Анна Швец

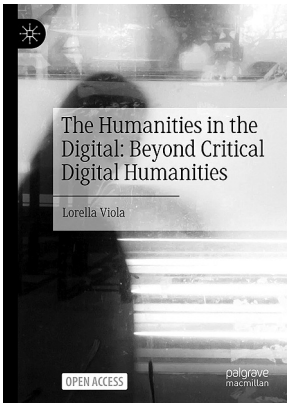
Эпистемология цифрового объекта

DOI: 10.53953/08696365_2023_183_5_337

Viola L. Humanities in the Digital: Beyond Critical Digital Humanities.

N.Y.; L.: Palgrave Macmillan, 2023. — XXVI, 173 p.

Тезис о кризисном состоянии гуманитарных наук встречается в англоязычной литературе так часто, что кажется уже данью конъюнктуре. Магистральный нарратив можно резюмировать так: гуманитарное знание как искусство интерпретации и контекстологической археологии теряет востребованность. Среди причин — запрос рынка труда на экспертную прогностику (вместо ретроспективного истолкования), заказ на производство инновационных продуктов на стыке больших данных и новых инженерных (цифровых) решений, неспособность гуманитарных наук полностью приспособиться к новым реалиям и недостаток бюджета у соответствующих институций. Нередко к этому добавляют: очертания кризиса особенно прояснились после исторической «точки невозврата» — пандемии коронавируса в 2020 г., когда все внимание сосредоточилось на статистических моделях, прогнозируемых и управляемых паттернах поведения множества людей и сервисах, смоделированных с учетом этих паттернов поведения.



С указанного тезиса начинается и книга Лорелы Виолы «Гуманитаристика в цифре: по ту сторону критико-ориентированных цифровых гуманитарных наук» — первая книга автора, заявка на дестабилизацию уже сложившегося исследовательского поля: попытка предложить новый взгляд на важную проблему, пусть и на локальном материале. Ее центральная тема — «производство знания в цифровой среде» (с. 29). Цифровая среда не является нейтральным контекстом в производстве знания, напротив: знание как таковое сейчас создается самим цифровым контекстом, поскольку именно он предопределяет специфику эпистемологического объекта. «Прозрачность, документируемость, воспроизводимость, предполагаемая надежность и объективность данных, их подлинность» (с. 9) — таковы качества эпистемологического объекта, которые ему, как предполагается, сообщает цифровая среда как пространство постановки вопроса и аналитической работы с данными.

Этот «цифровой поворот» противостоит укорененной в гуманитарных науках герменевтической традиции, в рамках которой источником знания о тексте становятся субъектность интерпретатора и его личное взаимодействие с текстом, зачастую не документируемое и не воспроизводимое. Разница между цифровыми (позитивистски точными, объективными) и «аналоговыми» (герменевтическими) гуманитарными науками поверяется демаркационным термином «данные» (data). Этот термин обозначает зафиксированные количественные показатели, описывающие тот или иной аспект объекта. Данные свободны от субъективности, считаются объективным доказательством (objective evidence). Данные принято анали-

зировать не по отдельности, а в рамках большого корпуса, чтобы опознать соотношения и корреляции и, опираясь на них, распознать отличительные черты референта данных.

В пример приводится громкий скандал 2015 г. вокруг компании «Cambridge Analytica», собравшей данные 87 миллионов пользователей социальной сети «Facebook»[1]: метки геолокации, лайки, время посещения социальной сети. Эти данные — потенциальная основа для «прогностических полицейских алгоритмов» (predictive policing), позволяющих, например, составить социальные профили и предсказать поведение потенциального преступника и его жертвы. Допустим, знание о том, что некто поставил много лайков под изображениями насилия, а затем оставил ряд геометок в оружейном магазине, является достаточным основанием для того, чтобы идентифицировать возможного нарушителя общественного порядка. Подобный вывод вроде бы очевиден, если собранные данные вписываются в ожидаемую закономерность; основная характеристика данных — подлинность, то есть неопосредованность субъективностью аналитика.

Схожим образом работает и методология «дальнего чтения», рассматривающая тексты как цифровые объекты. Если историк литературы, как иронически писал о нем Р. Якобсон, подобен «полицейскому, который, преследуя одного преступника, обыскивает и арестовывает всех подряд»[2], то представитель цифровой гуманитаристики способен разработать алгоритм поиска того самого преступника, и тоже не без прагматической пользы. Собираются объективные данные — содержащиеся в тексте языковые единицы и их сочетания, и «просвеченный» цифровым алгоритмом паттерн становится нередко самодостаточным свидетельством значения, лежащего на поверхности. Так, в недавней публикации интернет-издания «Системный Блокъ» был предложен проект сетевого анализа «Войны и мира», в котором система персонажей (упоминания их имен) была превращена в набор графов, и выяснилось, что ключевые персонажи (центральные узлы графов) — это представители двух сообществ, одно из которых дворянство (Безухов, Болконский), а другое — исторические участники военных действий (Кутузов, Наполеон). Графы показывают: «В центре сети, то есть в месте пересечения разных сообществ, оказываются Пьер, а также Наташа и Николай Ростовы»[3]. Если предиктивные полицейские алгоритмы позволяют установить личность шутера до теракта, то цифровое литературоведение — определить точные характеристики текста, разметить в нем статистически значимые и незначимые зоны и на основе этого, например, предложить модель маркетинговой реализации нового произведения.

Этот ход предсказуем — неожиданно то, что с этой отработанной линией аргументов Виола вступает в полемику. Вслед за известным медиаисториком Лизой Гительман она повторяет: «Сырые данные — оксюморон»[4]. Как и факты, данные фабрикуются — посредством интерпретативных актов отбора, упорядочивания и объединения в кластеры для анализа. За нейтральными и объективными корпусами стоят интерпретаторы — дизайнеры аналитических параметров их формиро-

---

1 Деятельность компании «Meta Platforms Inc.» по реализации продуктов — социальных сетей «Facebook» и «Instagram» запрещена на территории Российской Федерации Тверским районным судом 22.03.2022 г. по основаниям осуществления экстремистской деятельности. — *Прим. ред.*

2 *Якобсон Р.О.* Новейшая русская поэзия. Набросок первый. Прага: Политика, 1921. С. 5.

3 *Артемова Д.* 500 героев в одной схеме: о чем говорит сетевой анализ «Войны и мира» // Системный Блокъ. 2023. 15 июня (https://sysblok.ru/philology/500-geroev-v-odnoj-sheme-o-chem-govorit-setevoj-analiz-vojny-i-mira/).

4 Ср.: “Raw Data” Is an Oxymoron / Ed. by L. Gitelman. Cambridge, MA: MIT Press, 2013.

вания. Этих интерпретаторов Виола иронически сравнивает с шахматным автоматом Вольфганга фон Кемпелена (1769), механического устройства в виде восковой фигуры турка, которое могло обыграть даже сильных игроков; на самом деле внутри машины прятался гроссмейстер.

Более того, закономерности данных, определенные по массиву «великого Непрочитанного» (Ф. Моретти), сами по себе лишены смысла. «Компьютеры не запрограммированы находить смыслы (meanings), они запрограммированы находить закономерности (patterns); поскольку корреляции и закономерности сами по себе незначимы, по существу это означает, что базы данных дают нам акаузальный образ мира… современный [цифровой] мир предлагает нам бесчисленные закономерности, но не предлагает никаких объяснений, так что нам остается упорядоченный, но акаузальный способ познания действительности» (с. 90). Иначе говоря, функциональный смысл найденной закономерности может быть легко доступен, но ценностный смысл цифровой обсчет предоставить не в состоянии, это остается прерогативой человеческого сознания. Мы узнали, кто центральные персонажи романа, но у нас нет объяснения, что значит этот авторский выбор.

Как и в случае с шахматным автоматом, за данными скрыт невидимый игрок: человеческая субъектность и определяющие ее проявление контексты. Данные — это «абстракции, воображенные людьми»[5] и проявляющие их предубеждения, а значит, и контекстно-зависимые идеологические комплексы. В случае с потенциальными преступниками очевидно предубеждение: «Шутер тот, кто ходит в оружейный магазин», а в случае с центральными персонажами: «Самый часто упоминаемый персонаж в тексте и есть главный, а значит, его перспектива является ключевой для понимания романа». Подобные предубеждения позволяют маргинализировать определенные социальные группы (например, небелых индивидов, которые статистически чаще носят оружие, но не обязательно его применяют) и ряд аналитических ходов в отношении текста (например, анализ фигуры периферийного персонажа вроде Платона Каратаева).

Утопия объективного знания, производимого в «цифре», основана на квазиобъективных данных. Это, согласно Виоле, приводит к парадоксу междисциплинарности. Хотя междисциплинарный компонент является сегодня критерием успешности почти любой университетской программы или исследования, но на деле междисциплинарная работа выполняется в сотрудничестве предельно дисциплинарных компетенций. Инновации возникают на перекрестке дисциплин (например, литературоведения и теории графов) при работе с данными одной дисциплины методами другой, но мера убедительности инноваций определяется дисциплинарными критериями (в примере с графами персонажей — критериями литературоведения, поскольку производимое знание важно для литературоведов, а не для математиков).

Взаимодействие дисциплин оказывается поверхностным: сотрудничество происходит в условиях, когда каждая дисциплина преследует свою выгоду, а результатом становится усиление субспециализаций и дробление дисциплинарного поля на субдисциплины. В сотрудничестве математики и литературоведения решаются сугубо литературоведческие проблемы, и от литературоведения отпочковывается цифровое литературоведение, в результате чего мы получаем «насыщенное описание кейсов внутри одной дисциплины» (с. 28), осуществляемое с опорой на академическую конъюнктуру разделения влияния. Дробление на субдисциплины поддерживает академическую иерархию XIX в.: деление на точные, объективные

5 *Dobson J.E.* Critical Digital Humanities: The Search for a Methodology. Urbana Champaign: University of Illinois Press, 2019. P. 46.

науки и гуманитарные, не лишенные субъективности. Разве что намечается стратегический перекос баланса: гуманитарные науки признаются как менее склонные к производству позитивных, измеримых результатов, если только не произойдет междисциплинарное сотрудничество (на деле — дальнейшее дробление и прагматическая перенастройка дисциплины), и это де-факто возвращает нас к нарративу кризиса.

Виола предлагает мыслить производство знания в цифровой среде как «ощутимо недисциплинарное (и недисциплинируемое)» (там же), а цифровые объекты — как «объединение людей, структурных образований и процессов, соединенных друг с другом в соответствии с различными формами власти, которые укоренены в вычислительных техниках» (с. 32). Цифровые объекты — это не только объекты, ибо они не могут быть завершены, а еще и «пространство подвижных взаимодействий» людей и структурных единств, «никогда не нейтральное, включающее внешние, контекстно-обусловленные системы интерпретации и управления» (там же). Здесь угадывается концептуальный вектор акторно-сетевой теории (хотя Б. Латур в книге и не упоминается). Впрочем, автор не скрывает встроенность своего рассуждения в контекст постгуманистической теории с ее вниманием к сетям, где человек лишь один из множества равных акторов. В качестве подпорки используется определение материи, данное философом-постгуманистом Р. Брайдотти: материя — «сложный ассамбляж… сложносоставных единств, которые связаны со множеством силовых векторов, структурных образований и контактов» (с. 90).

Еще одна надстройка — постаутентичная парадигма (post-authentic framework), ставящая под вопрос подлинность основы цифрового объекта, то есть данных, якобы не опосредованных человеческим вмешательством. Человеческое вмешательство (интервенция) встроено в цифровой объект, если рассматривать его как историю контактов человека с технологическим приспособлением. Интервенция осуществляется в особом взаимодействии, петле обратной связи (feedback loop): человек как агент приспосабливается к технологической системе и вводит те данные, которые она может обработать; технологическая система выдает ответы; человек находит ошибки и уточняет запрос, и так уже система приспосабливается к человеку.

Цифровой объект определяется не как «всего лишь нематериальная копия оригинала», а как «не объект и не его репрезентация, а дистанция между ними»[6] (с. 39), или «одушевленные создания… которые создают определенные последствия» (с. 43). Важен не только и не столько сам цифровой объект, сколько инфраструктурная система отношений вокруг него. Объект и контекстуализирующая его коммуникативная инфраструктура мыслятся как состоящие в отношениях симбиоза (symbiosis) и обоюдности (mutualism). Симбиоз понимается как сосуществование, дающее взаимную выгоду, но это определение Виола дает обоюдности, симбиоз же трактуется как «постоянный пересмотр взаимодействий (в настоящем, прошлом и будущем) между системами, отношениями власти, инфраструктурами, актами кураторства и кураторами, программистами и разработчиками» (с. 45). Другими словами, симбиоз — это непрестанная ревизия отношений между разными агентами. Отношения симбиоза и обоюдности предполагают понимание знания как «подвижного», избавленного от соперничества дисциплин, но взаимовыгодного для них.

В этой связи первостепенным фокусом исследователя объявляются «отношения между отношениями» агентов и предметов (там же); эти реляционные сис-

6 Ср.: *Goriunova O.* The Digital Subject: People as Data as Persons // Theory Culture and Society. 2019. Vol. 36. No. 6. P. 124—145.

темы возникают на пересечении разных пространств, то есть в новом, гибридном пространстве (его предлагается называть трансверсальным — термин Р. Брайдотти и М. Фуллера[7]). Реляционная система отношений между отношениями обязывает исследователя документировать любое вмешательство в дизайн цифрового объекта (добавление новых данных, использование нового алгоритма и т.д.) как человеческую интервенцию, а интервенцию понимать как «сумму всех ранее и ныне принятых решений» всеми акторами. Комплекс «цифровой объект плюс инфраструктура» превращается в развернутый во времени коммуникативный палимпсест, незавершенный и незавершаемый; постоянный агон, где смыслы пересоздаются заново в симбиотическом взаимодействии человеческого актора и технологического инструмента.

В качестве иллюстрации подобного видения цифрового объекта используется оцифрованный (при участии Виолы) корпус эмигрантской прессы — выходивших в США с 1898 по 1936 г. италоязычных газет[8]. Этот корпус имел три версии; изначально целью проекта было «изучить, как… концепты путешествовали между Европой и США и как… результатом этих процессов стали транснациональные культурные и лингвистические контактные явления», выявив роль мигрантских сообществ как узлов культурного трансфера (с. 48). К третьей версии цель была уже такой: вычленить маркеры социокультурной идентичности и нарративы о становлении идентичности мигранта.

Если рассматривать этот корпус как коммуникативный палимпсест (корпус плюс инфраструктурные контексты), явным становится политико-прагматическое измерение исследования. Подобно тому как библиотеки оцифровывают книги исходя из внутреннего списка приоритетов (часто экономически обусловленных), выбор газет для оцифровки был обусловлен грантовой конъюнктурой. Первая версия проекта (ChronicItaly 1.0) должна была соответствовать условиям конкретного гранта, и был сформулирован критерий: отбирать только те источники, в которых содержатся данные о юридических постановлениях, политических решениях на уровне штата и о важных новостях локальных сообществ. Таким образом, источники были заранее профильтрованы через концептуальное сито грантовой повестки. Вторая версия (ChronicItaly 2.0) была разработана в Утрехтском университете; договор с нидерландским грантодателем подразумевал, что к уже собранному корпусу будут применены новые цифровые технологии разметки текста — с целью проявить «пространственное измерение» эмигрантской прессы, снабдив ее «географическими маркерами» и геометками. Третья версия (ChronicItaly 3.0) разрабатывалась уже в Люксембургском университете в рамках проекта «DeepteXTminER (DeXTER)». Задача, поставленная в переговорах с архитекторами DeXTER, предполагала использование технологий обработки естественного языка (natural language processing) и визуализацию полученных данных — так исследователи эмигрантского дискурса переключились на изучение маркеров идентичности.

Производство знания в цифровой среде неизбежно «переплетено с исследовательской повесткой отдельных институций» (с. 49), резюмирует Виола, подсвечивая симбиотический и обоюдный характер этой связи. Эпистемологическая выгода очевидна: грантодержатель нанимает ученых для освоения приоритетных направлений, а ученые решают интересные им задачи, сопрягая их с конъюнкту-

7 *Braidotti R., Fuller M.* The Posthumanities in an Era of Unexpected Consequences // Theory Culture and Society. 2019. Vol. 36. No. 6. P. 9.

8 См.: https://www.c2dh.uni.lu/de/data/chroniclitaly-and-chroniclitaly-20-digital-heritage-access-narratives-migration.

рой. Цифровой объект как история интеракций в череде контекстов предполагает вписанность определенных прагматических задач и предположений в собственную ценностную матрицу.

Большую часть книги составляет описание конкретных технологических интервенций и сопряженных с ними проблем. Это, во-первых, оптическое распознавание текста (optical character recognition, или OCR), то есть наложение текста на цифровое изображение. Во-вторых — распознавание имен объектов (named entity recognition, или NER), то есть создание геометок. В-третьих — анализ тональности текста (sentiment analysis, или SA), то есть контент-анализ, направленный на поиск эмоционально окрашенной лексики. Упоминается также тематическое моделирование (topic modelling, или TM), то есть репрезентация вероятностного распределения слов в текстах, модель того, как часто слова встречаются в тексте и насколько вероятно употребление определенных кластеров слов рядом (что и создает тему).

OCR — пожалуй, самая технически простая интервенция — проявляет «сложные взаимодействия между материальностью источника и цифровым объектом» (с. 65). Из-за пятен краски, клякс, стершегося изображения или низкого качества печати одна и та же буква может выглядеть по-разному и будет распознана как два разных символа, что приведет к возникновению нескольких версий одного и того же слова. Это обстоятельство значимо для процедуры токенизации (разбиения фразы, предложения, абзаца или всего текстового документа на более мелкие единицы, например отдельные слова, или термины — токены). Обработка текста может давать ряд разных токенов на одно и то же неверно распознанное слово, что требует дополнительной правки вручную. В итоге выясняется, что токен как вид данных — это не объективно данный в тексте феномен, а конструкт, возникающий во взаимодействии человека и инструмента.

Подобная опосредованность данных иллюстрируется на примере следующей интервенции — NER. Виола отмечает, что в 2019 г. при подготовке корпуса Chronic-Italy 3.0 посредством NER было выделено 547 667 упоминаний мест, однако 25 713 из них нуждались в ручной корректировке, что потребовало сочетания «экспертных знаний и технических возможностей» (с. 67). Ошибки состояли в том, что одно место из-за погрешностей сканирования распознавалось как два разных, некоторые имена собственные (Пятница) опознавались как названия места, а некоторые места не опознавались как таковые (Нью-Йорк). В итоге категории отбора должны быть скорректированы и откалиброваны концептуально, что уже являлось интервенцией на уровне смыслов, закладываемых в проект.

Схожая проблема — необходимость калибровки смыслов — наблюдалась и в случае с SA. Предъявляемая к этой интервенции претензия состоит в том, что «алгоритмы SA не обладают достаточными фоновыми знаниями местных социальных и политических контекстов», что создает «трудности выявления и интерпретации каламбуров, игры слов» (с. 71). Иначе говоря, алгоритмы поиска эмоционально окрашенной лексики не всегда способны заметить нетипичное эмоционально окрашенное слово в ряде контекстов со специфической прагматикой (например, выражение иронического отношения). Эти алгоритмы «дают скромные результаты на основе публицистических, экпрессивных текстов (opinionated texts)» (там же), частых в прессе. Однако даже когда SA находит эмоционально окрашенные слова, очевидна проблема с оценкой эмоционального контекста. Признав текст эмоционально заряженным, мы можем попытаться определить эмоции как позитивные или негативные, и алгоритмы оценки будут отражать наш ценностный выбор. Субъективность аналитика текста можно попытаться снизить путем расчета коэффициента «соглашения между комментаторами текста»

(inter-annotator agreement, с. 73), но нельзя изъять субъективность из акта анализа целиком: каждый раз коэффициент будет разным в силу разности восприятия комментаторов. Это обстоятельство ставит под вопрос мифологему воспроизводимости результатов исследования, проделанного при помощи использования цифровых технологий.

ТМ как интервенция проявляет нетождественность смысла и статистически установленной корреляции данных. Причина этого в том, что смысл присущ континуальным системам, а корреляция — дискретным. Дискретные системы состоят из изолированных и обособленных элементов, а континуальные — из элементов, неотделимых друг от друга в силу ряда взаимосвязей между ними. Язык — континуальная система, потому что слова связаны в рамках высказывания, а смысл высказывания спаян с прагматикой контекста. При этом, будучи подвергнут ТМ, язык предстает как дискретная система, то есть система, состоящая из отдельных элементов, слов.

В рамках континуальной системы возможно зафиксировать причинно-следственную связь (такая связь предполагает, что «каждое событие имеет уникальную, предшествующую ему причину», с. 86). В континуальной системе языка мы можем возвести смысл к контексту (игру слов — к иронической интенции и т.д.). В дискретной системе языка, где поиск производится по отдельному слову-токену, игнорируются многообразие контекстов и их прагматика ради поиска статистически повторяемых паттернов. Одно и то же слово может восприниматься поразному в разных контекстах, но это будет один и тот же токен (и игра слов, и обычное употребление будут распознаваться как одна единица). Более того, между двумя токенами в тексте можно установить статистически значимую связь, но это не всегда может иметь смысл; скорее можно сказать, что эти слова одновременно появляются рядом. Наконец, ТМ обнаруживает явную уязвимость в работе с нечастотными, редкими словами: вероятность их повтора в тексте ниже, а нередко такое неожиданное слово в ожидаемом контексте «может указывать на непредвиденный семантический сдвиг… что говорит о языковых изменениях» (с. 100).

Учитывая сказанное, ТМ признается полезной техникой, но работа с ней сравнивается с «гаданием на чайных листьях» (с. 101). Вычленять в тексте слова, которые с наибольшей вероятностью будут встречаться рядом, все равно что угадывать смысл по очертаниям гущи: между ними есть совпадение, но не обязательно есть причинно-следственные отношения. ТМ предлагается использовать как «инструмент поиска новых интерпретаций (reading), но не инструмент проявления смысла (meaning)» (с. 102), сознавая ограничения, присущие этой интервенции.

Продуктивным решением этих проблем видится разработка специфического пользовательского интерфейса, делающего интервенции в корпус исходных данных возможными, а сконструированность данных — видимой. Процедуры отбора, селекции, обработки данных в этом интерфейсе прозрачны. Каждая интервенция предлагается пользователю отдельно, и ее возможные ограничения учитываются заранее. Пример такого интерфейса, основанного на корпусе итальяноязычной прессы, приводится в книге; желаемый эффект прозрачности достигается благодаря системе вопросов, которые предваряют собственно исследование и позволяют рефлексивно подойти к отбору данных и их обработке. Такой интерфейс создает «несколько возможных способов видеть одни и те же данные» (с. 128).

«…Только активное, осознанное участие в процессе создания данных, их отбора, применения алгоритмов и методов позволит человеку… увидеть и распознать аналитические предрассудки и разобраться с ними» (с. 121), — подытоживает

Виола. Для подобного критического восприятия производства знания в цифровой среде необходимо расстаться с «удобными и успокаивающими» мифологемами позитивистской объективности и точности исследований при помощи цифровых технологий. Продуктивно мыслить знание как «постоянно меняющееся, где различия не отвергаются, а приветствуются в соответствии с принципами симбиоза и обоюдности» (с. 138). Иначе говоря, нам предлагают согласиться с перспективой, в которой знание относительно, симбиотически зависимо от человеческих интервенций и прагматических координат, а создающие это знание результаты принципиально невоспроизводимы в каждом новом исследовании. Ощущение кризисной ситуации в гуманитарных науках это не снимает, а скорее предлагает смириться со сложившимся статусом-кво.