых медиа в сохранении историко-культурного наследия? // Историческая информатика. 2012. № 1.

[3] Таллер М. Дискуссии вокруг Digital Humanities // Ист. информатика. 2012. № 1.

[4] Boonstra O., Breure L., Doorn P. Past, present and future of historical information science (Glasgow meeting, 25.04.2004). Amsterdam, 2004.

[5] Подробнее см.: McCrank L. J. Historical Information Science: An Emerging Unidiscipline. Information Today, 2002.

«цифровые гуманитарные науки» начали свое шествие по планете[1]. В 2016 году увидел свете уже «Новый компаньон по цифровым гуманитарным наукам»[2]. Оба компаньона вышли под редакцией трех исследователей — Сьюзан Шрейбман, Рея Сименса и Джона Ансворса. Еще в 2004 году редакторы называли «Компаньон» «поворотным пунктом в области цифровых гуманитарных наук», потому что «впервые широкий круг теоретиков и практиков, тех, кто работал в этой области в течение многих десятилетий, и тех, кто присоединился недавно, эксперты в разных дисциплинах, ученые-компьютерщики, специалисты в библиотечном деле и информационных исследованиях были объединены, чтобы рассмотреть цифровые гуманитарные науки как самостоятельную область знаний, а также задуматься о том, как она соотносится с традиционными гуманитарными исследованиями». Годы спустя редакторы замечают, что, хотя «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» (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 компьютинга 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».

«Новый компаньон» включает пять частей: инфраструктуры, создание, анализ, распространение и прошлое, настоящее, будущее

[1] A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004. URL: http://www.digitalhumanities.org/companion/

[2] A New Companion to Digital Humanities / Eds Susan Schreibman, Ray Siemens, John Unsworth. Wiley-Blackwell, 2016. 592 p. ISBN: 978–1–118–68059–9.

цифровых гуманитарных наук. (Для сравнения: в компаньоне 2004 года было четыре части: история, принципы, приложения и производство, распространение, 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 цифровых гуманитарных наук. Раздел о положении в области описывает существующий научный ландшафт, показывает влияние глобализации и интернетофикации, 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 пословицу «больше вкалывай — меньше болтай» (more hack less yack) в отношении таких создающихся областей, как цифровые гуманитарные науки (p. 538). У. Томас поддерживает эту идею следующим размышлением: «Историки, литературные критики, философы, филологи, ученые, открывшие для себя цифровые гуманитарные исследования, 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» (p. 524).

Любопытно, что руководства и дискуссионные тома, вышедшие вслед за «Компаньонами» уже фокусируются не на подходах и инструментах исследования, а на критическом рассмотрении контекстов, в которых эти исследования ведутся, погружая дискуссии о цифровых гуманитарных исследованиях в широкую философскую и политическую парадигму[1].

Цифровые методы получают все большее распространение в сфере гуманитарных наук и приобретают свои организационные формы в виде специализированных конференций, исследовательских центров и журналов.

Свои исследовательские центры, действующие в области цифровых гуманитарных наук, есть у множества крупных университетов по всему миру. Статус наиболее авторитетного в сфере цифрового литературоведения удалось получить Стэнфордской литературной лаборатории. В исторических исследованиях широкую известность приобрели Центр истории и новых медиа имени Роя Розенцвейга (RRCHNM), Люксембургский центр современной и цифровой истории (C2DH).

В России соответствующие институции имеются в Москве (МГУ, НИУ ВШЭ), Санкт-Петербурге (ИТМО, Институт русской литературы РАН), Барнауле (АлтГУ), Екатеринбурге (УрФУ), Калининграде (БФУ), Красноярске (СФУ), Перми (НИУ ВШЭ-Пермь), Ростове-на-Дону (ЮФУ), Томске (ТГУ).

К зарекомендовавшим себя авторитетам научной периодики относятся *Digital scholarship in the humanities* (с 2012 года, с 1986 года выходил под названием Literary and Linguistic Computing), *Digital humanities quarterly* (с 2007 года), Историческая информатика (с 2012), отчасти *Journal of Cultural analytics* присоединяются *International Journal of Digital Humanities* (с 2019 года), Квантитативная филология (с 2021 года), *Journal of Digital History* (с 2021 года), *Digital Orientalia* (с 2021 года), *Journal of Computational Literary Studies* (с 2023 года), Цифровые гуманитарные исследования (с 2023 года).

[1] The Bloomsbury Handbook to the Digital Humanities / Ed. James O'Sullivan. Bloomsbury Academic, 2022. 512 p.; Debates in the Digital Humanities 2023 / Eds. Matthew K. Gold and Lauren F. Klein. Univ Of Minnesota Press, 2023. 520 p.

Важными для определения границ научного поля становятся конференции, центральное место среди которых занимает ежегодная созываемая ассоциацией организаций цифровых гуманитарных наук (ADHO).

Состав секций и круглых столов этой конференции лучше всего описывает ключевые направления, в которых работают представители цифровых гуманитарных наук:

- самоопределение digital humanities;
- проблемы оцифровки, организации электронных изданий и публикации данных;
- методы, в частности, обработка естественного языка, сетевой анализ, компьютерное зрение.

С исследовательской точки зрения DH — это проектный подход к решению научных проблем, предполагающий в качестве итога исследовательского труда конкретный информационный цифровой продукт (набор данных, онлайн-ресурс, информационную систему). Проектный подход в том числе означает и соавторство, участие в исследовании нескольких авторов, каждый из которых вносит свой вклад, соразмерный его компетенциям. Действительно, обычная для цифровых гуманитарных наук картина — это наличие нескольких авторов у одной публикации (пример этому — данная монография). В то же время известно, что у абсолютного большинства статей на гуманитарную тематику только один автор[1].

С образовательной точки зрения цифровые гуманитарные науки можно рассматривать как привлекательное для студентов направление обучения. В таком смысле DH — это комплекс дисциплин, позволяющих представить специфику изучения гуманитарных проблем в современных условиях, то есть в эпоху «больших данных» и исследовательских облачных платформ. Как показывает практика, образовательные программы DH весьма популярны (как на бакалаврском, так и на магистерском уровне) в США, Великобритании, Германии, Франции, Японии, Австралии. При этом дисциплины, преподаваемые в соответствующих циклах, имеют существенный технологический уклон, не теряя очевидной возможности включения гуманитарных знаний и исследований в актуальную мультимедийную среду. Можно сказать, что уже складывается определенный образовательный DH-канон — оцифровка, модели и базы данных,

[1] Жэнгра И. Ошибки в оценке науки, или Как правильно использовать библиометрию. М.: НЛО, 2018. С. 44.

метаданные и разметка, интеллектуальный и сетевой анализ, визуализация и картографирование данных, трехмерное моделирование, веб-ресурсы и интерфейсы, проектный подход и интеллектуальная собственность[1]. При этом все перечисленные знания и умения не отменяют необходимости глубоко разбираться в конкретном предмете гуманитарного исследования.

И, наконец, с точки зрения профессионального сообщества, DH — это полезный бренд, позволяющий обращаться за финансированием и административной поддержкой, предлагая инновационные решения для вполне классических гуманитарных дисциплин. В последнее время, особенно в контексте цифровизации и успехов алгоритмических решений в повседневной жизни, наблюдается значительный общественный интерес к результатам реализации проектов в области DH, в том числе и потому, что присутствие такого рода проектов в интернете делает их более доступными любопытствующей публике. Вместе с тем онлайн-публикация результатов исследований способствует и международным дискуссиям. Такого рода перемены свидетельствуют о прямой практической пользе от «цифрового поворота».

Не будем скрывать скептического отношения многих традиционных гуманитариев, подмечающих и методологические слабости молодой компьютерной области, и легковесность стоящих за цифровыми исследованиями концепций, и неоднозначность продуцируемых выводов. Однако междисциплинарное направление DH за два десятилетия не только заявило о себе, но и утвердилось, нашло свою нишу в исследовательском сообществе и сформировало подход к реализации гуманитарных исследовательских проектов в цифровую эпоху. Необходимо подчеркнуть, что цифровые гуманитарные науки сформулировали цели и задачи в конкретный исторический момент — значительного увеличения вычислительных возможностей и расширения сетевого международного взаимодействия, при этом не превратились в закрытый клуб исследователей. По сути, современные цифровые гуманитарные науки предполагают широкую исследовательскую программу, которая включает вопросы, интересующие любого гуманитария. Цифровые исследовательские практики — это реальность любого ученого.

[1] См., например: Drucker J. The Digital Humanities Coursebook. An Introduction to Digital Methods for Research and Scholarship. Routledge, 2021.

Предсказывать вектор развития DH сложно. Эта область не вполне самостоятельна: ее будущее зависит от развития технологий, и траектории будущего дисциплины находятся в прямой связи с появлением и проработкой новых методов. Еще в начале 2010-х годов было крайне трудно распознать, что именно искусственные нейронные сети станут настолько эффективным инструментом решения множества интеллектуальных задач. Нет возможности предсказать, какие именно технологии появятся в ближайшем будущем. Но все же в общих чертах ясно, что главный курс дальнейшего развития DH связан с применением различных методов анализа гуманитарных данных и расширением разнообразия результатов таких исследований. Цифровые гуманитарные науки будут двигаться в сторону моделирования все более сложных и трудноформализуемых объектов и уровней. Значительный шаг в этом направлении — моделирование семантики, реализованное в форме векторных моделей (см. главу об анализе текста).

Глава 6

Компьютерный анализ текста

(*Б. В. Орехов*)

Цифровое исследование в гуманитарной науке оказывается возможным там, где удачным образом сходятся три компонента: данные, методы и исследовательский вопрос.

Наиболее удобным источником данных для количественных исследований является текст. В свернутом виде он содержит информацию, релевантную для историков, культурологов, лингвистов и литературоведов. Есть случаи, когда зафиксированная в текстовой форме информация оказалась востребована и представителями естественных наук[1].

Текстовые данные занимают не так много места, как мультимедийные файлы, они проще для оцифровки и обильно представлены в современном интернете.

Методы для машинного анализа текста начали разрабатываться задолго до современного компьютерного бума инженерной областью, которая называется компьютерной лингвистикой. В большинстве своем эти методы основываются на идее значимости частоты, с которой те или иные языковые единицы (например, слова) встречаются в тексте. Идея частотности делает осмысленной операцию подсчета, на которой основывается цифровой взгляд на гуманитарный материал.

С формированием современной ситуации, подразумевающей доступность больших вычислительных мощностей и свободно распространяемые крупные текстовые архивы, разработка методов пошла быстрее. Многие хорошо зарекомендовавшие себя технологии

[1] Neuhäuser R. et al. Colour evolution of Betelgeuse and Antares over two millennia, derived from historical records, as a new constraint on mass and age // Monthly Notices of the Royal Astronomical Society. 2022. Т. 516. № 1. С. 693–719.

оформлены в виде функций открытых программных библиотек, прежде всего, для языка Python.

Исследовательские вопросы вырастают из конкретной предметной области, из ее традиций и наработок. Часто это могут быть попытки уточнить какие-то положения, сформулированные без применения количественных методов, добавить дигитальные аргументы в теоретическую дискуссию, проверить релевантность ранее сделанных выводов на большом объеме данных.

Подготовка текстовых данных для машинного анализа

Работа с текстовыми данными предполагает, что мы подготовим их до применения компьютерных аналитических инструментов. Это касается не только текста, но и любых данных, у каждого типа данных есть своя специфика, отражающаяся на их предварительной обработке. Для изображений иногда требуется представить имеющуюся в нашем распоряжении коллекцию так, чтобы все они были одинакового размера и в одинаковом формате (например, содержали одинаковое число каналов для передачи цвета). Иногда нужно «убрать» все цвета и сделать изображение черно-белым. Такие же операции унификации данных (или, как еще говорят, препроцессинга) и их избавления от «шума» существуют для текста.

У всех операций, которые входят в препроцессинг, есть свой исследовательский смысл, поэтому их не следует рассматривать как чисто механические. Их выполнение полностью зависит от конечной исследовательской задачи и тех методов, которые мы избираем для ее решения. Например, лемматизация (см. ниже) уместна перед тематическим моделированием, но является избыточной для задач стилеметрии.

Данные, которые мы используем в исследованиях, существуют в своего рода «дикой природе». Это означает, что они отягощены контекстом своего бытования. Есть обстоятельства, при которых этот контекст нам не важен или даже мешает, и его следует отделить от данных или просто удалить до анализа.

Так, например, большое количество текстовых данных содержится в интернете. Развитие Всемирной сети вообще сильно повлияло на развитие науки о данных: до широкого распространения интернета

основное (если говорить образно) «топливо» науки найти было не так просто. Появлением многих современных технологий обработки текста мы тоже обязаны тому, что интернет стал огромным, и в нем хранится много текстовой информации. Такие технологии основаны на статистике распределения слов и поэтому чувствительны к объемам информации, с которыми они работают. Если данных мало, статистика дает сбои, показывает не тенденции, а шум, случайные значения. Если данных много, статистика с помощью измерений позволяет устанавливать закономерности. Поэтому, если бы у нас не было интернета в качестве почти неисчерпаемого источника текстов, мы бы не смогли обучать современные нейросетевые модели, полностью основанные на статистической индукции.

Но тексты берутся не только из интернета. Для цифровых гуманитарных исследований одним из источников текстовых данных является оцифровка документов. Документами в терминологии компьютерной лингвистики и информационного поиска называются и книги, и рукописи, и плакаты, то есть любые объекты, на которых может быть что-то написано или изображено.

Существуют очевидные и не слишком очевидные особенности предобработки текстов, с которыми исследователю приходится сталкиваться перед тем, как перейти к анализу.

Среди очевидного — ошибки распознавания символов, которые нужно исправлять. Оптическое распознавание символов, по-английски сокращенно называется OCR (optical character recognition), — это то, с чем сталкивается любой специалист, работающий с оцифровкой документов.

Менее очевидно, что в том же интернете тексты размещены на HTML-странице, а страницы содержат код, который помогает их отображению.

Например, вот такое представление имеет HTML-код вокруг содержательного текста на странице сайта rvb.ru (Русская виртуальная библиотека):

```
<hr class="color-19">

<h1>ПОСЛАНИЕ К ЖУКОВСКОМУ В ДЕРЕВНЮ</h1>
<div class="versusia6">
<span class="line" id="L1">Итак, мой милый друг, оставя
скучный свет</span><br>
```

```
<span class="line" id="L2">И в поле уклонясь от шума
и сует,</span><br>
<span class="line" id="L3">В деревне ты живешь, спокойный
друг природы,</span><br>
<span class="line" id="L4">Среди кудрявых рощ, под сению
свободы!</span><br>
<span class="line" id="L5">И жизнь твоя течет, как свет-
лый ручеек,</span><br>
<span class="line" id="L6">Бегущий по лугам, как легкий
ветерок,</span><br>
<span class="line" id="L7">Играющий в полях с душистыми
цветами</span><br>
<span class="line" id="L8">Или в тени древес пастушки
с волосами.</span><br>
```

Этот код тоже представляет собой текст, но для исследования он не важен. Терминологически такой код называется «обвязкой». С помощью кода текст на веб-странице перемежается с рекламой и видеоплеерами.

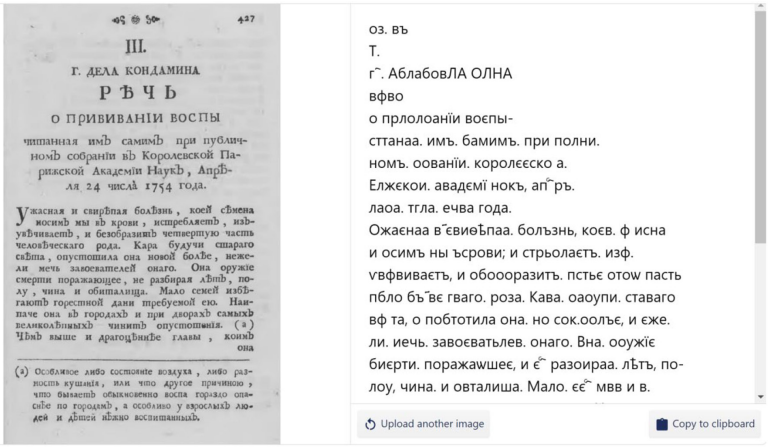
```
<span class="line" id="L27">То лирою своей Климену вос-
хищаешь,</span><br>
<span class="line" id="L28">То быстро на коне несешься
по полям,</span></div>
<div class="page" id="pg49">49</div>

<div id="yandex_rtb_R-A"></div>
<script type="81-text/javascript">window.yaContextCb.
push(()=>{
Ya. Context.AdvManager.render({
renderTo: 'yandex_rtb_R-A',
blockId: 'R-A-1281369-4'
})
})</script>
</div>

<div class="versusia6">
<span class="line" id="L29"><i>Как шумный ветр пустынь;</
i> то ходишь по утрам</span><br>
```

```
<span class="line" id="L30">С собакой и ружьем — и с пти-
цами воюешь;</span><br>
<span class="line" id="L31">То, сидя на холме, прелестный
вид рисуешь!</span><br>
```

При сканировании и распознавании печатных источников исследователь обязательно сталкивается с ошибками распознавания символов. Несмотря на то что есть типы текстов, в которых такие ошибки сводятся к минимуму (например, это современные стандартизованные книги), много проблем остается при оцифровке старых печатных источников. Например, программы для OCR плохо понимают дореволюционную орфографию и использовавшиеся в XVIII и XIX веках шрифты.



427

III.

Г. ДЕЛА КОНДАМИНА

РѢЧЬ

О ПРИВИВАНІИ ВОСПЫ

читанная имѣ самимѣ при публичномѣ собраніи вѣ Королевской Парижской Академіи Наукѣ, Апрѣля 24 числа 1754 года.

Ужасная и свирѣпая болѣзнь, коей сѣмена носимѣ мы вѣ крови, истребляетъ, извлеченіемъ, и безобразитъ четвертую часть человѣческаго рода. Кара будучи старагосвѣта, опустошила она новой болѣе, нежели мечь завоевателей онаго. Она оружіе смерти поражающее, не разбирая лѣтъ, полу, чина и обиталища. Мало семей избѣгаютъ горестной дани требуемой ею. Наипаче она вѣ городахъ и при дворахъ самыхъ великолѣпныхъ чинитъ опустошенія. (a) Чѣмъ выше и драгоцѣнѣе главы, коимъ она

(a) Особливое либо состояніе воздуха, либо разность кушанія, или что другое причиною, что бываетъ обыкновенно оспа гораздо опаснѣе по городамъ, а особливо у взрослыхъ людей и дѣтей нѣжно воспитанныхъ.

оз. въ
Т.
г͡. АблабовЛА ОЛНА
вфво
о прлолоанїи воєпы-
сттанаа. имъ. бамимъ. при полни.
номъ. оованїи. королєєско а.
Елжєкои. авадемї нокъ, ап͡ръ.
лаоа. тгла. ечва года.
Ожаєнаа в͡євиеѢпаа. болъзнь, коєв. ф исна
и осимъ ны ъсрови; и стрьолаєтъ. изф.
vвфвиваєтъ, и обоооразитъ. пстьє отоw пасть
пбло бъ͡вє гваго. роза. Кава. оаоупи. ставаго
вф та, о побтотила она. но сок.оолъє, и єже.
ли. иечь. завоєватьлев. онаго. Вна. ооужїє
биєрти. поражавшее, и є͡ разоираа. лѢтъ, полоу, чина. и овталиша. Мало. є͡є мвв и в.

Upload another image
Copy to clipboard

Анализировать насыщенные ошибками тексты так, как если бы это были исправные тексты, некорректно. Некоторые специалисты высказывали идеи, которые бы позволяли работать с плохо распознанными текстами[1], но пока что это направление остается в области теоретических рассуждений.

Ошибки нужно исправлять, и делать это приходится вручную. Здесь нет хороших программных решений, хотя перспективным представляется спелл-чекер для текстов в старой орфографии,

[1] Jiang M. et al. Impact of OCR quality on BERT embeddings in the domain classification of book excerpts // Proceedings http://ceur-ws. org ISSN. 2021. Т. 1613. С. 0073.

разработанный в рамках учебного проекта в школе лингвистики НИУ ВШЭ[1]. Вычитка распознанных текстов — самая затратная (человеческий труд обычно стоит больших денег, чем машинное время) и нетворческая часть исследовательской работы.

В текстах из интернета мы можем столкнуться с тем, что вместо привычных нам символов обнаружим т.н. HTML-entities (например, « вместо «). В чем опасность того, что исследователь не позаботится об удалении HTML-entities? Релевантное слово склеится со словом «laquo», и мы неправильно посчитаем частотность первого. Это может повлиять на наши выводы. Допустим, сразу после слова «map» (‘карта’) в тексте будет следовать закрывающая кавычка: *map«*. Наш код предобработки удалит знаки амперсанда и точки с запятой, получится квазислово *maplaquo*. Таким образом, подсчитывая частотность слова «map», мы упустим его вхождение в связке с HTML-entity ««».

В процессе предобработки необходимо превратить HTML-entities в соответствующие им символы. В Python это можно сделать следующим способом:

```
import html
html_decoded = html.unescape(html_string)
```

Кроме того, на странице как минимум два раза представлено название текста, имя автора: первый раз они отражены в части, которая называется header, например, в теге <title>, содержимое которого выводится в заголовке окна браузера; второй раз в части страницы, которая называется <body> и видима для пользователя в рабочей области окна. Это тоже может повлиять на последующее вычисление частотности слов. Если мы не удалим навигационные элементы сайта, то самым частотным словом у нас может стать слово «главный»: на веб-страницах обычно есть ссылка, ведущая на *главную* страницу сайта.

Для исследователя полезно, чтобы текст, с которым он работал, был структурированным. Как и в случае с обвязкой сайта, важно разделять в тексте основное его тело и заголовок. Это позволит в дальнейшем не путаться в распределениях лексики и не допускать перекоса при составлении частотного словаря. Например, известно, что название романа Стендаля «Красное и черное» никак

[1] https://github.com/dhhse/Otechestvennie_zapiski/tree/master/kate_data

не подкрепляется в самом тексте упоминанием этих цветов[1]. Понятия, стоящие за этими цветами, важны, но сами цвета при этом выражены именно в заглавии. Существуют и исследования, опирающиеся при этом только на данные анализа заглавий без обращения к тексту романов[2]. Если текст большого романа будет разбит на главы, то исследователь получит возможность выяснить, чем эти главы с точки зрения статистики данных отличаются друг от друга.

Такие исследования могут проводиться не только на романах. В статье «Почему ошибался Жуковский»[3] речь идет о метрических «сбоях» в переводе «Одиссеи», то есть о таких строках, в которых поэт использовал семь или пять стоп вместо требуемых шести. Такие строки встречаются только в нескольких песнях поэмы, и эти песни противопоставлены «безошибочным». Благодаря структуризованности текста, выделенности границ песен, оказалось возможным провести все нужные подсчеты и сделать вывод о том, что метрические аномалии появляются главным образом в «морских» песнях поэмы и отсутствуют в «сухопутных».

Но для предобработки текста существенны и более мелкие членения. Так, для некоторых типов аналитических операций важно бывает разбить текст на отдельные предложения. Предложение — это минимальный смысловой контекст, опираясь на который мы можем смоделировать значение слова.

Не всегда ясно, как простым способом произвести это членение. Вроде бы у нас есть в тексте для этого хороший маркер — знак конца предложения. Это точка, а в некоторых случаях — восклицательный или вопросительный знак. Но есть и множество случаев, когда все не так просто. Например, та же точка используется и для сокращения, после инициалов. Слово «Вячеслав» сокращается до «Вяч.» и если бы машина решила, что точка всегда разделяет предложения, то граница предложений прошла бы здесь между именем и отчеством человека: «Заглядывает “в башню” Вяч. Иванова, когда там водят “хороводы”

[1] Лотман Ю. М. Несколько слов к проблеме «Стендаль и Стерн» (Почему Стендаль назвал свой роман «Красное и черное»?) // Лотман Ю. М. Избранные статьи: в 3 т. Т. 3. Таллинн, 1993. С. 428–429.

[2] Моретти Ф. Корпорация стиля: размышления о 7 тысячах заглавий (британские романы 1740–1850) // Моретти Ф. Дальнее чтение / пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Изд-во Института Гайдара, 2016. С. 248–287.

[3] Орехов Б. В. Почему ошибался Жуковский: о внутритекстовых причинах метрических сбоев в «Одиссее» // М. Л. Гаспарову — стиховеду. In Memoriam / сост. М. В. Акимова, М. Г. Тарлинская. М.: Языки славянской культуры, 2017. С. 73–89.

и поют вакхические песни, в хламидах и венках» [З. Н. Гиппиус. Задумчивый странник (о Розанове) (1923)].

Инициалы — это не единственный случай такого рода трудностей. Вообще-то в русской типографике не принято ставить точку после сокращения «миллион», но в реальных текстах (особенно из интернета) мы, разумеется, встретимся с тем, что точка в этом месте ставится. Поэтому разделение на предложения должно быть устроено более умным способом, и для этого есть готовые программные решения. О них см. ниже.

Разбивать тексты нужно и на более мелкие единицы, например, «токены». Токен — это и слово, и знак препинания. Например, слово «так» в некоторых контекстах существует вместе со следующей за ним пунктуацией:

Так!.. Но, прощаясь с римской славой,
С Капитолийской высоты
Во всем величье видел ты
Закат звезды ее кровавый!..

Важно разделить их и представить отдельно, потому что в противном случае у нас будет две единицы для подсчета — слово «так» с «прилипшей» к нему пунктуацией и слово «так» без пунктуации. Нас почти наверняка будет интересовать объединенная частотность слова «так» в обоих этих случаях, поэтому пунктуацию нужно отрезать.

В Python есть переменная, содержащая почти все нужные нам небуквенные символы, которые имеет смысл отрезать от слов.

```
from string import punctuation, digits
punct = punctuation + '""--…“”\n\t’ + digits
```

Если мы работаем не с русским языком, а с какими-нибудь нестандартными для европейского культурного пространства письменностями, то ситуация еще сложнее. Например, китайская или тайская графика вообще не подразумевают деления на слова с помощью пробелов. Так выглядит первая строфа перевода на китайский язык стихотворения Ф. И. Тютчева «Silentium!»:

别声张，要好好地收起
自己的感情，自己的向往；

任凭它们在心灵深处
默默地升起，悄悄地沉落，
像繁星，在夜空中
任你观赏，可别声张！

Носитель языка при чтении легко находит границы слов, а для компьютера это проблема.

Наконец, если в языке слова изменяются (например, по падежам — как в русском языке), значит, перед нами будет стоять задача лемматизации, то есть автоматического нахождения леммы, проще говоря — словарной формы слова. Скорее всего, частотный словарь, который мы хотим получить, это словарь именно лемм, а не конкретных словоформ, из которых состоит текст. Иначе говоря, нас будет интересовать не частотность отдельных форм «окнами, окна, окну», а частота всех их сразу.

Хороший инструмент для той самой умной сегментации текста на предложения (или, как еще говорят, сплиттинга, от слова *split* — разрезать) — это программная библиотека natasha[1], написанная на языке программирования Python. Сплиттингом ее функциональность не ограничивается, но нас сейчас интересует именно он.

```
from natasha import (
Segmenter,
Doc
)

segmenter = Segmenter()
doc = Doc(text)
doc.segment(segmenter)
```

Библиотека умеет разбивать текст на предложения и предложения на отдельные токены. Числовые значения у параметров «старт» и «стоп» — это номер символа, на котором начинается или заканчивается соответствующее предложение в тексте.

Структурирование текста позволяет получать информацию автоматически. Структуру фиксирует разметка, то есть система специальных указаний для компьютера на то, чем являются те или иные сегменты текста. Такая разметка может существовать, например,

[1] https://github.com/natasha/natasha

в виде XML-тегов. Так структурирован текст стихотворения в поэтическом корпусе НКРЯ:

```
<?xml version="1.0" encoding="utf-8"?>
<html><head>
<title>Тютчев Ф.И. Какое дикое ущелье!.. (1835)</
title>
</head>
<body>
<p class="verse"><line meter="Я4ж"/>Какоѐ дѝкоѐ
<rhyme-zone/>ущѐлье!<br/>
<line meter="Я4м"/>Ко мнѐ навстрѐчу клю̀ч <rhyme-
zone/>бежѝт —<br/>
<line meter="Я4ж"/>Он в дол̀ спешѝт на <rhyme-
zone/>ново̀сѐлье, —<br/>
<line meter="Я4м"/>Я лѐзу ввѐрх, где ѐль <rhyme-
zone/>стоѝт.</p>
<p class="verse"><line meter="Я4ж"/>Вот взобра̀лся᷄
я на᷄ <rhyme-zone/>вершѝну,<br/>
<line meter="Я4м"/>Сижу᷄ здесь ра̀достѐн и <rhyme-
zone/>тѝх —<br/>
<line meter="Я4ж"/>Ты к лю̀дям, клю̀ч, спешѝшь в <rhyme-
zone/>долѝну, —<br/>
<line meter="Я4м"/>Попробу̀й, ка̀ково᷄ у <rhyme-
zone/>нѝх!</p>
<p class="date"><noindex>&lt;1835&gt;</noindex></p>
</body></html>
```

Здесь заглавие отделено от собственно текста. Это позволяет отдельно подсчитать частотность слов, входящих в поэтическую строку и оставшихся за ее пределами, связать конкретные лексемы и метр строки[1]. С помощью тега <p> отделены друг от друга строфы. Выделены и рифмующиеся слова: это тег <rhyme-zone>.

Если для наших аналитических целей мы составляем частотный словарь, то больше информации нам даст именно подсчет лемм, а не отдельных словоформ. Распределение по словоформам часто

[1] Подробнее см.: Orekhov B. Lexis Meets Meter: Attraction of Lexical Units in Russian Verse // CLLS2016. Computational Linguistics and Language Science. Proceedings of the Workshop on Computational Linguistics and Language Science. 2017. Vol-1886. P. 110–121.

имеет случайный характер, и информация о частотах словоформ «окнами» и «окнах» нам скорее всего ничего полезного не даст, а вот если мы будем знать частотность всех этих форм, то эта информация уже будет иметь системный характер для, например, тематики текста.

Если мы работаем с большим текстом или даже большой коллекцией текстов, то вручную для каждого «окнами» словарную форму не пропишем — это будет слишком большая и бессмысленная работа. Существуют программные решения для лемматизации текстов. В простых случаях они работают эффективно, обычно это частотные и регулярно образующие свои формы слова вроде «земля» (земли, земле, землей и под.), «дерево» (дерева, дереву, деревом) и подобных. Но есть и сложные случаи. Во-первых, это имена собственные, особенно пришедшие из других языков. Их облик затрудняет машине распознавание способа формоизменения. Во-вторых, это формы, которые теоретически могут восходить к разным леммам, и без контекста будет непонятно, к какой именно. Например, слово «мыла» может быть и глаголом «мыть» в форме прошедшего времени, и существительным «мыло» в родительном падеже. Только в контексте «мама мыла раму» мы поймем, что речь именно о глаголе. Такие трудности называются грамматической омонимией. Некоторые программы умеют «смотреть» на контекст и реконструировать наиболее вероятную для такого контекста лемму. Такое умение называется «снятием омонимии».

Одно из лучших решений для автоматической лемматизации — программа под названием mystem[1]. Она умеет и восстанавливать словарную форму слова, и снимать омонимию. Программа доступна в виде бинарного исполняемого файла для разных операционных систем, но существует и обертка на языке Python[2].

Другой популярный инструмент для лемматизации — это библиотека pymorphy2[3], она написана на чистом языке Python, но не умеет учитывать контекст и снимать омонимию. В омонимичных случаях программа выдает весь возможный набор вариантов в порядке от наиболее частотного в языке к наименее частотному. Естественно, что исследователю может не повезти и тогда программа выдаст частотный, но неподходящий для данного контекста вариант.

---

[1] Зобнин А.И., Носырев Г.В. Морфологический анализатор MyStem 3.0 // Труды Института русского языка им. В.В. Виноградова. 2015. Т. 6. С. 300–310.

[2] https://pypi.org/project/pymystem3/

[3] https://pymorphy2.readthedocs.io/

Эти программы умеют не только восстанавливать словарную форму, но и делать морфологический разбор, то есть сообщать пользователю, к какой части речи принадлежит слово, в какой грамматической форме оно стоит (то есть в каком падеже или форме какого времени и лица).

Число таких инструментов множится. Лемматизировать текст можно и с помощью той же natasha, которая упоминалась выше.

```
for token in doc.tokens:
token.lemmatize(morph_vocab)
for _ in doc.tokens:
print(_.lemma)
```

Как мы уже говорили, программы неизбежно ошибаются. Имена собственные, которые похожи на какие-то косвенные формы русских слов, могут лемматизироваться, исходя из ложных оснований. Например, фамилия бывшего президента Франции Николя Саркози напомнила программе форму русского повелительного наклонения вроде «своди», «замени». От этих форм восстанавливается лемма «сводить», «заменить», а от «Саркози» — «Саркозить». Это немного напоминает языковую игру, в которой слово «крокодил» рассматривается как глагол: «крокодил, крокожу и буду крокодить». Американский политический деятель Дик Чейни рассматривается программой как родительный падеж от существительного «чейня», а сокращение «Изд.» (то есть «издательство») как родительный падеж множественного числа от слова «изда».

Частотность слова

Частотность слова — это ценный ресурс, с которым полезно работать и филологу, и тому, кто осуществляет цифровое исследование художественного текста. Полезность этого параметра доказывается и его использованием в разнообразных технологиях обработки текста, о которых мы скажем несколько позже. Частота слова — это тот самый «крючок», за который может зацепиться компьютер в подходах к «пониманию» информации, которая содержится в тексте.

Частотные словари — это по сути один из способов моделирования текста. Моделирование — это научное упрощение объекта,

попытка представить его в более общем виде, чем в реальном бытовании. Через частоту слов можно реконструировать и тематику текстов, и их стилистические особенности. Но при этом не следует трактовать частотный словарь слишком прямолинейно, то есть предлагать простые и наивные интерпретации частотных списков. Существует множество разнообразных факторов, которые влияют на частоту появления слова в тексте, но при этом плохо отражаются в словаре.

Существуют старые, выходившие еще в 1970-е, исследования частотности слов в поэзии, например, за авторством Гейра Хьетсо. Этот скандинавский филолог-русист стал особенно известен благодаря инициированному им количественному исследованию романа «Тихий Дон», позволившему, как казалось автору, утверждать, что авторство этого текста принадлежит Шолохову[1]. В том исследовании можно обнаружить заметное количество методологических ошибок, не позволяющих считать тему исчерпанной. Более подробно этот случай разбирается в более современной публикации, посвященной тому же предмету[2].

По данным Хьетсо[3], у Баратынского первые места в частотном словаре занимают слова «мятежный» и «счастливый». Как кажется интуитивно, ни то ни другое слово не отражают особенности поэтики Баратынского. У Тютчева на первое место попадает слово «великий», что тоже довольно далеко от главных для нас мотивов в творчестве этого поэта[4]. Еще раз подчеркнем, что к интерпретации частотности нужно подходить осторожно. Это важный способ описания текста, но далеко не универсальный.

Мы не всегда знаем, частотность отражает распространенность слова во всех текстах нашей коллекции или только внутри одного текста. Частотное слово может получить такой статус благодаря просто повторению, скажем, в рефрене одного поэтического текста. Тогда, конечно, это случайный «выброс», и служить характеристикой

---

[1] Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / пер. А. В. Ващенко, Н. С. Ноздриной. М.: Книга, 1989.

[2] Великанова Н. П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. Научно-просветительский общенациональный журнал. 2019. № 1. С. 70–82.

[3] Хетсо Г. Лексика стихотворений Лермонтова. Опыт количественного описания. Oslo, 1973. 44 с.

[4] Орехов Б. В. Принципы организации мотивной структуры в лирике Ф. И. Тютчева: автореф. дис. … канд. филол. наук. Воронеж, 2008.

всего набора текстов этот выброс не может. Поэтому было бы неправильно на основе таких подсчетов говорить, что частотность репрезентирует художественный мир того автора, количественное исследование текстов которого мы проводим.

На распределение частотностей слов влияет и жанр. Если, к примеру, мы обозреваем жанр писем, то частотными у нас окажутся слова, которые входят в традиционные формулы приветствия и прощания, повторяемые в письмах: «Привет», «Будь здоров», «Всего доброго». Это тоже не черта авторского стиля, а особенность жанра, заставляющая автора писать так, а не иначе. Аналогичным образом влияет на частотность и тематика текста, которая в большей степени будет отражать не авторскую стилистику, а законы построения художественного мира. На частотности отразиться то, на космическом корабле происходит действие романа или в провинциальном селе.

Важно помнить, что самыми частотными в тексте всегда будут служебные слова. Первые 10–20, а то и 30 слов в частотном словаре, ранжированном от большего к меньшему, это всегда предлоги, союзы, частицы[1]:

1. и
2. в
3. не
4. на
5. я
6. быть
7. он
8. с
9. что
10. а
11. по
12. это
13. она
14. этот
15. к

Причем это не особенность исключительно русского языка и не особенность какого-то конкретного автора, а универсальная закономерность.

[1] Ляшевская О.Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Чтобы данные в разных частотных словарях были сопоставимы, то есть чтобы их можно было сравнивать, частотность нужно измерять не только в абсолютных числах, но и в единицах IPM (instance per million). Абсолютное значение — это, например, 5, или 15, или 45 — сколько раз некоторое слово встретилось в нашем корпусе. IPM — это та же самая частотность в расчете на один миллион словоупотреблений. Это значит, что если длина нашего корпуса равна миллиону слов, то IPM для слова, встретившегося в таком корпусе один раз, будет 1, три раза — 3 и т.д. А если длина нашего корпуса два миллиона, то IPM для абсолютной встречаемости 1 будет 0,5, для абсолютной частотности 3 будет 1,5. Это позволит нам получить такие цифры, которые дадут возможность для сравнения — того же Тютчева с Баратынским или Лермонтовым, даже если объем их текстов будет меньше миллиона слов.

Важно помнить, что частотный словарь всегда подчиняется так называемому закону Ципфа. Закон Ципфа — это частный случай того, что в математике называется степенным распределением. Если грубо упростить, то степенное распределение означает, что частотность первых в списке слов будет очень высокой, но станет быстро падать. Примерно четверть всех слов из словаря будет употреблена только два раза, а половина всех слов будет иметь частотность 1. Важно помнить, что это не особенность какого-то одного текста или автора. Так будет всегда для текстов, созданных на естественном языке. Если озвученное выше правило с наиболее частотными служебными словами, или закон Ципфа, не выполняется, это значит, что с текстом или нашими подсчетами «что-то не так». Например, в некоторых языковых разделах Википедии есть не написанные человеком, а сгенерированные по определенному шаблону статьи. Чтобы их обнаружить, достаточно составить частотный словарь этой Википедии и увидеть там в десятке наиболее частотных слова вроде «река», «бассейн» и подобные[1]. Это будет значить, что Википедия состоит в том числе и из десятков тысяч единообразных статей о реках и других водных объектах. Это уже не вполне естественный язык и не вполне естественный набор текстов.

Один из способов посчитать частотность правильно — это соотнести ее с частотностью того же самого слова в других документах

[1] Орехов Б.В., Решетников К.Ю. Государственные языки России в Википедии: к вопросу о сетевой активности миноритарных языковых сообществ // Настройка языка: управление коммуникациями на постсоветском пространстве: коллективная монография. М.: Новое литературное обозрение, 2016. С. 263–281.

текстовой коллекции. Это важно для тех случаев, когда нам важно понять, что частотность информативна и ее можно привлекать к научному анализу. Речь идет о метрике, которая называется TF-IDF. Компьютерная лингвистика — это дисциплина, которая занимается подсчетами в текстах и строит на основе выявленных закономерностей технологии обработки больших массивов текстовой информации.

Одна из задач, которые решают компьютерные лингвисты, — это выделение в тексте ключевых слов. Эти слова способны представлять весь смысл документа в сжатом виде, они выделяют для нас в тексте главное. Их поиск можно алгоритмизировать, если принять, что они часто встречаются в некотором одном интересующем нас документе, но при этом редко во всех остальных. Такой подход кажется очень осмысленным.

Что стоит за этой метрикой? Если мы возьмем абстрактный текст про Францию, ключевыми словами будут сама Франция и Париж, потому что они встретятся в этом тексте, а в текстах про другие страны или вообще не про страны Франция и Париж встречаются реже. Но все-таки встретятся, потому что это известная страна и известная ее столица.

В статье о Франции из Википедии, разумеется, есть и слово «Франция», и слово «Париж».

Но в другой статье из Википедии, не про Францию, а про русского поэта Владислава Ходасевича, видно, что несмотря на то, что текст в целом посвящен другому предмету, слова «Франция» и «Париж» тоже присутствуют. Что нам нужно сделать, чтобы понять, является ли для данного текста слово «Франция» ключевым? Правда ли слово «Париж» выражает важную информацию из текста, с которым мы имеем дело?

Для вычисления того, насколько интересующие нас слова являются ключевыми, и существует метрика tf-idf. Если сказать упрощенно, то tf — это сокращение от term frequency, то есть это значение абсолютной частотности, с которой некоторое слово встретилось в тексте или документе. Текст и документ — это в терминах компьютерной лингвистики одно и то же.

Idf — вторая часть этого термина — это inverted document frequency. То есть насколько редко это слово встречается во всех остальных документах. С помощью этого параметра мы можем вычислить не просто частотность слова, но и понять, насколько эта частотность важна на фоне остальных текстов.

Нам совершенно не обязательно пытаться самостоятельно записать формулы вычисления метрики в коде, потому что существуют уже готовые библиотеки на Python, которые включают в себя функционал вычисления нужных нам значений. Приведем код для вычисления idf с помощью библиотеки sklearn.

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
texts[0]
```

```
'купаться в шторм запрещать . \n заплывать -- не возвращаться . \n волна накатный бревно \n расплющива
ть бедный артист ! \n но среди бешеный вал \n быть тихий волна -- / пасат ,\n как среди гром к
аблук \n стопа / неслышный / босой . \n ты от берег влечь \n не удалой бесшабашность ,\n а ужасать
расчет --\n в открытый море / безопасный . \n артист , над мировой волна \n ты носиться от жизнь
к смерть ,\n как ограниченный дуга \n латунный / сгорбленный / рейсфедер ! \n но слышать зоркий спин
а \n среди безвыходный сальто ,\n как зарождаться волна \n с протяжный имя -- пасат . \n « пасат , во
звращать волна , пасат ,\n запретный мой заплыв , но хлынуть тишина возврат ,\n я обожать вода --
/ но что она без земля ? / пустой ! \n я обожать свобода -- / но что она без любовь , / пас
ат ! \n нести я , пока носить , оставлять на берег , -- быть святой ! \n я вставать / и , пошатыва
ться , / ты поблагодарить ,\n но ты растворяться в море , / не поглядеть , / пасат … »\n'
```

```
пасат 0.5448942548714197
волна 0.3863655788838256
среди 0.18024593980104597
артист 0.16964603222642136
но 0.16143537093042773
обожать 0.13717149335466636
берег 0.13405707874582715
море 0.11356792593812406
ты 0.09958267454732944
поблагодарить 0.09081570914523662
пошатываться 0.09081570914523662
возврат 0.09081570914523662
хлынуть 0.09081570914523662
```

```
tfidf_vectorizer = TfidfVectorizer(
max_df=0.95, min_df=2, max_features=n_features
)
```

Здесь мы видим результат вычислений с помощью кода sklearn. Это одно из стихотворений Андрея Вознесенского[1]:

Купаться в шторм запрещено.
Заплывшему — не возвратиться.
Волны накатное бревно
расплющит бедного артиста!

[1] См. также: Орехов Б. В. Метрическое и лексическое разнообразие в стихах А. А. Вознесенского // Труды Института русского языка им. В. В. Виноградова. 2022. № 3. С. 50–58.

Слова уже лемматизированы, а значение idf ранжировано для корпуса его ранних произведений. Мы видим, что такие слова, как «пасат» и «волна» значимы именно для этого стихотворения, а в остальных текстах раннего периода эти слова или встречаются один раз, или не встречаются вовсе, что показывает их ключевой характер для текста, с которым мы работаем.

Совместная встречаемость слов

Мы разобрали случаи, когда частотный словарь способен подсказать аналитику что-то о том, как устроен текст, на котором этот частотный словарь был построен. Единицами частотного словаря по умолчанию являются слова. В то же время многие частотные слова попадают в этот класс случайно, и в целом интерпретация частотного списка требует аккуратности. Чтобы избежать традиционных ловушек, связанных с частотностью, компьютерная лингвистика разработала несколько более умных подходов, нивелирующих влияние случайных факторов на позицию в частотном списке.

Но еще более важным количественным источником информации о тексте является совместная встречаемость слов. К сожалению, и здесь мы не можем быть уверены, что данные о совместной встречаемости, с которой мы работаем, не случайны. В одном тексте могут оказаться какие угодно слова, и их сочетание не будет иметь никакого семантического наполнения. Поэтому и для количественной оценки неслучайности совместного появления лексических единиц в тексте есть компьютерно-лингвистические методики.

Слова, появляющиеся вместе в тексте не случайно, могут отражать тему, которой посвящен текст. Мы уже видели это на примере статей Википедии о Франции. Слова, появляющиеся вместе в тексте, могут отражать особенности авторского стиля, то есть сознательное или бессознательное стремление к такому словоупотреблению. Обычно стилистика особенно ярко отражается на так называемых n-gram'ах. Gram — это единица текста. В зависимости от задач исследователей «грамом» может быть и буква, и слово. Мы далее будем говорить именно о словах, хотя совершенно естественны и такие контексты, в которых под «грамом» говорящий будет понимать и более мелкую единицу письма. N — это буква,

означающая неопределенность, переменную. На место N можно подставить какое-то число, например двойку, и тогда речь будет идти о биграмме — комплексе из двух стоящих друг за другом слов. Но можно пойти дальше и работать с триграммой, тетраграммой, то есть последовательностью, состоящей из трех или четырех слов.

N-gram'ы важны для разных исследований текстов, они являют собой частный случай совместной встречаемости слов, такую разновидность, когда имеют в виду именно стоящие друг за другом слова, которые не разрываются дистантным расположением в тексте.

N-gram'ы — это не случайные совпадения; если одни и те же слова часто стоят рядом друг с другом, это отражает в том числе и авторскую стилистику.

Еще один важный в контексте этого разговора термин — коллокации, то есть такие n-gram'ы, которые встречаются вместе в тексте значимо часто по сравнению с их индивидуальными частотами. Кроме того, иногда говорят о том, что коллокации обязательно должны представлять собой грамматически связанные слова, а не просто случайные. То есть «золотой осени» — это коллокация, а «вода серебряный» уже нет.

Что значит, что слова встречаются вместе в тексте статистически значимо по сравнению с их индивидуальными частотами? Это значит, что мы сможем судить о том, случайно или не случайно слова оказались рядом, только сделав некоторые вычисления, в которых будут учтены частотности слов, входящих в возможную коллокацию, по отдельности. Одна из метрик, которая дает нам представление о том, с чем мы имеем дело, это PMI, то есть piecewise mutual information, «покомпконентная взаимная информация». Не стоит путать ее с PMI, то есть instance per million, единицей, в которых следует отражать частотность частотного словаря. О ней мы говорили раньше.

PMI показывает «необычность», «непредсказуемость» явления, состоящего в том, что происходит сразу несколько событий одновременно, и у каждого из этих событий по отдельности есть своя вероятность. Хитрость в том, чтобы оценить их совместную вероятность. PMI показывает, насколько сильна связь в слов в сочетании, вне зависимости от частности самих слов. То есть один и тот же показатель PMI могут иметь две комбинации, при этом одна из них составлена из низкочастных слов, а другая из высочастных. Тогда PMI показывает, насколько часто два слова объединяются вместе относительно их собственных частот. Коллокация с высоким

PMI — это редкое, неожиданное явление, коллокация с низким PMI — обычное, частое, предсказуемое, рядовое явление. Важно при этом принять во внимание частотности самих слов в анализируемом корпусе.

Как и в других случаях, нам не обязательно воспроизводить эту формулу самостоятельно, поскольку существуют готовые программные решения. PMI можно вычислить с помощью соответствующей функции внутри программного пакета для Python под названием NLTK, то есть natural language toolkit, «набор инструментов для естественного языка».

Вот результат вычисления PMI для произвольного текста:

4,486 Федор Достоевский
4,481 Иван Тургенев
4,481 многолетний растение
—
2,376 велосипед ветка
1,4875 задание колодец

Высокое значение PMI оказалось у слов, которые действительно часто встречаются вместе, поскольку являются именем и фамилией известных писателей. Еще один пример высокого PMI — это коллокация «многолетнее растение». Поскольку в тексте уже проведена лемматизация и словарной формой для прилагательного в русском языке является форма мужского рода, то в списке словосочетание выглядит несогласованным по роду. Но это просто результат препроцессинга текста.

Под чертой — ряд сочетаний слов с низким PMI. Это слова, совместное появление которых в тексте выглядит случайным и не сообщает полезной информации. Важно понимать, что значимость совместной встречаемости слов в тексте можно оценивать не на глазок, а с помощью конкретных вычислительных методик.

В области цифрового литературоведения совместная встречаемость слов (называемая лексическими комбинациями) стала основой для ряда исследований смоленской филологической школы[1].

[1] См., например: Павлова Л. В., Романова И. В. Лексические комбинации в «Кормчих звездах» Вячеслава Иванова (из опыта применения компьютерного комплекса «Гипертекстовый поиск слов-спутников в авторских текстах») // Новый филологический вестник. 2019. № 3 (50).

Тематическое моделирование

Одна из технологий, которые базируются на идее совместной встречаемости слов как информационного ресурса, называется тематическим моделированием (topic modeling). Тематическое моделирование применяется в информационном поиске, то есть встроено в поисковые машины, в анализе новостного потока и других практических приложениях компьютерной обработки текста.

Но тематическое моделирование хорошо вписывается и в концептуальную рамку digital humanities. Особенно удачно подходит для осмысления места и вклада этой технологии в цифровые гуманитарные науки понятие distant reading, которое на русский язык переведено как «дальнее чтение»[1]. Термин придуман американским литературоведом итальянского происхождения Франко Моретти. Моретти назвал так свою книгу, под обложкой которой собрал несколько важных для последующего развития digital humanities эссе. «Дальнее чтение» (или, как его еще можно передать по-русски: «отвлеченное чтение») — это понятие, не имеющее точного и строгого определения, то есть так называемый «зонтичный» термин, удобный для обозначения с его помощью разных подходов, которые все же имеют между собой что-то родственное. В данном случае то, что объединяет разные подходы, это оппозиция традиционному медленному чтению (по-английски — close reading). При медленном чтении в классическом литературоведении внимательное изучение текста совершается самим исследователем, который пытается интерпретировать объект своего интереса. Distant reading — это попытка изучать текст, не читая его. Как это возможно? Между текстом и исследователем появляется новая сущность, научная модель, создаваемая компьютером. Машина по определенной программе извлекает из текста важные для его изучения параметры и представляет их исследователю в сжатом виде. Этот вид можно, например, визуализировать, то есть отрисовать на графике, или представить в виде таблицы. Такими параметрами могут быть и частотные списки слов, и система взаимодействий между персонажами, и даже длина заглавия произведения. Именно с этой промежуточной моделью исследователь и работает, не имея необходимости заглядывать в исходный текст.

[1] Моретти Ф. Дальнее чтение / пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Изд-во Института Гайдара, 2016.

Лучший способ понять тематическое моделирование в контексте digital humanities — представить его в виде одного из подходов distant reading, поскольку тематическое моделирование не подразумевает чтения человеком текстов, к которым моделирование применяется.

Как можно догадаться благодаря названию, тематическое моделирование — это технология, которая позволяет выделять темы текста, то есть, использовав компьютер, узнать, о чем текст или текстовая коллекция. Правда, это не значит, что темы предстают в привычном нам со школы виде, где учителя предлагают нам написать сочинения на тему природы, родины или любви. Темы в рамках topic modeling — это ряды слов, которые заранее никак не озаглавлены. Машина, которая реализует технологию тематического моделирования, может представить нам список слов по типу «финансы, деньги, транш, вклад, счет, банк». И уже человек, а не сама машина может увидеть в этом списке объединяющее начало и присвоить ему ярлык, например «тема финансов».

Как получаются эти «темы» или, вернее сказать, списки слов? Фактически слова, образующие темы, — это слова, которые встречаются в текстах вместе. Так реализует себя в конкретной технологии концепция совместной встречаемости слов как информационного ресурса. Но в данном случае речь не идет о простых биграммах или даже коллокациях. Слова, объединяемые в темы, могут располагаться в тексте дистантно, и, чтобы определить их взаимное тематическое тяготение, используется специальный математический аппарат.

У нас есть два наиболее распространенных алгоритма со своими преимуществами и недостатками, которые обычно используются для тематического моделирования. Самый популярный алгоритм, который хорошо подходит для прозаических повествовательных текстов, это LDA (Latent Dirichlet allocation, Латентное размещение Дирихле). Второй используется реже, но хорошо себя показывает на поэтических текстах: NMF (Non-negative Matrix factorization, неотрицательное матричное разложение). Оба алгоритма подразумевают составление так называемой терм-документной матрицы, то есть, упрощенно говоря, таблицы, в которой прописано, в каком тексте какое слово и с какой частотой встретилось. Затем эта таблица подвергается математическим преобразованиям. Но, как и в прочих случаях, нам не нужно пытаться реализовать эти алгоритмы самостоятельно, так как они уже встроены в разнообразные программные библиотеки Python: sklearn и gensim.

В работе со стихами специалисты обычно говорят, что имеет смысл предпринимать максимально дробное членение текстов, желательно по одной строфе, а не анализировать стихотворение целиком[1].

У тематического моделирования уже есть несколько имеющихся применений в digital humanities. Одна из известных работ на эту тему — это анализ 45 тысяч стенограмм французского парламента, который начал работать в конце XVIII века[2]. Исследователей интересовало, какие темы в выступлениях депутатов парламента обладают, скажем так, «наибольшей живучестью». Иными словами, какие темы продолжают обсуждаться, а какие, наоборот, затихают и не получают продолжения в дискуссиях. Оказалось, что наиболее интересными для последующих обсуждений оказывались темы, которые первоначально поднимались депутатами от левых фракций. Темы, которые предлагались правыми депутатами, были не очень долговечными.

Еще один пример применения тематического моделирования — это анализ классической французской драмы в статье, опубликованной в журнале *Digital Humanities Quarterly*[3]. В этой работе показано, что тематика комедий и трагедий во французской драматургии очень сильно отличается. Если попробовать отрисовать анализировавшиеся тексты на графике, что и сделано в статье, то видно, что комедии, которые обозначены красными точками, находятся довольно далеко от трагедий, обозначенных синими точками. Видно также, что есть и немногочисленные тексты, которые близки по тематике и кластеру комедий, и кластеру трагедий, но в целом это все же сильно различающиеся классы. Нечто среднее представляют собой трагикомедии. Это как минимум свидетельствует о том, что тематическое моделирование хорошо отвечает заявленному предмету, то есть в самом деле позволяет увидеть тематическую разницу между жанрами.

---

[1] Haider T., & Eger S. (2019). Semantic change and emerging tropes in a large corpus of new high German poetry. In N. Tahmasebi, L. Borin, A. Jatowt, & Y. Xu (Eds.). Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change (pp. 216–222). Association for Computational Linguistics, 2019

[2] Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo Individuals, institutions, and innovation in the debates of the French Revolution // PNAS May 1, 2018 115 (18) 4607–4612.

[3] Schöch C. Topic modeling genre: An exploration of French classical and enlightenment drama // Digital Humanities Quarterly. Vol. 11. No. 2. 2017. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Можно проводить исследования с помощью тематического моделирования и на русскоязычном материале. При этом важно учитывать то, о чем мы говорили раньше: нужна правильная обработка текстов, лемматизация, избавление от знаков препинания. В данном случае лемматизация особенно важна, так как носителем темы текста является не конкретная словоформа в косвенном падеже, а именно лемма.

Пример применения тематического моделирования на русскоязычном материале — статья Лейбова и Орехова о поэтической топике Крыма[1]. Здесь исследуются многочисленные, доходящие до десятков тысяч, стихотворения, посвященные Крыму, написанные непрофессиональными поэтами, публикующимися на сайте stihi.ru. Прочесть все это текстовое богатство глазами было бы невозможно, а исследовательская задача состояла в том, чтобы узнать, с какими темами в сознании поэтов связывается образ Крыма. Помочь в решении этой задачи могло только тематическое моделирование, которое действительно открыло, что Крым — это прежде всего необычный пейзаж, сочетающий в себе летние и зимние элементы, своеобразная природа с морской доминантой и тема любви.

Примером работы тематического моделирования с художественной прозой могут быть статьи Т. Ю. Шерстиновой и соавторов[2].

Векторные модели

Одним из современных способов анализа текста является построение векторных моделей слов. Это технология, которая восходит к лингвистической идее дистрибутивной семантики (дистрибутивный от *distribution* — «распределение»).

Уже как минимум с 50-х годов XX века эта идея обсуждалась в фундаментальной науке. Она заключается в том, что слова получают свое значение только в контексте. Мы уже касались этой темы

[1] Лейбов Р. Г., Орехов Б. В. Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике // Шаги/Steps. Т. 8. 2022. № 2. С. 205–232.

[2] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic modelling with NMF vs. expert topic annotation: The case study of Russian fiction // Advances in computational intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020: Proceedings / Ed. by L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, F. A. Castro-Espinoza. Pt. 2. Cham: Springer, 2020. P. 134–151.

в связи с грамматической омонимией при лемматизации. Действительно, некоторые слова становятся понятными только в конкретной фразе, как форма «мыла» — существительное или глагол. Но это означает и обратное: контекст позволяет опознать семантику слова. Афористически это формулируется так: «Вы узнаете слово по той компании, в которой оно придет»[1].

Эта идея имеет несколько практических следствий. В частности, это должно означать, что слова, которые встречаются в похожих контекстах, имеют близкое значение. Вот два почти идентичных предложения: «я приду туда через несколько часов» и «я приду туда через несколько минут». Отличаются они только словами «час» и «минута», остальной контекст одинаков. Значит, и слова «час» и «минута» должны быть похожи по значению. И это действительно так, речь идет о двух единицах измерения времени. Правда, значения эти не идентичны.

Как мы можем рассчитать контексты слов, присутствующих в собрании текстов, представляющих для нас исследовательский интерес? Для таких вычислений строится таблица, в которой прописывается частота совместной встречаемости слов, то есть насколько часто слова попадают в один и тот же контекст.

Эти подсчеты позволяют представить контексты слов, а значит, и их семантику в виде вектора. Над векторами мы можем производить вычислительные операции — складывать их, вычитать один вектор из другого и т.д. Но главное — мы можем находить в векторном пространстве ближайшие векторы к данному. В практическом смысле это означает, что мы можем находить ближайшие по значению слова. При этом векторное пространство для слов многомерно, и число измерений зависит от числа слов, которые мы учитываем при анализе контекста.

Таким образом, представив семантику слова в виде математического объекта, мы можем вывести из этого несколько полезных следствий. Например, мы можем находить слова, близкие по значению, так называемые квазисинонимы. Это не настоящие синонимы, не то, что лексикографы помещают в синонимические словари. Хотя и настоящие синонимы среди квазисинонимов тоже могут присутствовать. Но квазисинонимами могут быть и антонимы — у антонимов часто очень близкие контексты словоупотребления. Вспомним также пример про «час» и «минуту». Это тоже не синонимы в строгом смысле этого термина.

[1] Firth J. R. Papers in Linguistics 1934–1951. London: Oxford, 1957. P. 11.

Правда, срабатывает такой поиск хорошо только тогда, когда слово частотно и когда мы подсчитали контексты для интересующего нас слова на достаточно большом корпусе. Ну и для более корректной модели мы должны лемматизировать тексты, поскольку носителем значения слова является все-таки лемма, а не отдельная словоформа. Иными словами, слово «окнами» не противопоставлено по значению слову «окном».

Еще раз подчеркнем, что речь идет о похожих с точки зрения векторных пространств словах, но не тождественных по значению.

Чем эта технология может быть интересна при работе с текстами на естественном языке, которые становятся объектами исследования в цифровых гуманитарных науках? Если построить такие векторные модели на контекстах, формируемых в цикле рассказов Конан-Дойла о Шерлоке Холмсе, а затем визуализировать векторное пространство, то видно, что Холмс и Ватсон из первого рассказа «Этюд в багровых тонах» не похожи на Холмса и Ватсона из других произведений цикла. Не похожи — значит, что эти имена попадают в отличные контексты. Объяснение очевидно: автор в начале пути еще нечетко представлял себе жанровые рамки и образы персонажей[1].

Кроме того, если провести аналогичное исследование для персонажей классических романов XIX века, то векторы для имен протагонистов этих романов оказываются близки друг к другу. Это означает, что авторы подбирают для описания своих протагонистов очень похожие слова, формируя близкие контексты. Точно такую же плотную группу векторов составляют имена персонажей, которые являются романтическим интересом главного героя.

С помощью векторов можно проводить и исследования поэтических текстов. Создав векторную модель на заранее заготовленной большой текстовой коллекции, можно затем оценить семантическую связность слов в нескольких жанрах. Можно взять, во-первых, статьи из Википедии, во-вторых, случайно сгененированные тексты, в-третьих, поэзию и прозу, написанные настоящими авторами.

Связность нескольких слов оценивается автором исследования[2] как значение их попарного сходства с точки зрения близости в векторном пространстве. Если упростить: слова, близкие по значению, будут связными, далекие — не связными. Оказалось, что Википедия

[1] Grayson S. et al. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings // AICS. 2016. С. 68–79.

[2] Herbelot A. The semantics of poetry: A distributional reading // Digital Scholarship in the Humanities. 2015. Т. 30. № 4. С. 516–531.

демонстрирует наиболее высокую семантическую связность текста, случайно сгенерированные произведения — наименьшую. А поэзия в этом ряду занимает срединное положение, демонстрируя определенную неожиданность появления слов в стихах.

На русском материале такие исследования тоже проводились[1]. Если построить две векторные модели: одну на прозаических русских текстах, а другую на стихотворных, а потом запросить у векторной модели для каждого слова его квазисинонимы, то некоторые слова в той и другой модели не будут иметь общих квазисинонимов. Это будет означать максимальную разницу в семантике слов внутри прозаического и поэтического словоупотребления. Иначе говоря, слова эти выглядят как похожие, но по реальной семантике будут очень сильно отличаться: земля, любовь, человек, час.

Квазисинонимы слова «земля» показывают, что в поэтическом контексте обычно актуализируются символические значения, касающиеся погребального обряда (отпевание, погост, привал, территория), а в прозаическом — конкретные, связанные с сельскохозяйственной деятельностью (грунт, семя, перегной, пашня, почва). Так поэзия и проза актуализируют разные смыслы, а векторные модели позволяют это установить.

Стилеметрия

Стилеметрия (или *стилометрия* — сейчас под влиянием англ. *stylometry*) — это субдисциплина цифровых гуманитарных наук, переводящая стилевое своеобразие текста в исчислимые параметры, позволяющие измерять и количественно сравнивать стиль разных авторов.

Проблематичным при этом является и само понятие стиля, и набор параметров, и способ их квантификации.

Под стилем в контексте цифровых исследований неявно подразумевается неповторимая авторская манера письма, являющаяся текстовой репрезентацией личности пишущего. Иными словами, каждый человек уникален благодаря своему особому опыту, взглядам, навыкам. Эта уникальность, по мнению тех, кто использует

[1] Орехов Б.В. Стихи и проза через призму дистрибутивной семантики // Острова любви БорФеда: сборник к 90-летию Бориса Федоровича Егорова. СПб.: Росток, 2016. С. 652–655.

стилеметрию, должна проявляться и в сочиняемых им текстах, причем не на уровне тематики, а на уровне собственно языковой материи: выбора слов, использования грамматики, устойчивых сочетаний. То есть стилистическая особенность человека не в том, что он, будучи ихтиологом, пишет про рыб, или, будучи кинорежиссером, пишет о кино. Особенность в том, что один чаще, чем другие, употребляет слово «либо», а другой — оборот «на самом деле». В литературе о стилеметрии такая «особенность» называется авторским сигналом.

Такие стилистические особенности порой заметны невооруженным глазом и позволяют отгадывать авторство текста или адресата пародии.

Например, для речевой манеры И. Бродского характерно использование слова «суть» в качестве глагола-связки, равнозначной «есть»:

> Призрак бродит бесцельно по Каунасу. Он
> *суть* твое прибавление к воздуху мысли
> обо мне,
> *суть* пространство в квадрате, а не
> энергичная проповедь лучших времен.

«Суть» действительно является устаревшей формой глагола «быть» в 3 л. мн.ч., но Бродский, противореча языковым правилам, использует ее и для ед.ч., как в приведенном выше примере.

Этот характерный параметр помогает узнать авторскую манеру Бродского, но бесполезен для абсолютного большинства остальных авторов. Какие параметры в таком случае представляются продуктивными для стилеметрии?

Идейный прорыв в этом направлении произошел в последней трети XIX века. Он был связан с потребностями атрибуции, то есть установления авторства — и живописных полотен, и текстов. В области живописи идею, сходную с той, которая потом окажется высказана и для вербальных произведений, сформулировал итальянский искусствовед Джованни Морелли: «необходимо научиться отличать подлинники от копий. Однако для этого, утверждал Морелли, не следует брать за основу, как это обычно делается, наиболее броские и потому воспроизводимые в первую очередь особенности полотен: устремленные к небу глаза персонажей Перуджино, улыбку персонажей Леонардо и так далее. Следует, наоборот, изучать самые второстепенные детали, наименее затронутые влиянием той школы,

к которой художник принадлежал: мочки ушей, ногти, форму пальцев рук и ног»[1]. То есть наиболее репрезентативными становятся детали, которые меньше всего контролируются волей художника, те, которые он рисует машинально, по привычке. Такое внимание к деталям, перенос на них фокус исследовательского внимания культуролог К. Гинзбург называет «уликовой парадигмой», сравнивая деталь Морелли с уликой в криминальном расследовании.

Не случайно, что примерно в это же время (то есть в том же интеллектуальном контексте) идеи, очень сходные с основными постулатами «уликовой парадигмы», высказывает Томас Менденхолл[2]. Для установления авторства текстов, приписываемых Шекспиру, американский ученый предлагает использовать такой параметр, как длина слова в тексте. Достоинство этой характеристики опять-таки в ее неконтролируемости: в самом деле, вряд ли можно в обычной ситуации (то есть не включая сюда ситуацию авангардного творчества в духе литературы формальных ограничений[3]) представить себе автора, который возьмется выстраивать свой текст, высчитывая, сколько в нем присутствует слов той или иной длины.

Вторым важным обстоятельством является то, что длину слова и соответствующее ему количество в тексте мы можем перевести в числовые параметры и сравнить для нескольких произведений. Так, сопоставляя проблемный текст Шекспира с показателями Ф. Бекона и К. Марло, Менденхолл приходит к выводу о полном соответствии числовых показателей Шекспира и Марло.

Идея использовать длину слова для установления авторства себя не оправдала, то есть была позднее экспериментально опровергнута, но стремление ориентироваться на неконтролируемые автором параметры текста оказалось правильным. Вокруг такого рода «улик» и строилась в дальнейшем стилеметрия.

Вместо длины слова в качестве подсчитываемого параметра, отождествляемого со стилем, исследователи пробовали разные варианты:

– частотности слов и словоформ;
– цепочки символов;

[1] Гинзбург К. Мифы — эмблемы — приметы: Морфология и история. Сборник статей. М.: Новое издательство, 2004. С. 190.

[2] Mendenhall T.C. The characteristic curves of composition // Science. 1887. № 214s. P. 237–246.

[3] Бонч-Осмоловская Т.Б. Введение в литературу формальных ограничений. Литература формы и игры от античности до наших дней. Самара: Издательский дом «Бахрах-М», 2009.

– частотность и распределение частей речи;
– частотность грамматических конструкций;
– стиховедческие параметры (для поэтических текстов);
– длины слов и предложений;
– знаки препинания (зависит от уровня редакторского вмешательства в текст, может отражать представление о расстановке знаков препинания публикатором, а не автором).

Но и сам подсчет этих параметров может проводиться поразному. Например, одна из самых известных отечественных методик стилеметрического определения авторства («авторский инвариант» Т.Г. и В.П. Фоменко) предполагала вычисление процента служебных слов от числа всех слов в тексте. Но то же самое присутствие служебных слов в тексте может подсчитываться и иначе, например, с усреднением на определенный текстовый отрывок, скажем, в 10 тыс. слов.

Существенно, что серьезная проверка работоспособности этого набора параметров стала возможна только в последние десятилетия благодаря современному уровню оцифрованности текстов и компьютерным технологиям. В XX веке тестирование возможностей стилеметрии было вынужденно ограниченным: вручную проверялись сравнительно небольшие выборки текстов.

Еще одной проблемой стилеметрии стало то, что методики в ее рамках часто создавались под конкретные задачи определения авторства того или иного текста. С одной стороны, широко известные проблемы авторства культурно значимых произведений (диалоги Платона, трагедии Шекспира, «Конек-горбунок» Ершова, «Тихий Дон» Шолохова, спорные научные тексты Бахтина) подстегивали интерес исследователей к стилеметрическим проблемам. С другой стороны, многие методики создавались ad hoc, то есть для специального случая, и имели целью, таким образом, подкрепить точку зрения автора на конкретную историческую проблему. Работает ли та же самая методика для другого случая, оставалось второстепенным вопросом.

Выгодно выделяющимся на этом фоне универсальным (то есть не привязанным к конкретной историко-литературной проблеме) способом определения авторства стала Delta Берроуза, статья о котором была опубликована в 2002 году[1]. Дж. Берроуз предложил действовать следующим образом. Для исследования нужно собрать

[1] Burrows J.F. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing 2002. 17(3): 267–287.

корпус текстов, в число которых входил бы текст, авторство которого было бы под вопросом, тексты, которые уверенно атрибутируются авторам, подозреваемым в том, что это они написали проблемный текст, а также некоторое количество текстов «фоновых» авторов той же эпохи. Заметим, что распространенной ошибкой многих стилеметрических исследований является привлечение произведений другой эпохи[1]. Стилеметрия чувствительна к изменению языка даже на коротких временных дистанциях. Берроуз поступил исследовательски чрезвычайно грамотно. Он выбрал текст, в авторстве которого никто не сомневается: «Потерянный Рай» Джона Мильтона. Правда, забавный факт в том, что в XVIII веке чрезвычайно недружелюбно настроенный к Мильтону интеллектуал Уильям Лоудер с помощью подлога пытался доказать, что «Потерянный рай» — это плагиат. Лоудер обвинял Мильтона в намеренных заимствованиях из сочинений авторов, писавших на латинском языке[2]. Другими текстами в исследовательском корпусе Берроуза стали другие произведения Мильтона и поэзия его современников.

Методика предполагала, что для расчетов задается некоторое количество наиболее частотных словоформ. Например, 100. Все дальнейшие операции производятся только для этих слов. Самыми частотными всегда оказываются служебные, а не полнозначные слова (см. об этом выше). Следствием этого обстоятельства является то, что Delta работает не с тематикой текста, которая выражается именами и глаголами, а с теми самыми неконтролируемыми параметрами.

Далее для каждого из выбранных слов в каждом из текстов корпуса вычисляется z-score — отношение разницы, взятой в процентах, от общего числа слов в тексте частотности слова в данном тексте и общей частотности слова по всему корпусу (то есть вычисленной для всех текстов выборки сразу, как если бы они вместе составляли один текст) к стандартному отклонению частотности слова по корпусу. Среднее арифметическое взятых по модулю разниц между z-score у двух сравниваемых текстов — это и есть значение стилистического расстояния между ними.

Эмпирически установлено, что это значение оказывается меньше у текстов, принадлежащих одному автору, и больше у текстов, принадлежащих разным авторам. В этом состоит основание для

[1] Например, Marina Iosifyan, Igor Vlasov And Quiet Flows the Don: the Sholokhov-Kryukov authorship debate // Digital Scholarship in the Humanities, Volume 35, Issue 2, June 2020, Pages 307–318.

[2] Уайтхед Дж. Серьезные забавы. М.: Книга, 1986. С. 112.

атрибуции. Действительно, подсчеты Берроуза подтвердили, что Мильтон — самый вероятный кандидат для авторства «Потерянного Рая», то есть методика прошла проверку.

Позднее выяснилось, что у метода, в целом достаточно надежного и работоспособного для разных языков, есть ряд ограничений.

Во-первых, сравниваемые тексты должны быть жанрово однородными. Сопоставлять вперемежку художественную, дневниковую, деловую прозу, стихи и драму с помощью Delta нецелесообразно, она не показывает осмысленных результатов.

Во-вторых, для надежного исследования нужны тексты не меньше 5 000 слов (а лучше 10 000 слов) каждый. На меньшем материале стилистически значимые статистики слов показывают случайные значения, поскольку не успевают стабилизироваться. Это, кстати, означает, что никакая методика определения авторства, основанная на статистике, не будет показательной для текстов небольшого объема.

Из-за недостаточности текстового материала исследователи, решающие задачи атрибуции, порой добавляют к ставшим традиционными подсчетам с помощью Delta дополнительные параметры, в частности, стиховедческие. Так, известный в филологической среде случай текстов, приписываемых поэту-декабристу Г. Батенькову, разбирается современными учеными с помощью добавления в анализируемые данные сведений о метрическом оформлении стихотворений[1]. В то же время благодаря масштабному исследованию, которое провел на материале поэтического корпуса НКРЯ Борис Орехов, известно, что стиховедческие параметры редко имеют в творчестве автора какой-то устойчивый тренд[2].

С начала 2000-х годов Delta (а также ее модификации вроде Delta Эдера) стала своего рода стандартом области цифровых гуманитарных исследований. Ее популярности способствовала не только вновь и вновь подтверждаемая надежность, но и низкий порог входа, обесепеченный программным пакетом Stylo для языка R[3].

---

[1] Шеля А., Плехач П., Зеленков Ю. Феномен Батенькова и проблема верификации авторства: многомерный статистический подход к нерешенному вопросу // Acta Slavica Estonica XI. Пушкинские чтения в Тарту. 2020. № 6. С. 131–165.

[2] Орехов Б. В. Микродиахрония стиховедческих параметров у русских поэтов // ВАПросы языкознания: Мегасборник наностатей. Сборник статей к юбилею В. А. Плунгяна / ред. А. А. Кибрик, Кс. П. Семенова, Д. В. Сичинава, С. Г. Татевосов, А. Ю. Урманчиева. М.: Буки Веди, 2020. С. 161–164.

[3] Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis // R Journal. 2016. 8(1): 107–121.

Значимыми достоинствами этого пакета стали бесплатный характер его распространения и наличие графического интерфейса (а не только интерфейса командной строки, который создает психологические сложности для гуманитариев).

Для того чтобы воспользоваться этим программным пакетом, нужно установить на компьютер интерпретатор языка R (можно добавить к нему дружественную к пользователю R Studio), в командной строке установить пакет stylo:

```
install.packages("stylo")
```

Сделать это достаточно один раз.

Тексты для исследования имеет смысл подготовить следующим образом:

– каждый текст поместить в отдельный файл в формате plain text (с расширением.txt);

– все тексты должны быть в одной кодировке (предпочтительно: UTF-8);

– название файлов стоит выдержать в шаблоне: ИмяАвтора_НазваниеТекста.txt, то есть имя автора и название разделить знаком подчеркивания, для одного автора имя следует написать единообразно, это поможет при визуализации результата;

– не следует допускать пробелов и специальных символов внутри имени файлов;

– все файлы исследовательского корпуса нужно поместить в директорию с именем corpus.

После того как эта подготовка будет завершена, в интерпретаторе R нужно выполнить код, привязывающий программный пакет stylo к сессии:

```
library(stylo)
```

Этот код, в отличие от операции установки пакета (выше), нужно выполнять после каждого перезапуска интерпретатора.

Следующим шагом нужно установить рабочую директорию. Рабочей должна стать директория на уровень выше, чем директория corpus с файлами исследовательского корпуса, то есть та папка, в которой лежит папка corpus:

```
setwd('тут должен быть путь к директории, содержащей директорию corpus')
```

Далее нужно вызвать нужную функцию и работать уже с графическим интерфейсом:

```
stylo()
```

Stylo способен рассчитывать значение Delta для пар текстов в исследовательском корпусе и представлять результаты в виде таблицы попарных расстояний. Если текстов много, такая таблица будет сложна для визуального восприятия и анализа. Поэтому разработчики в качестве результата по умолчанию выбрали визуализацию в виде дендрограммы кластеризации, на которой тексты являются листами, а мера их стилистического сходства определяется дальностью или близостью ветвей. Наиболее близкие стилистически тексты (скорее всего, принадлежащие перу одного автора) в норме должны находиться на соседних ветках. Если выдержать шаблон именования файлов (см. выше), то тексты, заявленные как тексты одного автора (содержащие один и тот же префикс до знака подчеркивания) в цветном варианте графика будут помечены одним цветом.

Первая публикация, содержащая результаты применения Delta на русском языке, состоялась в 2016 году[1]. Исследователи Д. Скоринкин и А. Бонч-Осмоловская подтвердили работоспособность метода для русскоязычного материала. С тех пор с помощью Delta применительно к русскому языку решено несколько научных проблем, самой известной из которых можно считать проблему авторства «Тихого Дона».

«Антишолоховская» версия происхождения текста «Тихого Дона» появилась еще в первой трети XX века и к настоящему моменту получила широкую популярность. В контексте наличия множества частных фактов, которые можно трактовать как подкрепляющие любую из конкурирующих позиций, количественная атрибуция является для этой проблемы модельным случаем. Гораздо более надежным аргументом могли бы быть документальные свидетельства, подтверждающие ту или иную точку зрения, но таких

[1] Скоринкин Д.А., Бонч-Осмоловская А.А. «Особые приметы» в речи художественных персонажей: количественный анализ диалогов в «Войне и мире» Л.Н. Толстого // Электронный научно-образовательный журнал «История». 2016. Т. 7. № 7 (51).

свидетельств не существует. Количественные исследования также не могут подвести черту под дискуссиями, зачастую инспирированными вненаучными интенциями. Применение же непроверенных и не заслуживших авторитета (и скорее всего просто не работающих) цифровых методик, как это случилось в практике коллектива зарубежных русистов[1], только дискредитирует сам по себе цифровой подход к атрибуции.

В этой ситуации особенно любопытны были бы результаты, которые можно было получить для проблемы авторства «Тихого Дона» с помощью надежной методики Delta. Такое исследование провели отечественные ученые Н. П. Великанова и Б. В. Орехов[2]. Проанализировав межтекстовые расстояния для отдельных томов «Тихого Дона», «Донских рассказов» Шолохова, текстов В. Севского и Ф. Крюкова как наиболее вероятных авторов романа, с точки зрения антишолоховедов, а также художественных произведений современников Шолохова, филологи пришли к выводам, что:

– все тома «Тихого Дона», вероятнее всего, написаны одним человеком (в разных работах антишолоховского направления это положение ставилось под сомнение);

– наиболее вероятным является то, что «Донские рассказы» и «Тихий Дон» написаны одним человеком (не все антишолоховеды готовы признать и за «Донскими рассказами» авторство Шолохова, так что здесь нужны аккуратные формулировки);

– ни Севский, ни Крюков не являются вероятными авторами романа «Тихий Дон»;

– нет разницы между текстологически корректной версией романа и опубликованной с ошибками: Delta оказалась нечувствительна к правке текста[3].

Стилеметрию иногда отождествляют с атрибуцией. Это неверно. Когда мы говорим о стилеметрии, речь идет именно об измерении текстовых параметров, которые могут быть сопоставлены со стилем.

---

[1] Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / пер. А. В. Ващенко, Н. С. Ноздриной. М.: Книга, 1989. 186 с.

[2] Великанова Н.П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. Научно-просветительский общенациональный журнал. 2019. № 1. С. 70–82.

[3] Данные для воспроизведения этого исследования опубликованы в Репозитории открытых данных по русской литературе и фольклору: Орехов, Борис, 2020, «Стилеметрические данные “Тихого Дона” и современной ему прозы», https://doi.org/10.31860/openlit-2020.05-R001, Репозиторий открытых данных по русской литературе и фольклору, V1.

Такое измерение может иметь прикладное значение для задач определения авторства, но спектр его применения шире и включает не только прикладные, но и академические задачи, формулируемые вокруг сопоставления текстов. В этом смысле стилеметрия наглядно показывает статус количественных исследований вообще: цифровые исследования не гарантируют точности (и, соответственно, не обеспечивают истинности выводов в задачах атрибуции). Они позволяют сравнивать объекты изучения. Без стилеметрии мы, оставаясь в научном поле, не можем сказать, насколько стилистически близок «Тихий Дон» Крюкову или Шолохову, а благодаря стилеметрическому подходу эта близость получает числовое выражение.

Следует заметить, что значимость атрибуции в случае художественных текстов сильно преувеличена. Центральным для культуры является текст, а не его автор. Именно благодаря тексту у читателя появляется интерес к автору, а не наоборот. Концепция смерти автора[1] добавляет оснований считать проблему авторства второстепенной. Дискуссионным является и вопрос о том, насколько один художественный текст может помочь при медленном чтении и интерпретации другого: художественный мир в каждом новом произведении конструируется автором заново. Поэтому предоставление об одном стихотворении методологически проблемно при переносе на другое стихотворение. Так что установление авторства некоторого произведения, вызывающего сомнения в своей атрибуции, не обязательно имеет научную ценность.

В то же время нельзя отрицать, что публику за пределами академического сообщества вопросы авторства живо интересуют. Ученым же представляется, что атрибуция — это наиболее естественный круг задач в гуманитарной сфере, где возможно применение количественных методов, поскольку предполагает вычислимый и проверяемый результат.

Однако гораздо более важной для гуманитарной науки является проблема понимания текста, его внутреннего устройства и механизмов взаимодействия с другими текстами внутри поля культуры.

Особую ситуацию как будто должен представлять философский текст. Философские произведения репрезентируют цельную философскую систему, освещая ее с разных сторон. От того, принадлежат ли автору философской системы те или иные слова, вроде бы зависит полнота нашего представления об этой системе, то есть

[1] Барт Р. Избранные работы: Семиотика. Поэтика. М., 1994 С. 384–391.

как раз задача атрибуции тут должна быть связана с проблемой понимания.

Но и здесь вопросы авторства находятся в подчиненном положении, как видно из следующей цитаты: «Диалог “Феаг”, который входит в платоновский корпус, но очевидно не принадлежит Платону, играет важную роль в развитии учения о божестве Сократа. Если в подлинных платоновских диалогах божество упоминается достаточно коротко, то в “Феаге” ему посвящена заключительная часть диалога»[1]. Центральным оказывается не вопрос об авторстве, а о принадлежности школе, исповедующей одну философскую традицию.

Одним из возможных направлений применения стилеметрии за пределами задач определения авторства оказывается исследование переводов. Известен эффект, при котором авторский сигнал при переводе проявляет себя ярче, чем сигнал переводчика. Иными словами, стилеметрически переводы текстов одного автора, выполненные разными переводчиками, объединяются в один кластер, а переводы разных авторов, выполненные одним переводчиком, нет.

На русском и английском материале этот эффект показан в статье о стилеметрии Набокова[2]. Б. В. Орехов с помощью Delta проанализировал, насколько похожи русскоязычные романы Набокова и русские переводы его англоязычных романов. Методика позволила увидеть два разных противопоставленных друг другу кластера, и даже переведенный Набоковым самостоятельно роман «Лолита» оказался в кластере переводных текстов и не смешался с русскоязычными. Кроме того, исследованию подвергся выполненный Набоковым перевод «Героя нашего времени» на английский язык. В сопоставлении с оригинальными романами Набокова и другими переводами Лермонтова на английский место набоковского перевода выявилось довольно определенно — в одном кластере с остальными вариантами англоязычного Лермонтова. Даже писатель с такой индивидуализированной авторской манерой, как Набоков, превращаясь в переводчика, приглушает свой сигнал в тексте, подчиняя его авторскому сигналу.

---

[1] Беликов Г. С. Речи Максима Тирского, посвященные божеству Сократа, в литературном и философском контексте I–II вв. н.э.: дис. … канд. философ. наук. М., 2020. С. 68.

[2] Орехов Б. В. Текст и перевод Владимира Набокова через призму стилеметрии // Новый филологический вестник. 2021. № 3. С. 200–213.

Еще одно исследование перевода с помощью стилеметрических инструментов выполнено на материале последней по времени попытки передать на русском языке «Илиаду» Гомера[1]. А. И. Любжин продолжил и завершил неоконченный перевод XVIII века, принадлежащий Е. Кострову. Исследовательская задача Б. В. Орехова состояла в том, чтобы сравнить эти переводы и измерить лежащую между ними стилистическую дистанцию. Работа показала, что дистанция эта значительна, и современному переводчику не удалось мимикрировать под стилистику поэта XVIII века, что может положительно сказаться на восприятии этой версии русского Гомера для современных читателей.

Кроме переводов, предметом стилеметрического исследования может быть текстовое оформление речи различных персонажей, масок, псевдонимных авторов. Так, в рамках диссертационного исследования, посвященного роману «Война и мир», Д. А. Скоринкин выявил индивидуальную стилистику речи персонажей[2]. В отдельной статье Д. А. Скоринкин и Б. В. Орехов[3] обнаружили, что Delta оказалась чувствительна к разнице в стиле т.н. гетеронимов, то есть вымышленных «авторов» с глубоко проработанной — в отличие от простых псевдонимов — биографией, которых создавал для подписи своих текстов португальский поэт Фернандо Пессоа. Анализ результатов, которые демонстрирует Delta, показывает, что стилеметрически тексты гетеронимов Р. Рейша, А. де Кампуша и А. Каэйра должны быть интерпретированы как тексты других людей, а не самого Пессоа.

[1] Орехов Б.В. «Илиада» Е. И. Кострова и «Илиада» А. И. Любжина: стилеметрический аспект // Аристей. 2020. Т. XXI. С. 282–296.

[2] Скоринкин Д. А. Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа «Война и мир» Л. Н. Толстого): дис. … канд. филол. наук. НИУ ВШЭ. М., 2018.

[3] Skorinkin D., Orekhov B. Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta // Digital Scholarship in the Humanities. 2023. Volume 38, Issue 3. P. 1247–1266.

Послесловие

Отдайте же человеку — человеческое, а вычислительной машине — машинное.

Норберт Винер

Даже самый беглый обзор цифровых гуманитарных наук позволяет понять, что компьютеры открыли перед учеными пространство больших возможностей. Слово «большой» в контексте разговоров о настоящем и будущем науки не случайно. В зависимости от угла зрения, наше время называют и эпохой больших данных, и эпохой больших языковых моделей. Большими они называются не только ради рекламной гиперболы, но и потому, что эти сущности больше, чем каждый человек в отдельности.

Традиционный образ ученого-гуманитария рисует нам одиночку, который, запершись в кабинете наедине с текстом, погружен в герменевтику или метафизику.

Большие данные и компьютерные методы их анализа, во-первых, возвращают гуманитария к эмпирическому материалу, позволяют ему нарисовать перед собой большую картину знания, масштаб которой раньше был недоступен; во-вторых, позволяют ученому преодолеть конечность собственных исследовательских возможностей; и, в-третьих, позволяют объединять в научном диалоге многих специалистов, становящихся участниками общего проекта.

Благодаря новой конфигурации научных коллективов, благодаря новым исследовательским вопросам и электронным копиям недоступных ранее объектов изучения меняется сам портрет гуманитария.

Если ни один человек не может прочесть за свою жизнь миллионы томов, то компьютер — может. Если ни один человек не может собрать в своем кабинете полную коллекцию археологических находок, то электронное хранилище — может. Содружество гуманитария и компьютера не только позволяет развернуть исследовательские направления, которые раньше представляли собой чистую фантазию,

но и заставляет эволюционировать стоящего в центре этой сферы субъекта.

Если еще несколько десятилетий назад имели право на публикацию работы, в которых были выполнены только какие-то подсчеты, то теперь количественные характеристики без интерпретаций публиковать неприлично. Считать умеют уже все. А вот объяснять результаты подсчетов, извлекать из них новое знание способны только специально подготовленные к таким задачам исследователи.

Цифровые исследования для гуманитария — это окно из его кабинета в большой мир, такой, который до появления компьютерных помощников трудно было охватить даже мысленным взором. Теперь охватить можно гораздо больше, но и научную оптику (тот самый «взор») нужно перенастраивать. Недаром возникают такие понятия цифровой гуманитарной оптики, как, например, «дальнее чтение» или «макроскоп». В нашей книге для этой перенастройки не всегда можно найти готовые инструкции, но вдумчивый читатель обязательно сможет нащупать направление, в котором ему следует двигаться.

Digital humanities все еще переживают процесс становления. Далеко не все проблемы решены, далеко не все пути найдены. Поэтому цифровые гуманитарные науки заинтересованы в привлечении в свои ряды свежих сил. Для этого нужно стараться рассказывать на всех языках потенциальным союзникам о возможностях и ограничениях доступных в рамках этой дисциплины методов, о достижениях и слабостях ведущихся исследований, о решаемых и малоперспективных научных вопросах. Именно такой диалог на русском языке мы и выстраиваем с читателем (и, как мы рассчитываем, в скором времени коллегой) на этих страницах. Книга на русском языке, но при этом она фиксирует мировые тренды и ссылки, которые в ней можно найти, чаще на иностранных языках. Этот принцип можно выразить с помощью англоязычного афоризма: think global, act local.

За последние годы уже сложился определенный «канон» компьютерных методов, который необходимо иметь в виду, обращаясь к машиночитаемым данным или рассчитывая увидеть смысл в оцифрованных коллекциях, превышающих возможности нашего физического восприятия. К такому «канону» можно отнести базы данных, компьютерный анализ текста, геоинформационный анализ, сетевой анализ данных, компьютерное моделирование. При этом в большинстве случаев эти методы — в исконном значении путь исследования, который каждый гуманитарий проходит по-своему,

собирая уникальное соотношение нужных подходов и умений, существенно обогащающих присущие исследователям навыки вдумчивого чтения, пристального наблюдения и интуитивной классификации. Чем более сложным станет инструментарий исследователя, тем сильнее он будет себе казаться. Но не стоит обольщаться, полагая, что методы заменят знания. Важно понимать, что цифровые методы и данные ни в коем случае не заменяют глубокого знания и понимания предмета исследования. Пути формализации и концептуализации могут быть разными, но главным остается приращение нового знания. Для того чтобы методы применить успешно, нужно хорошо знать предмет, который предполагается измерить, изучить, понять.

В книге мы не раз подчеркиваем, что цифровые гуманитарные исследования — это молодая и динамично развивающаяся область. В то же время мы наблюдаем, как пионеры этой сферы быстро становятся пенсионерами. Стало быть, сама дисциплина уже выходит из детского возраста. Ее состояние хорошо описывается цитатой из классического романа «Мельмот Скиталец» Ч. Р. Метьюрина, серьезно повлиявшего на «Евгения Онегина» и «Портрет Дориана Грея»: « …его нежное румяное лицо, стройная, словно точеная фигура и переливы его мягкого голоса пробуждали в людях тот смешанный интерес, с каким мы обычно наблюдаем, как в юноше сквозь еще отроческую незрелость пробиваются первые побеги силы, которым в будущем суждено вырасти и окрепнуть, и наполняли сердца родителей той ревнивой тревогой, с какой мы следим за погодой теплым, но сумрачным весенним утром: мы радуемся разлитому в небе спокойному сиянию зари, однако боимся, что еще до полудня лазурь его будет затянута тучами». И хотя до зрелости цифровым гуманитарным наукам еще далеко, сейчас она определенно находится в точке юношеской равновесности, в том моменте, когда еще рано подводить окончательные итоги, но уже нужно фиксировать промежуточные. Надеемся, что наша книга послужит и этому. Carpe diem, quam minimum credula postero.

Андрей Володин,
Борис Орехов

Орехов Борис Валерьевич — кандидат филологических наук, доцент Школы лингвистики факультета гуманитарных наук и ведущий научный сотрудник Международной лаборатории языковой конвергенции Национального исследовательского университета «Высшая школа экономики», старший научный сотрудник Лаборатории цифровых исследований литературы и фольклора Института русской литературы (Пушкинского дома) РАН.

Румянцев Максим Валерьевич — кандидат философских наук, ректор Сибирского федерального университета, научный руководитель лаборатории «Digital Humanities» Сибирского федерального университета.

Сметанин Андрей Владимирович — кандидат исторических наук, доцент историко-политологического факультета Пермского государственного национального исследовательского университета.