Александр Александрович Вильховенко

аспирант факультета социологии
Европейского университета в Санкт-Петербурге,
Санкт-Петербург, Россия;
e-mail: avilhovenko@eu.spb.ru

Методологические сценарии исследования академических текстов в социальных науках

УДК: 316.1
DOI: 10.24412/2079-0910-2025-1-240-261

В статье представлен обзор научных работ, в которых применяется линейка методов для анализа текстов научных статей и книг. В обзоре приводятся примеры из областей, в которых исследуются условия производства и распространения науки, что охватывает наукометрию, социологию науки, науку о науке (science of science), метанауку и другие направления. Актуальность обзора связана с развитием за последние десять лет новых методов анализа текстовых данных, которые стали применяться к большим массивам научной информации. С опорой на качественные и количественные ориентации методов анализа текстов автор выделяет вариативность в этих двух подходах на уровне выборки и конкретных инструментов, что проявляется при изучении таких объектов, как научные темы, парадигмы и каноны, концепты, методы, дисциплины и специализации. При исследовании канонов и концептов исследователи опираются на стратегию глубокого чтения, тогда как по отношению к остальным объектам чаще используются более обширные выборки статей, к которым возможно применение сетевых инструментов и методов компьютерной лингвистики.

***Ключевые слова*:** методы анализа текстовых данных, социология науки, сетевой анализ, кластеризация текстов, анализ лексики, методологические ландшафты, исследования науки, подходы к изучению сущностей, вычислительная социология.

Благодарность

Выражаю благодарность директору Центра институционального анализа науки и образования Катерине Сергеевне Губе за неоценимую помощь в систематизации знания.

Введение

Научные статьи и книги обычно рассматривают как главный результат научной работы, ведь именно в научных текстах содержится научное знание — итоговые высказывания ученых, претендующие на новизну. В этой статье будет представлен обзор научных работ, в которых применяется линейка методов для анализа научных статей и книг. Обзор охватывает эмпирические исследования, в которых главным объектом являются научные тексты из статей по социальным наукам, к анализу которых применяются различные методологические сценарии. Мы не привязываемся к одному конкретному направлению, так как интерес к исследованию научного знания и условий его производства сейчас объединяет различные области — наукометрию, социологию науки, науку о науке (science of science), метанауку и др. В этих областях научные тексты являются источником данных для постановки вопросов, направленных на понимание развития науки.

В начале 1950-х гг. появились первые работы, в которых применялись статистические методы для анализа текстов. Это классический период в развитии направления компьютерной лингвистики, когда исследователи обращали внимание на частоту слов и статистические закономерности в них. В середине 1980-х гг. с развитием компьютерных технологий и появлением методов обработки естественного языка[1] стали активно применяться методы машинного обучения для анализа и классификации текстов. Научный интерес сместился к автоматизированным методам анализа текстов. В конце XX — начале XXI в. наблюдается рост популярности методов анализа текстов на основе глубокого обучения [*Johri et al.*, 2021], позволяющих автоматически извлекать семантику из текстов и строить сложные модели для анализа содержания. Они позволяют дать ответы на вопросы, касающиеся структуры текста, используя при этом большие объемы данных. С распространением новых методов машинного обучения, вслед за вычислительными дисциплинами, проекты сферы социальных наук стали пересобираться. В новых технических условиях большие данные и инструменты для работы повлияли и на то, что методология начала догонять теорию, расширяя возможности для объяснения социальных паттернов, в том числе и условий производства научного знания [*Губа*, 2021; *Губа, Семенов*, 2012].

Анализ научных текстов фокусируется как на текстах статей или книг, так и на информации о контексте появления текста — метаданных статей (такую информацию также называют знанием о знании) [*Evans, Foster*, 2011; *McKenzie*, 1999]. К метаданным относятся год издания, авторство, страна публикации, цитирования, примечания, заголовки и названия, ключевые слова, рубрики, год публикации, место публикации и т. д. Все, что относится к такой информации, дает дополнительный контекст для формирования содержания текста. К примеру, по названию журнала можно определить его влиятельность в дисциплине, а имя автора может указать на научную репутацию исследователя, так же как и его аффилиация — работает ли он в престижном университете [*Губа, Семенов*, 2012]. Библиографическая информация стала систематизирующими данными не только для библиотекарей, но и для социальных ученых [*McKenzie*, 1999; *Evans, Foster*, 2011]. Однако с развитием инстру-

[1] В международном научном сообществе эти методы называются “Natural Language Processing” (NLP).

ментов компьютерных наук стали анализироваться сами тексты научных работ, что значительно расширило возможности исследователей.

Актуальность обзора связана с развитием за последние десять лет новых методов анализа текстовых данных [*Baden et al.*, 2022]. В предлагаемом тексте описывается разнообразие методов, затрагивающих как более классические качественные способы анализа, так и количественные, стремительно набирающие популярность в науках о человеке и обществе. Обзор актуален для читателей и исследователей, интересующихся анализом организации научного знания, текстов или применяющих подобные методы в эмпирической работе.

Методологические сценарии для анализа академических текстов

Последовательность приемов в анализе текстов можно представить в форме избираемого сценария. Его выбор зависит от параметров самого исследования: исследовательские вопросы, дизайн исследования, объекты интереса и сами данные. Методологию в социальных исследованиях принято делить на качественную и количественную. Каждая парадигма методов закрепила за собой определенные способы анализа данных. При этом их граница до конца не выявлена. Нередко встречаются дискуссии в кругах исследователей, как определить методологическую границу [*Hammersley*, 2017]. С развитием компьютерных технологий и машинного обучения появились новые ответвления методов, дающих возможность сделать содержательные выводы, похожие на качественные, но при этом воспроизведенные при помощи количественной оценки. Далее мы рассмотрим особенности применения качественных и количественных методов к анализу научных текстов, при этом наибольшее внимание будет уделено количественному подходу[2].

Качественная методология

Начать стоит с качественного анализа, так как он представляет собой меньшую вариативность, нежели количественные исследования. По отношению к академическим текстам в социальных науках качественный анализ базируется на нарративном понимании происходящего в тексте. Для глубинного погружения в содержание материала достаточно частая стратегия — качественная сплошная вычитка материала [*Twine*, 2020; *Wu et al.*, 2020; *Kuehn, Rohfling*, 2022; *Fain*, 2024]. Это нередко предпринимается для создания концептуальных обзоров.

Разные способы качественной работы с текстом можно классифицировать как подстратегии для качественного анализа. Первый класс — это сплошная вычитка источников и ее комбинация в единый описательный текст. Проходя по этому сценарию, исследователь отбирает источники и на основе прочитанного суммирует полученную информацию в последовательный литературный обзор [*Wu et al.*, 2022]. Это достаточно универсальный способ систематизировать специфичную область,

[2] Выделяется также отдельный тип исследований, в которых происходит совмещение этих двух перспектив. При объединении качественной и количественной перспективы образуются смешанные дизайны исследований. Такие совмещения позволяют достигать наиболее комплексного результата.

выбранную для описания. Как правило, в статьях такого жанра описывается генеалогия идей, которая служит описанием всей выбранной рамки исследования.

Вторая методика предполагает кодирование статей по интересующим признакам с использованием контент-анализа [*Kuehn et al.*, 2022]. Задавая исследовательский вопрос, ученый определяет, на что будет обращать внимание во время чтения. Представление результатов исследования, произведенного по такому сценарию, задействует дескриптивные техники, представляющие распределение кодов и выявляющие частоту упоминания объекта интереса в зависимости от критериев источника.

Количественная методология

Технологии значительно повлияли на производство и распространение научных текстов. Распространение и поиск научного знания стали менее трудозатратными и позволили увеличить взаимопроникновение научных полей [*Barry*, 2008]. С увеличением компьютерных мощностей и разработкой новых инструментов появилась возможность обрабатывать источники большого объема. Распространение технологий позволило исследователям не только облегчить доступ к материалам для чтения, но и значительно развить научные области, оказавшие ключевое воздействие на инструменты текстового анализа. Особенно заметные изменения произошли в компьютерных науках, лингвистике и информационных технологиях [*Macanovic*, 2022]. Пересечение дисциплин и их столкновения произвели на свет целые ветки средств анализа — когда математика и компьютерная наука пришли в лингвистику, образовалась вычислительная лингвистика.

Количественные методы чаще всего ассоциируются со счетными моделями и сложными алгоритмами, основанными на математике и программировании. Исследователь при этом зачастую имеет выбор между двумя способами обработки данных. Первый путь — готовые программы[3], основанные на сложных алгоритмах анализа данных, «обернутых» в готовый инструмент. Второй путь — инструменты, реализованные через пакеты для языков программирования. Оба способа являются легитимными, а выбор между готовым инструментом или языком программирования зависит от задач исследования.

Популярными базами выгрузки метаданных статей являются *Web of Science* и *Scopus*. Это большие агрегаторы с высокой степенью систематизации [*Mongeon*, 2016]. Специально под создаваемые этими базами цитирований массивы адаптированы готовые инструменты для быстрого распаковывания и анализа подобных форматов. Распространенный метод анализа и визуализации — это сетевой анализ. Библиометрические данные позволяют построить сети взаимосвязей текстов по разным измерениям. Графы могут показывать связи: авторов [*Ianchuk*, 2021], аффилиаций, ключевых слов, терминов [*Fuhse*, 2011]. Давно существующие *VOSviewer* [*Çolak*, 2023] или *Sci2* [*Siler*, 2013] хорошо справляются с быстрым анализом подобных выгрузок.

Альтернативой готовым инструментам является программирование. Оно открывает возможности построения детализированной методологии, с расширен-

[3] В среде цифровых исследований это может называться «инструменты без кода», то есть готовое программное обеспечение с полнофункциональным пользовательским интерфейсом без использования программирования.

ным контролем фильтрации данных и параметров алгоритмов, с помощью которых выполняется анализ. Стоит кратко обозреть, какие при выборе этой ветки инструментов появляются пробелы, ставящие под сомнение общую достоверность научных результатов, — в работе Бадена и соавторов выделено их три [*Baden et al.*, 2022]. Первый пробел заключается в приоритете технологических аспектов над проблемами достоверности, которые связаны с операционализацией социальных научных измерений. Во-вторых, существует несоответствие между направленностью анализа текстов на извлечение конкретного содержания и шаблонов на уровне документа и необходимостью исследователей социальных наук измерять множественное, часто сложное содержание в тексте. И третье, более техническое, — это доминирование инструментов английского языка [*Ibid.*]. Однако появляется все больше моделей, адаптированных для различных языков. Это проблема, которой занимается современная компьютерная лингвистика [*Kondratyuk, Straka*, 2019]. Перечисленные пороги можно преодолеть как с помощью технического совершенствования инструментов, так и с помощью профессиональной подготовки.

Общая перспектива использования текстового анализа имеет несколько ответвлений. Первое — использование классических методов статистики и машинного обучения для текстовых данных. К ним может относиться кластеризация — разделение данных по выявленным признакам. На основе сформированных корпусов ученые классифицируют статьи, журналы, книги в разные группы и пытаются объяснить сложившуюся группировку [*Parodi*, 2014; *Schwemer*, 2022; *Beckert, Suckert*, 2021; *Ginnerskov*, 2024]. Для присвоения групп используются привычные в текстовом анализе методы кластеризации и снижения размерности: методы опорных векторов (SVM) [*Rona-Tas et al.*, 2019], методы главных компонент (PCA) [*Beckert, Suckert*, 2021]. Анализируя образ будущего в социологии [*Beckert, Suckert*, 2021], авторы одного из подобных исследований применили метод главных компонент (PCA) и проанализировали кластеры текстов по методологическим и дисциплинарным измерениям. Так они пытались ответить на вопрос, как и при помощи каких инструментов социология анализирует будущее. Дистанция между компонентами позволила определить близость дисциплин посредством конструирования будущего в социальной науке.

Известны сценарии, когда кластеризация и классификация текстов идет дальше обзорных потребностей, стремящихся описать природу кластеров. К примеру, можно понять более глубокую разницу, посмотрев на употребляемые в кластерах слова. Существует несколько ветвей методов, взаимодействующих с пониманием разницы академических текстов. Первая явная ветка основана на выявлении общих тем в дисциплинах, книгах и научных областях. Исследователей интересует обобщенное содержание, и при помощи инструментов тематического моделирования они выявляют это напрямую [*Zougris*, 2019]. Главная идея таких методов — выявление групп слов, которые наиболее часто встречаются друг с другом в одном документе. На основе этой зависимости создается тема. Через наблюдение лексических наборов можно обозреть содержание интересующего массива с разных сторон. К примеру, в публикации, посвященной различиям британской и американской социологии, подобный анализ применялся, чтобы выявить структурную разницу между двумя национальными кластерами. Одним из инструментов был алгоритм латентно-семантического анализа (LSA), позволяющий просмотреть содержание в корпусах текстов этих двух национальных перспектив [*Zougris*, 2019].

Когда информация аккумулирована в сжатой теме, исследователям легче описать ее содержание.

Содержание языка может анализироваться не только тематическими моделями, но и методами выявления разницы лексики. Это принципиально другой метод, производящийся, например, с помощью регрессионных оценок. Такие модели опираются на идею дистанции между словами, шкалой измерения которых служат независимые переменные. Выбираемая регрессионная модель зависит от состава данных и допущений, необходимых для корректной работы алгоритма. Как иллюстрацию следует рассмотреть публикацию об анализе методологического консенсуса в социологии [*Schwemmer, Wieczorek*, 2020]. Считается, что социология — это наука с низким уровнем консенсуса [*Turner*, 2006, 2016]. Такая разобщенность происходит прежде всего в инструментах. Для того чтобы ее измерить, авторы применили пуассоновскую модель[4] для измерения дистанции текстов академических статей по лексическим единицам. Чтобы выявить разобщенность — нужно наблюдать расползающуюся дистанцию между словами. Из результатов видно, что она есть и характерная лексика явно наблюдается как для количественной методологии, так и для качественной. Измеряя дистанции по классам методологий, можно дополнительно получить информацию, какой журнал из выборки будет больше направлен на прием публикаций с применением количественной методологии, а какой — качественной [*Schwemmer, Wieczorek*, 2020]. Методы выявления лексического разнообразия могут не только служить метрикой для проведения сравнительного исследования отдаленности дисциплин, но и выявлять общее состояние когнитивного расширения науки [*Milojević*, 2015].

В качестве отдельного методологического направления необходимо выделить применение больших языковых моделей (Large Language Models), породившее качественный скачок в области использования технологий обработки текста для широкого спектра исследовательских задач. Языковые модели являются наиболее эффективной итерацией предшествующих инструментов компьютерного анализа текстов. Их преимущества строятся на больших базах данных для предобучения, которые убирают необходимость самостоятельно обучать алгоритмы под каждую узкую задачу. За сравнительно небольшой срок с начала роста популярности языковых моделей в социальных науках уже появились отдельные инструменты для аннотирования [*Weber, Reichardt*, 2023] и суммаризации [*Radha*, 2024] академических текстов. Уже на текущем этапе технологического развития существуют инструменты, позволяющие легко обрабатывать и классифицировать тексты из академических баз данных [*Arhiliuc et al.*, 2024; *Shen et al.*, 2023]. Этот процесс можно обозначить как автоматизацию аналитического процесса, что, несомненно, может вывести как социальные науки, так и другие дисциплины в целом на новый уровень [*Bail*, 2024; *Ziems et al.*, 2024].

В количественной и качественной методологической перспективе мы выявили большое разнообразие методов, направленное на разные объекты сущностей и исследовательских вопросов, исследуемых на примере академических текстов. Следующим этапом нужно понять другие компоненты исследования, использующиеся для заявленных исследовательских задач.

---

[4] Пуассоновская модель — счетная регрессия, позволяющая работать со шкальными или целочисленными значениями.

Стратегия выборки данных

Выборка данных определяется методологическим сценарием. В качественных исследованиях, основанных на вычитке текстов, превалирует стратегия ограниченной выборки, сфокусированной в основном на высокорейтинговых журналах с высокими требованиями и порогом вхождения, состоящих в первых двух квартилях [*Schwemer*, *Wieczorek*, 2020; *Beckert, Suckert*, 2021; *Twine*, 2020; *Ferragina, Deeming*, 2023]. Эта выборка имеет свою логику: отбираются тексты, прошедшие многоступенчатую редактуру, и сам факт публикации в подобных журналах может расцениваться как признание в научном сообществе [*Hargens, Hagstrom*, 1982]. При этом она имеет явное ограничение, так как такой материал сокращает выводы до узких представлений, которые базируются на статьях, подпадающих под конвенцию «хорошей науки». Такое ограничение обычно обосновывается техническими возможностями — небольшая группа исследователей физически не может обработать большое количество текстов сплошным просмотром. Количественные методы разрешают эту проблему: так как методы обработки расширяют возможности для анализа большого количества текстов, это позволяет делать более масштабные выводы.

Базы данных академических текстов, сформированные в большое хранилище и доступные для использования, предоставили достаточно удобный доступ для исследователей со всего мира выгружать метаданные статей, включая цитирования. В короткие сроки по запросам с интересующими ключевыми словами можно выгрузить большие массивы метаданных статей. Большие массивы при этом легко обрабатываются инструментами. Достаточно часто ученые могут прибегать к готовым инструментам, которые направлены на работу конкретных форм записи в конкретных базах данных [*Bar-Ilan*, 2007]. Самые большие — это базы цитирований *Web of Science* и *Scopus*, вбирающие в себя огромное количество журналов и статей по очень широким диапазонам. Если можно обойтись простым анализом, не требующим детального контроля выборки и сложной визуализации, то готовые инструменты — это лучшее решение для осуществления анализа. На построенных графах цитирований авторов регулярно пишутся статьи, но если говорить о самом научном тексте как о главном источнике — в форме выгрузки чаще всего фигурируют ключевые слова или аннотация, являющиеся доступным, но проблемным объектом для анализа.

Явным ограничением аннотации как источника выступает ее краткость. Зачастую она очень сжато раскрывает суть статьи, концептуальные основы и используемую лексику, дающую понимание о структуре знания. Как правило, она не раскрывает методологическую суть статьи, и в этом ее главный недостаток. Не во всех областях существуют единые фиксированные требования по написанию аннотации статьи для журнала. Больше возможностей тем самым появляется при полнотекстовом анализе, строящемся не на терминологических сетях, а на применении методов машинного обучения. Этот сценарий можно проследить, например, в статьях, описывающих методологическую принадлежность в социальных науках. И нередки статьи, которые используют корпуса с полными текстами, задействуя алгоритмы машинного обучения [*Schwemer, Wieczorek*, 2020; *Beckert, Suckert*, 2021].

Важный фактор выборки — это временна́я рамка. Знание динамично, и содержание статей меняется с течением времени. Различные социальные изменения влияют на форму и содержание текста. Например, через исследование изменения

текста в зависимости от времени публикации текста [*Bohr*, 2014; *Keith*, 2004; *Zougris*, 2019] можно прослеживать динамические трансформации содержания. И подобные пересечения служат ценным основанием для исследования взаимосвязи условий появления публикации и ее содержания. Динамика тем, теоретических концепций, структуры текста следует за нормами современной науки, воспроизводящейся в режиме реального времени. К примеру, важность определенной области исследования может меняться из-за внешних факторов, происходящих в социально-политической реальности [*Gel'man*, 2022] или в результате внутренних процессов науки, связанных с исследовательской актуальностью. Это обостряет потребность во временном разграничении. Исследование большого временного периода как монолитного отрезка может провоцировать неточности. Для этого статьи, анализирующие динамику тем, дисциплин или иной сущности, прибегают к дроблению на более короткие отрезки [*Giordan*, 2018; *Zougris*, 2019] и последующему их детальному рассмотрению.

Концептуализация и методы

Текстовые данные и инструменты текстового анализа используются исследователями для концептуализации различных сущностей, охватывающих широкий спектр объектов исследований в социальных науках. Синоним сущности (scientific artefacts) — объект интереса, по отношению к которому авторы статей применяют методы с целью получения ответа на заданные в тексте вопросы [*Kang*, *Evans*, 2023]. В этой части текста мы систематизируем набор исследуемых сущностей и постараемся ответить на вопрос, с помощью каких методологических средств авторы извлекают сущности из научных текстов. В качестве подобранных примеров здесь выбраны статьи, не упомянутые в тексте ранее (см. табл. 1). Это намеренный шаг для демонстрации более широкой вариативности способов анализа, соответствующих при этом общей логике методологических сценариев.

Темы

Первая сущность — обсуждаемые в тексте темы, которые позволяют определить основные направления исследования, акценты в тексте и общую линию рассуждений. Для обозревания тем может использоваться качественная вычитка, чтобы определить приблизительный ландшафт возможных предметов для обсуждения [*Twine*, 2020; *Ferragina*, 2023], в том числе и используя коды. Для структурного изложения, как правило, используются методы тематического моделирования. Темы можно наблюдать в зависимости от разных факторов. В данном случае факторы — это классы, внутри которых варьируются темы. И здесь все может зависеть от временных рамок [*Giordan*, 2018] и рассматриваться в динамике или, к примеру, в зависимости от дисциплин [*Goldenstein*, 2019] или областей социальных наук [*Sbalchiero*, 2018]. Алгоритмический вывод тем показывает нам общую картину в выборке.

Каноны и парадигмы

Каноны — установленные нормы, правила и стандарты, которые присутствуют в тексте и определяют его структуру, стиль и логику представления информации. В количественном текстовом анализе каноны — не самая популярная сущность,

хотя в некоторых примерах и исследованиях задействуются дескриптивные инструменты, демонстрирующие каноны с помощью частотной иллюстрации их изменения [*Silver et al.*, 2022]. Гораздо чаще можно встретить исследования, основанные на качественной вычитке. Поскольку канон — ориентация на классический образец, часто можно встретить жанр эссе, в которых подробно разбирается генеалогия мысли внутри дисциплины со стороны признанных работ [*Atkinson*, 2001; *Gruning*, 2021]. Другой вариант исследований — качественная глубинная вычитка. В одной из таких работ на выборке из 250 книг по социологической теории авторы воспроизвели и суммировали риторику канона в социологии [*Silver et al.*, 2022].

Парадигмы — общие концептуальные рамки и подходы, которые определяют способы восприятия и анализа проблематики в тексте. Концепция парадигмальности науки была привнесена Томасом Куном при попытке объяснить научные изменения с социологической перспективы. Согласно Куну, в основе производства наук живет парадигма, определяющая теоретическую рамку конкретной дисциплины и представления о том, что является актуальным, а что таковым не является. Теоретическая рамка служит фундаментом для последующего обоснования результатов исследования [*Kuhn*, 1962]. В одной из статей, посвященных исследованию парадигмальности, авторы используют лексический анализ при помощи косинусного сходства — одного из методов машинного обучения, рассчитываемого на основе векторного представления текста[5], для измерения сходства лексики из разных документов. Подсчет производился на нескольких уровнях. Первый уровень — общая направленность наук: социально-поведенческие и естественные. Второй уровень — точечная дисциплина — физика, социология, психология и т. д. В результате авторы подсчитали близость текстов (которая концептуализировалась как степень консенсуса) для каждого из уровней [*Evans et al.*, 2016].

Методы

Средства анализа — такой же важный аналитический элемент научной работы, нуждающийся в осмыслении академического текста. В одной из публикаций, анализирующих эту сущность, исследователи фокусируются на компаративных дизайнах в разрезе методологии и общих тенденций в социологии, политологии и социальной политике.

Собрав 12 483 текста из журналов высокого квартиля по интересующим дисциплинам, а также метаинформацию о статьях и их содержании, авторы закодировали качественную и количественную парадигмы методологий. Основываясь на этой информации, авторы попытались понять, как меняются методологические тренды. Естественным образом регрессия является одним из популярнейших методов в количественной перспективе. Нередко используются кластерные анализы и качественно-сравнительные анализы QCA (Qualitative Comparative Analysis). В исследовательских дизайнах преобладают стратегии смешанных методов в кейс-стади, структурные регрессионные модели. Достаточно популярно дескриптивное доказательство, демонстрирующее результаты моделирования или частотности в распре-

[5] В основе любого компьютерного метода анализа текстов лежит пространственное представление текста. Для проведения операций со словом нужно представить его в виде кодируемой последовательности символов. Это помогает производить со словом математические операции, в том числе представление текста в векторном пространстве.

делениях данных. Иллюстрируя свои выводы количественно, авторы сами прибегают к этому методу. В работе также использовался классический метод качественной вычитки 50 случайных статей из каждого собранного журнала [*Ferragina*, *Deeming*, 2022]. Это не единственный сценарий исследований методов в академических текстах. Ранее уже разбирались статьи, основанные на измерении дистанций между кластерами [*Schwemmer*, 2020], а также кластеризация [*Beckert*, 2021] и тематическое моделирование [*Zougris*, 2019].

Теории и концепты

Концепцией можно считать конкретную дискуссию, собранную вокруг конкретной идеи внутри дискурса. Например, в социологии климатических изменений — достаточно широкой области исследования — присутствует вопрос об отношении человека и животного. Хотя в общей перспективе климат воспринимается как совокупность природных явлений, он не предоставляет достаточного места в научном обсуждении для этой проблемы, хотя это является отдельной концепцией в научной ветви исследований о климате. Основываясь на трех выборках из 12, 24 и 28 статей из высококвартильных журналов, авторы поставили задачу изучить конструирование знания о климатических изменениях. Частным объектом для этой демонстрации стало отношение «человек — животное» в трех разных областях социологии климатических изменений. Это пример текста, использующего качественный контент-анализ, — на основании подразделов с разными темами фиксировался и анализировался гендер автора, фокус журнала, фокус самой статьи. В работе анализировалось наличие дискурса отношений «человек — животное» в работах о климатических изменениях. Вычитав содержание из нескольких выборок, автор заключил, что этот дискурс не сильно учитывается, и обозначил это как недостаток [*Twine*, 2020]. С помощью дескриптивных инструментов и качественной вычитки можно выяснить представительность той или иной темы. У такого подхода есть существенные преимущества: подобный анализ глубоко погружает читателя в содержание и может продемонстрировать проблему отсутствия конкретной дискуссии в научном дискурсе. Теоретические и концептуальные вопросы гораздо удобнее исследуются качественными методами с помощью вычитки или качественного контент-анализа, так как этот сценарий позволяет максимально погрузить читателя в проблематику и разобраться в тонкостях теоретического употребления.

Научные школы

Научные школы — традиции и направления мысли, которые отражают определенные теории, методы и подходы к исследованию какой-либо области знаний. В качестве примера приведем исследование, в фокусе которого находятся Чикагская школа экономики, продвигающая неолиберальный подход в экономике, и «Группа Чарльза Ривера», которая следует кейнсианскому направлению и предлагает подход, больше опирающийся на государственное регулирование в хозяйственной политике государства. На основе библиометрического подхода анализа цитирований образуется два кластера, которые взаимодействуют друг с другом. Паттерны цитирования в группе Чарльза показывают, что связанность сети ссылок представителей направления ниже, чем у Чикагской школы. К тому же в Чикагской школе степень взаимного цитирования значимо выше, чем у кейнсианцев. Это структурно объяс-

няет, почему к 1980-м гг. более популярной в американских кругах оказалась неолиберальная мысль [*Henriksen*, 2022].

Дисциплины и специализации

Дисциплины — научные области знаний, в рамках которых происходит постановка научных вопросов. Примеры, релевантные для анализа текста в социальных науках, основаны на анализе влиятельности дисциплин через анализ социальных сетей на основе цитирования [*Moody*, 2004; *Moody, Light*, 2006] или семантических графов [*Varga*, 2011]. В работе Д. Муди и Р. Лайта на основе анализа аннотаций рассматривается место социологии в ландшафте остальных социальных наук. Авторы не только рассматривают, как поменялась тематическая структура социологии, — так, они показали, что социология отошла от фундаментальных тем к исследованиям социальных проблем, — но также на основе анализа цитирований рассматривают, как устроены границы между дисциплинами.

Специализации — различные области знаний внутри научных дисциплин. Они могут иметь как тематическую, так и методологическую основу (в этом они пересекаются с такими рассмотренными сущностями, как темы и методы). В работе Зугрэса [*Zougris*, 2019] рассматривается тематическая структура национальных социологий. Составив корпус текстов из 11 793 статей, написанных в течение 40-летнего периода, автор попытался выявить эпистемологический разрыв между социологическими дисциплинами в британской и американской социологии, демонстрируя их разность с помощью тематического моделирования. Выявленная дистанция обусловлена разностью построения исследований и академических акцентов. В ходе анализа различных тем неравенства было выявлено, что американская социология больше склонна делать упор на методы исследования, тогда как британская больше фокусируется на теории. Но этот разрыв не постоянен для всех дисциплин. Существуют как интегрирующие области знания, так и разобщающие.

Улучшение и модернизация компьютерных методов может решить множество проблем, связанных с анализом сущностей, важных для анализа академических источников. В случае применения больших языковых моделей, являющихся примером реализации для задач классификации текстов, можно привести в пример работу К. Арилиюк с коллегами, в которой была произведена попытка разделения публикаций на дисциплины, основываясь на текстах аннотаций [*Arhiliuc et al.*, 2024]. Это пример инструментального применения современных методов компьютерной лингвистики для анализа в социальных науках. Эта работа дает нам определенный взгляд на то, как современные технические решения позволяют выполнять задачи распознавания текстов и выделения их свойств. Руководствуясь этими способами анализа, можно существенно увеличить эффективность работы по определению границ дисциплин на больших выборках текстовых данных, в сравнении с классическими методами анализа [*Ibid.*].

Табл. 1. Подбор методологий к изучаемым сущностям

Table 1. Sampling of methodologies for the entities

Статья	Методы	Стратегия выборки	Сущность	Исследовательский вопрос
Ferragina E., Deeming C. Comparative Mainstreaming? Mapping the Uses of the Comparative Method in Social Policy, Sociology and Political Science since the 1970s // Journal of European Social Policy. 2023. Vol. 33. No. 1. P. 132–147. DOI: 10.1177/0958928722112843	Дескриптивная статистика + Качественная вычитка	12 483 статей из высококвартильных журналов для дескриптивного описания По 50 статей из каждого выбранного журнала для вычитки	Темы	Достигли ли сравнительные методы «зрелости» к 1990-м гг.? (для дескриптивного доказательства) Каковы содержательные характеристики наиболее цитируемых сравнительных исследований? (для качественной вычитки)
Giordan G., Saint-Blancat C., Sbalchiero S. Exploring the History of American Sociology through Topic Modelling // Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. 2018. P. 45–64. DOI: 10.1007/978-3-319-97064-6_3	Тематическое моделирование	Корпус из 3 992 аннотаций статей с 1895 по2016 г.	Темы	Как дисциплина растет и развивается во времени, учитывая социальные изменения?
Atkinson P., Coffey A., Delamont S. A Debate about Our Canon // Qualitative Research. 2001. Vol. 1. No. 1. P. 5–21. DOI: 10.1177/146879410100100101	Качественная вычитка	~ 70 статей значимых высокоцитируемых авторов	Канон	Каков канон качественных исследований?
Silver D. et al. The Rhetoric of the Canon: Functional, Historicist, and Humanist Justifications // The American Sociologist. 2022. Vol. 53. No. 3. P. 287–313. DOI: 10.1007/s12108-022-09529-0	Качественная вычитка	250 учебников по социологии	Канон	Как авторы рационализируют исключения авторов из учебников по социологии? Каковы способы и правила такой рационализации?
Evans E.D., Gomez C.J., McFarland D.A. Measuring Paradigmaticness of Disciplines using Text // Sociological Science. 2016. Vol. 3. No. 32. P. 757–778. DOI: 10.15195/v3.a32	Метод информационной энтропии[6]	167 959 статей, выгруженных в WoS по названиям	Парадигма	Какова разница между парадигмами «точных» и «мягких» наук?

[6] На английском обозначается “Shannon entropy”.

Продолжение табл. 1

Статья	Методы	Стратегия выборки	Сущность	Исследовательский вопрос
Schwemmer C., Wieczorek O. The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications // Sociology. 2020. Vol. 54. No. 1. P. 3–21. DOI: 10.1177/0038038519853146	Контрастивный анализ	Корпус из 8 737 аннотаций журналов, рецензируемых “Social Science Citation Index” (SSCI) с 1995 по 2017 г.	Парадигмы + методы	Отражается ли методологический разрыв в публикациях универсальных журналов по социологии? Если да, то в какой степени методологический разрыв отражается на предпочтениях определенных парадигм в различных социологических журналах и тенденциях публикаций с течением времени?
Beckert J., Suckert L. The Future as a Social Fact. The Analysis of Perceptions of the Future in Sociology // Poetics. 2021. Vol. 84. No. 3. P. 101499. DOI: 10.1016/j.poetic.2020.101499	Кластерный анализ + дескриптивные методы	571 публикация с 1950 по 2019 г.	Методы	Как социологи обращают внимание на будущие ориентации в широком спектре социологических областей, используя разнообразные методы и задавая широкий набор вопросов об оценках будущего?
Zougris K. Detecting Topical Divides and Topical “Bridges” Across National Sociologies // The American Sociologist. 2019. Vol. 50. No. 1. P. 63–84. DOI: 10.1007/s12108-018-9392-2	Тематическое моделирование	11 793 аннотации статей из четырех британских и четырех американских социологических журналов за 40-летний период	Методы + специализации	Какие темы способствуют разделению американской и британской социологии? Какие темы способствуют «наведению мостов» между американской и британской социологиями?
Twine R. Where Are the Nonhuman Animals in the Sociology of Climate Change? // Society & Animals. 2020. Vol. 31. No. 1. P. 105–130. DOI: 10.1163/15685306-BJA10025	Качественная вычитка	Три выборки из 12, 24 и 28 статей в высококвартильных журналах	Теории и концепты	Каково присутствие дискуссии вокруг отношений «человек – животное» в социологии изменения климата?

Окончние табл. 1

Статья	Методы	Стратегия выборки	Сущность	Исследовательский вопрос
Henriksen L.F., Seabrooke L., Young K.L. Intellectual Rivalry in American Economics: Intergenerational Social Cohesion and the Rise of the Chicago School // Socio-Economic Review. 2022. Vol. 20. No. 3. P. 989–1013. DOI: 10.1093/ser/mwac024	Сетевой анализ	Данные о цитированиях в двух противостоящих школах на основе текстов экономистов и их учеников + Качественные данные о профессиональной сплоченности профессоров и студентов	Научные школы	Как возникла и укрепилась неолиберальная доктрина в американской экономической мысли?
Moody J. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999 // American Sociological Review. 2004. Vol. 69. No. 2. P. 213–238. DOI: 10.1177/000312240406900204	Сетевой анализ	База данных “Sociological Abstracts”, охватывающая период с 1963 по 1999 г.	Дисциплины	Как разные модели сотрудничества в социологии влияют на структуру производства знания?
Varga A.V. Measuring the Semantic Integrity of Scientific Fields: a Method and a Study of Sociology, Economics and Biophysics // Scientometrics. 2011. Vol. 88. No. 1. P. 163–177. DOI: 10.1007/s11192-011-0342-9	Сетевой анализ Семантические сети	Аннотации из статей журналов по социологии (n = 5 852), экономике (n = 41 924), естественных наук и биофизике (n = 33 416)	Дисциплины + Парадигмы	Какая ветвь наук имеет большую интеграцию, социальные или естественные?
Arhiliuc C. et al. Journal Article Classification Using Abstracts: a Comparison of Classical and Transformer-Based Machine Learning Methods // Scientometrics. 2024. P. 1–30. DOI: 10.1007/s11192-024-05217-7	Классификация при помощи больших языковых моделей	Аннотации всех статей, опубликованных в 2022 г. и проиндексированных в “Science Citation Index Expanded (SCIE), “Social Sciences Citation Index-”(SSCI) и “Arts & Humanities Citation Index” (AHCI) *Web of Science* (WoS) (n = 2 077 486)	Дисциплины	Как разные методы классификации научных публикаций выполняют свою задачу на основе аннотаций?

Заключение

В статье были рассмотрены различные методы анализа академических текстов, отражающие важность глубокого изучения содержания и структуры научных работ. Применение методов анализа текстов позволяет не только более полно понять суть текста, но также выявить ключевые аспекты и особенности исследуемой темы. В современных исследованиях социальных наук, использующих академические тексты в качестве материала, существует ряд сценариев, релевантных по отношению к широкому диапазону исследовательских сущностей, воплощенных в исследуемом материале. С опорой на качественные и количественные ориентации методов мы выделили разную вариативность в этих двух парадигмах. В качественной перспективе анализа текстов ярко выражены две составляющие: 1) контент-анализ, позволяющий кодировать нужные нарративы и подмечать вариацию в исследуемых категориях источников, и 2) качественная вычитка и последующее сведение в последовательное изложение состояния исследуемой дисциплины или области знания. При исследовании канонов авторы могут полагаться на несистематизированный нарратив, представляющий генеалогию канона в виде последовательности значимых и признанных работ. Тогда как по отношению к остальным сущностям чаще используются выборки статей, которые прочитываются с целью найти исследуемые объекты интереса.

В количественных методах наблюдается большая вариативность, обусловленная развитием и экспансией компьютерных методов в другие области наук. Диапазон используемых инструментов анализа тянется от дескриптивных методов, стремящихся описать частоты, до регрессионных и специфичных оценок, кластеризации и тематических моделей, позволяющих обобщать тексты до групп или дробить их на темы. Появляются более мощные инструменты, открывающие возможность к выражению сплоченности дисциплин через дистанцию слов. С помощью нее можно описывать динамику, общность разных направлений внутри дисциплинарных и методологических кластеров.

Вместе с тем библиографическая информация играет очень важную роль в оценке влияния теорий, концептов и дисциплин. С появлением баз цитирований и систематизации библиографической информации открываются возможности для анализа массивов и корпусов текста с помощью построения семантических сетей, включающих информацию о цитировании. Внутренняя сетевая структура дает много полезной информации для понимания процесса организации и формирования знания. При помощи тех же библиометрических данных как в сетевую методологию, так и в количественную может быть включена контекстуальная информация, дающая дополнительные возможности для проведения кластеризации и моделирования текстовых массивов научной литературы.

В российском контексте также существует возможность проводить эмпирические исследования науки с применением инструментов анализа текстов — существуют полнотекстовые базы научных текстов и источник наукометрических данных в виде Российского индекса научного цитирования. В этом смысле российские данные гораздо более открыты для исследователя в отличие от многих западных баз, которые доступны по подписке. Вместе с тем российские базы редко предполагают удобную выгрузку, что требует от исследователя владения инструментами скрэпинга данных. Опора на российские данные может помочь преодолеть существенный

пробел в виде доминирования инструментов английского языка, что ограничивает сравнительные исследования и инклюзивность научных сообществ, изучающих другие языки, кроме английского [*Baden et al.*, 2022]. Российскими исследователями уже активно используются метаданные научных статей с применением сетевого анализа [*Сафонова*, *Винер*, 2013; *Губа, Семенов*, 2012; *Мальцева и др.*, 2023a, 2023b], однако анализ полных текстов пока не представлен достаточно заметно. Мы надеемся, что этот обзор станет шагом к тому, чтобы больше опираться на новые инструменты анализа.

Литература

Губа К.С. Большие данные в исследовании науки: новое исследовательское поле // Социологические исследования. 2021. № 6. С. 24–33. DOI: 10.31857/S013216250013878-8.

Губа К.С., Семенов А.В. Западная теория в петербургской социологии: между Максом Вебером и Эрвином Гоффманом // Социологические исследования. 2012. № 6 (338). С. 83–96. EDN: PBTUPT.

Мальцева Д.В., Ващенко В.А., Капустина Л.В. Методология обработки библиографических данных на русском языке для построения сетей коллаборации (на примере базы данных eLibrary) // Социология: методология, методы, математическое моделирование (Социология: 4М). 2023a. № 54–55. С. 45–78. DOI: 10.19181/4m.2022.31.1-2.2. EDN: GRRLBQ.

Мальцева Д.В., Павлова И.А., Капустина Л.В., Ващенко В.А., Фиала Д. Сравнительный анализ возможностей WoS и eLibrary для анализа библиографических сетей // Социология: методология, методы, математическое моделирование (Социология: 4М). 2023b. № 56. С. 7–68. DOI: 10.19181/4m.2023.32.1.1. EDN: ZBAAGN.

Сафонова М.А., *Винер Б.Е.* Сетевой анализ социтирований этнологических публикаций в российских периодических изданиях: предварительные результаты // Социология: методология, методы, математическое моделирование. 2013. № 36. С. 140–176. EDN: RCFOWT.

Abbott A. Chaos of Disciplines. University of Chicago Press, 2010.

Arhiliuc C. et al. Journal Article Classification Using Abstracts: a Comparison of Classical and Transformer-Based Machine Learning Methods // Scientometrics. 2024. P. 1–30. DOI: 10.1007/s11192-024-05217-7.

Atkinson P., Coffey A., Delamont S. A Debate about Our Canon // Qualitative Research. 2001. Vol. 1. No. 1. P. 5–21. DOI: 10.1177/146879410100100101.

Baden C., Pipal C., Schoonvelde M., van der Velden M.A.G. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda // Communication Methods and Measures. 2022. Vol. 16. No. 1. P. 1–18. DOI: 10.1080/19312458.2021.2015574.

Bail C.A. Can Generative AI Improve Social Science? // Proceedings of the National Academy of Sciences. 2024. Vol. 121. No. 21. P. e2314021121. DOI: 10.1073/pnas.2314021121.

Bar-Ilan J., Levene M., Lin A. Some Measures for Comparing Citation Databases // Journal of Informetrics. 2007. Vol. 1. No. 1. P. 26–34. DOI: 10.1016/j.joi.2006.08.001.

Barry A., Born G., Weszkalnys G. Logics of Interdisciplinarity // Economy and Society. 2008. Vol. 37. No. 1. P. 20–49. DOI: 10.1080/03085140701760841.

Beckert J., Suckert L. The Future as a Social Fact. The Analysis of Perceptions of the Future in Sociology // Poetics. 2021. Vol. 84. No. 3. P. 101499. DOI: 10.1016/j.poetic.2020.101499.

Bohr J., Dunlap R.E. Key Topics in Environmental Sociology, 1990–2014: Results from a Computational Text Analysis // Environmental Sociology. 2018. Vol. 4. No. 2. P. 181–195. DOI: 10.1080/23251042.2017.1393863.

Çolak K., Koç S. Bibliometric Analysis and Mapping with Vosviewer in Neet-Head Research in Social Sciences // Journal of Ekonomi. 2023. Vol. 5. No. 2. P. 7–91. DOI: 10.58251/ekonomi.1380379.

Evans J.A., Foster J.G. Metaknowledge // Science. 2011. Vol. 331. No. 6018. P. 721–725. DOI: 10.1126/science.1201765.

Evans E.D., Gomez C.J., McFarland D.A. Measuring Paradigmaticness of Disciplines using Text // Sociological Science. 2016. Vol. 3. No. 32. P. 757–778. DOI 10.15195/v3.a32.

Fain N., Vukašinović N., Kastrin A. Scientometric Exploration of Responsible Innovation: Mapping the Knowledge Landscape // Proceedings of the Design Society. 2024. Vol. 4. P. 265–274. DOI: 10.1017/pds.2024.29.

Ferragina E., Deeming C. Comparative Mainstreaming? Mapping the Uses of the Comparative Method in Social Policy, Sociology and Political Science since the 1970s // Journal of European Social Policy. 2023. Vol. 33. No. 1. P. 132–147. DOI: 10.1177/0958928722112843.

Fuhse J., Mützel S. Tackling Connections, Structure, and Meaning in Networks: Quantitative and Qualitative Methods in Sociological Network Research // Quality & Quantity. 2011. Vol. 45. No. 5. P. 1067–1089. DOI: 10.1007/s11135-011-9492-3.

Gel'man V. Exogenous Shock and Russian Studies // Post-Soviet Affairs. 2023. Vol. 39. No. 1–2. P. 1–9. DOI: 10.1080/1060586X.2022.2148814.

Ginnerskov J. Quest for Sociology: Revisiting Prevailing Understandings of a Discipline with Computational Text Analyses of Dissertations. Dissertation. Acta Universitatis Upsaliensis, 2024.

Giordan G., Saint-Blancat C., Sbalchiero S. Exploring the History of American Sociology through Topic Modelling // Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Springer, 2018. P. 45–64. DOI: 10.1007/978-3-319-97064-6_3.

Goldenstein J., Poschmann P. Analyzing Meaning in Big Data: Performing a Map Analysis using Grammatical Parsing and Topic Modeling // Sociological Methodology. 2019. Vol. 49. No. 1. P. 83–131. DOI: 10.1177/00811750198527.

Grüning B., Santoro M. Is There a Canon in This Class? // International Review of Sociology. 2021. Vol. 31. No. 1. P. 7–25. DOI: 10.1080/03906701.2021.1926674.

Hagstrom W. The Scientific Community // Human Resource Management. 1967. Vol. 6. No. 1. P. 29.

Hargens L.L., Hagstrom W.O. Scientific Consensus and Academic Status Attainment Patterns // Sociology of Education. 1982. Vol. 55. P. 183–196.

Hammersley M. Deconstructing the Qualitative-Quantitative Divide 1 // Mixing Methods: Qualitative and Quantitative Research. Routledge, 2017. P. 39–55. DOI: 10.4324/9781315248813-2.

Hassard J., Wolfram Cox J. Can Sociological Paradigms Still Inform Organizational Analysis? A Paradigm Model for Post-Paradigm Times // Organization Studies. 2013. Vol. 34. No. 11. P. 1701–1728. DOI: 10.1177/01708406134950.

Henriksen L.F., Seabrooke L., Young K.L. Intellectual Rivalry in American Economics: Intergenerational Social Cohesion and the Rise of the Chicago School // Socio-Economic Review. 2022. Vol. 20. No. 3. P. 989–1013. DOI: 10.1093/ser/mwac024.

Ianchuk S. Bibliometric Analysis and Visualization of Funding Social Housing: Connection of Sociological and Economic Research // SocioEconomic Challenges. 2021. Vol. 5. No. 1. P. 144–153. DOI: 10.21272/sec.5(1).144-153.2021.

Kang D., Evans J. Scientific Networks // The Sage Handbook of Social Network Analysis. SAGE Publications, 2023. P. 232–234.

Kondratyuk D., Straka M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally // arXiv preprint arXiv:1904.02099. 2019. DOI: 10.48550/arXiv.1904.02099.

Kuehn D., Rohlfing I. Do Quantitative and Qualitative Research Reflect Two Distinct Cultures? An Empirical Analysis of 180 Articles Suggests "No" // Sociological Methods & Research. 2022. P. 00491241221082597. DOI: 10.1177/00491241221082597.

Kuhn T.S. The Structure of Scientific Revolutions. Chicago; London: University of Chicago Press, 1962. 210 p.

Leydesdorff L. The Knowledge Content of Science and the Sociology of Scientific Knowledge // Journal for General Philosophy of Science. 1992. Vol. 23. No. 2. P. 241–263. DOI: 10.1007/BF01801451/.

Lynch M., Bogen D. Sociology's Asociological "Core": An Examination of Textbook Sociology in Light of the Sociology of Scientific Knowledge // American Sociological Review. 1997. Vol. 62. No. 3. P. 481–493.

Macanovic A. Text Mining for Social Science — The State and the Future of Computational Text Analysis in Sociology // Social Science Research. 2022. Vol. 108. P. 102784. DOI: 10.1016/j.ssresearch.2022.102784.

McKenzie D.F. Bibliography and the Sociology of Texts. Cambridge University Press, 1999. 130 p.

Milojević S. Quantifying the Cognitive Extent of Science // Journal of Informetrics. 2015. Vol. 9. No. 4. P. 962–973. DOI: 10.1016/j.joi.2015.10.005.

Mongeon P., Paul-Hus A. The Journal Coverage of Web of Science and Scopus: a Comparative Analysis // Scientometrics. 2016. Vol. 106. P. 213–228. DOI: 10.1007/s11192-015-1765-5.

Moody J. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999 //American Sociological Review. 2004. Vol. 69. No. 2. P. 213–238. DOI: 10.1177/000312240406900204.

Moody J. Trends in Sociology Titles // The American Sociologist. 2006. Vol. 37. No. 1. P. 77–80. DOI: 10.1007/s12108-006-1016-6.

Moody J., Light R. A View from Above: The Evolving Sociological Landscape // The American Sociologist. 2006. Vol. 37. No. 2. P. 67–86. DOI: 10.1007/s12108-006-1006-8.

Parodi G. Genre Organization in Specialized Discourse: Disciplinary Variation across University Textbooks // Discourse Studies. 2014. Vol. 16. No. 1. P. 65–87. DOI: 10.1177/1461445613496355.

Radha N. et al. AI-Driven Summarization of Academic Literature using Transformer Model // 2024 Second International Conference on Inventive Computing and Informatics (ICICI). IEEE, 2024. P. 359–364. DOI: 10.1109/ICICI62254.2024.00065.

Rona-Tas A. et al. Enlisting Supervised Machine Learning in Mapping Scientific Uncertainty Expressed in Food Risk Analysis // Sociological Methods & Research. 2019. Vol. 48. No. 3. P. 608–641. DOI: 10.1177/00491241177297.

Sbalchiero S. et al. What's Old and New? Discovering Topics in the American Journal of Sociology // Proceedings of 14th International Conference on Statistical Analysis of Textual Data. Rome: UniversItalia Editore, 2018. P. 724–732. DOI: 10.1007/s11135-020-00976-w.

Schwemmer C., Wieczorek O. The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications // Sociology. 2020. Vol. 54. No. 1. P. 3–21. DOI: 10.1177/0038038519853146.

Shen S. et al. SsciBERT: A Pre-Trained Language Model for Social Science Texts // Scientometrics. 2023. Vol. 128. No. 2. P. 1241–1263. DOI: 10.1007/s11192-022-04602-4.

Siler K. Citation Choice and Innovation in Science Studies // Scientometrics. 2013. Vol. 95. No. 1. P. 385–415. DOI: 10.1007/s11192-012-0881-8.

Silver D. et al. The Rhetoric of the Canon: Functional, Historicist, and Humanist Justifications // The American Sociologist. 2022. Vol. 53. No. 3. P. 287–313. DOI: 10.1007/s12108-022-09529-0.

Turner J.H. Explaining the Social World: Historicism versus Positivism // The Sociological Quarterly. 2006. Vol. 47. No. 3. P. 451–463. DOI: 10.1111/j.1533-8525.2006.00053.x.

Turner J.H. Academic Journals and Sociology's Big Divide: A Modest but Radical Proposal // The American Sociologist. 2016. Vol. 47. No. 2. P. 289–301. DOI: 10.1007/s12108-015-9296-3.

Twine R. Where Are the Nonhuman Animals in the Sociology of Climate Change? // Society & Animals. 2020. Vo. 31. No. 1. P. 105–130. DOI: 10.1163/15685306-BJA10025.

Varga A.V. Measuring the Semantic Integrity of Scientific Fields: a Method and a Study of Sociology, Economics and Biophysics // Scientometrics. 2011. Vol. 88. No. 1. P. 163–177. DOI: 10.1007/s11192-011-0342-9.

Weber M., Reichardt M. Evaluation Is All You Need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer Using Open Models // arXiv preprint arXiv: 2401.00284. 2023. DOI: 10.48550/arXiv.2401.00284.

Wu L. et al. Metrics and Mechanisms: Measuring the Unmeasurable in the Science of Science // Journal of Informetrics. 2022. Vol. 16. No. 2. P. 101290. DOI: 10.1016/j.joi.2022.101290.

Ziems C. et al. Can Large Language Models Transform Computational Social Science? // Computational Linguistics. 2024. Vol. 50. No. 1. P. 237–291. DOI: 10.1162/coli_a_00502.

Zougris K. Detecting Topical Divides and Topical "Bridges" Across National Sociologies // The American Sociologist. 2019. Vol. 50. No. 1. P. 63–84. DOI: 10.1007/s12108-018-9392-2.

Methodological Scenarios for Researching Academic Texts in Social Sciences

Alexandr A. Vilkhovenko

European University at Saint-Petersburg,
St. Petersburg, Russia;
e-mail: avilhovenko@eu.spb.ru

The article provides an overview of scientific works using a range of methods for analyzing the texts of scientific articles and textbooks. Using examples from scientometrics, sociology of science, science of science, metascience and other areas, the paper demonstrates the conditions and dissemination of science studying nowadays. The relevance of the review is associated with the development over the past ten years of new methods for analyzing text data, which have begun to be applied to large amounts of scientific information. Based on the qualitative and quantitative orientations of text analysis methods, the variability in these two approaches at the level of sampling and specific tools is highlighted, which is manifested in the study of such objects as scientific topics, paradigms and canons, concepts, methods, disciplines and specializations. When studying canons and concepts, researchers rely on the strategy of deep reading. Whereas in relation to other objects, more extensive samples of articles are more often used, to which it is possible to apply network tools and methods of computational linguistics.

***Keywords*:** methodology of text analysis, sociology of science, network analysis, clusterization, lexical analysis, methodological landscapes, science of science, scientific artefacts, computational sociology.

Acknowledgment

With great respect I express my gratitude to the director of the Center for Institutional Analysis of Science and Education, Katerina Sergeevna Guba, for her invaluable assistance in systematizing knowledge.

References

Abbott, A. (2010). *Chaos of Disciplines*, University of Chicago Press.

Arhiliuc, C. et al. (2024). Journal Article Classification Using Abstracts: a Comparison of Classical and Transformer-Based Machine Learning Methods, *Scientometrics*, 1–30. DOI: 10.1007/s11192-024-05217-7.

Atkinson, P., Coffey, A., Delamont, S. (2001). A Debate about Our Canon, *Qualitative Research, 1* (1), 5–21.

Baden, C., Pipal, C., Schoonvelde, M., van der Velden, M.A.G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda, *Communication Methods and Measures, 16* (1), 1–18. DOI: 10.1080/19312458.2021.2015574.

Bail, C.A. (2024). Can Generative AI Improve Social Science? *Proceedings of the National Academy of Sciences, 121* (21), e2314021121. DOI: 10.1073/pnas.2314021121.

Bar-Ilan, J., Levene, M., Lin, A. (2007). Some Measures for Comparing Citation Databases, *Journal of Informetrics, 1* (1), 26–34. DOI: 10.1016/j.joi.2006.08.001.

Barry, A., Born, G., Weszkalnys, G. (2008). Logics of Interdisciplinarity, *Economy and Society, 37* (1), 20–49. DOI: 10.1080/03085140701760841.

Beckert, J., Suckert, L. (2021). The Future as a Social Fact. The Analysis of Perceptions of the Future in Sociology, *Poetics*, *84* (3), 101499. DOI: 10.1016/j.poetic.2020.101499.

Bohr, J., Dunlap, R.E. (2018). Key Topics in Environmental Sociology, 1990–2014: Results from a Computational Text Analysis, *Environmental Sociology, 4* (2), 181–195. DOI: 10.1080/23251042.2017.1393863.

Çolak, K., Koç, S. (2023). Bibliometric Analysis and Mapping with Vosviewer in Neet-Head Research in Social Sciences, *Journal of Ekonomi, 5* (2), 77–91. DOI: 10.58251/ekonomi.1380379.

Evans, J.A., Foster, J.G. (2011). Metaknowledge, *Science, 331* (6018), 721–725. DOI: 10.1126/science.1201765.

Evans, E.D., Gomez, C.J., McFarland, D.A. (2016). Measuring Paradigmaticness of Disciplines using Text, *Sociological Science*, *3* (32), 757–778. DOI: 10.15195/v3.a32.

Fain, N., Vukašinović, N., Kastrin, A. (2024). Scientometric Exploration of Responsible Innovation: Mapping the Knowledge Landscape, *Proceedings of the Design Society*, vol. 4, 265–274. DOI: 10.1017/pds.2024.29.

Ferragina, E., Deeming, C. (2023). Comparative Mainstreaming? Mapping the Uses of the Comparative Method in Social Policy, Sociology and Political Science since the 1970s, *Journal of European Social Policy, 33* (1), 132–147. DOI: 10.1177/0958928722112843.

Fuhse, J., Mützel, S. (2011). Tackling Connections, Structure, and Meaning in Networks: Quantitative and Qualitative Methods in Sociological Network Research, *Quality & Quantity*, *45* (5), 1067–1089. DOI: 10.1007/s11135-011-9492-3.

Gel'man, V. (2023). Exogenous Shock and Russian Studies, *Post-Soviet Affairs, 39* (1–2), 1–9. DOI: 10.1080/1060586X.2022.2148814.

Ginnerskov, J. (2024). *Quest for Sociology: Revisiting Prevailing Understandings of a Discipline with Computational Text Analyses of Dissertations*, Dissertation, Acta Universitatis Upsaliensis.

Giordan, G., Saint-Blancat, C., Sbalchiero, S. (2018). Exploring the History of American Sociology through Topic Modelling, in *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (pp. 45–64), Springer. DOI: 10.1007/978-3-319-97064-6_3.

Goldenstein, J., Poschmann, P. (2019). Analyzing Meaning in Big Data: Performing a Map Analysis using Grammatical Pparsing and Topic Modeling, *Sociological Methodology, 49* (1), 83–131. DOI: 10.1177/00811750198527.

Grüning, B., Santoro, M. (2021). Is There a Canon in this Class? *International Review of Sociology, 31* (1), 7–25. DOI: 10.1080/03906701.2021.1926674.

Guba, K.S. (2021). Bol'shiye dannyye v issledovanii nauki: novoye issledovatelskoye pole [Big data in scientific research: A new research field], *Sotsiologicheskiye issledovaniya*, no. 6, 24–33 (in Russian). DOI 10.31857/S013216250013878-8.

Guba, K.S. Semenov, A.V. (2012). Zapadnaya teoriya v peterburgskoy sotsiologii: mezhdu Maksom Veberom i Ervinom Goffmanom, *Sotsiologicheskiye issledovaniya*, no. 6 (338), 83–96 (in Rusian). EDN PBTUPT.

Hagstrom, W. (1967). The Scientific Community, *Human Resource Management, 6* (1), 29.

Hargens, L.L., Hagstrom, W.O. (1982). Scientific Consensus and Academic Status Attainment Patterns, *Sociology of Education*, vol. 55, 183–196.

Hammersley, M. (2017). Deconstructing the Qualitative-Quantitative Divide 1, in *Mixing Methods: Qualitative and Quantitative Research* (pp. 39–55), Routledge. DOI: 10.4324/9781315248813-2.

Hassard, J., Wolfram Cox, J. (2013). Can Sociological Paradigms Still Inform Organizational Analysis? A Paradigm Model for Post-Paradigm Times, *Organization Studies, 34* (11), 1701–1728. DOI: 10.1177/01708406134950.

Henriksen, L.F., Seabrooke, L., Young, K.L. (2022). Intellectual Rivalry in American Economics: Intergenerational Social Cohesion and the Rise of the Chicago School, *Socio-Economic Review, 20* (3), 989–1013. DOI: 10.1093/ser/mwac024.

Ianchuk, S. (2021). Bibliometric Analysis and Visualization of Funding Social Housing: Connection of Sociological and Economic Research, *SocioEconomic Challenges, 5* (1), 144–153. DOI: 10.21272/sec.5(1).144-153.2021.

Kang, D., Evans, J. (2023). Scientific Networks, in *The Sage Handbook of Social Network Analysis* (pp. 232–234), SAGE.

Kondratyuk, D., Straka, M. (2019). 75 Languages, 1 Model: Parsing Universal Dependencies Universally, *arXiv preprint arXiv: 1904.02099*. DOI: 10.48550/arXiv.1904.02099.

Kuehn, D., Rohlfing, I. (2022). Do Quantitative and Qualitative Research Reflect Two Distinct Cultures? An Empirical Analysis of 180 Articles Suggests "No", *Sociological Methods & Research*, p. 00491241221082597. DOI: 10.1177/00491241221082597.

Kuhn, T.S. (1997). *The Structure of Scientific Revolutions*, Chicago; London: University of Chicago Press.

Leydesdorff, L. (1992). The Knowledge Content of Science and the Sociology of Scientific Knowledge, *Journal for General Philosophy of Science*, *23* (2), 241–263. DOI: 10.1007/BF01801451.

Lynch, M., Bogen, D. (1997). Sociology's Asociological "Core": An Examination of Textbook Sociology in Light of the Sociology of Scientific Knowledge, *American Sociological Review*, *62* (3), 481–493.

Macanovic, A. (2022). Text Mining for Social Science — The State and the Future of Computational Text Analysis in Sociology, *Social Science Research*, vol. 108, p. 102784. DOI: 10.1016/j.ssresearch.2022.102784.

Mal'tseva, D.V., Pavlova, I.A., Kapustina, L.V., Vashchenko, V.A., Fiala, D. (2023). Sravnitel'nyy analiz vozmozhnostey *WoS* i *eLibrary* dlya analiza bibliograficheskikh setey [Comparative analysis of the capabilities of *WoS* and *eLibrary* for bibliographic network analysis], *Sotsiologiya: metodologiya, metody, matematicheskoye modelirovaniye (Sotsiologiya: 4M)*, no. 56, 7–68 (in Russian). DOI: 10.19181/4m.2023.32.1.1. EDN: ZBAAGN.

Mal'tseva, D.V., Vashchenko, V.A., Kapustina, L.V. (2023). Metodologiya obrabotki bibliograficheskikh dannykh na russkom yazyke dlya postroyeniya setey kollaboratsii (na primere bazy dannykh *eLibrary*) [Methodology for processing bibliographic data in Russian for building collaboration networks (based on the *eLibrary* database)], *Sotsiologiya: metodologiya, metody, matematicheskoye modelirovaniye (Sotsiologiya: 4M)*, no. 54–55, 45–78 (in Russian). DOI: 10.19181/4m.2022.31.1-2.2. EDN: GRRLBQ.

McKenzie, D.F. (1999). *Bibliography and the Sociology of Texts*, Cambridge University Press.

Milojević, S. (2015). Quantifying the Cognitive Extent of Science, *Journal of Informetrics, 9* (4), 962–973. DOI: 10.1016/j.joi.2015.10.005.

Mongeon, P., Paul-Hus, A. (2016). The Journal Coverage of Web of Science and Scopus: a Comparative Analysis, *Scientometrics*, vol. 106, 213–228. DOI: 10.1007/s11192-015-1765-5.

Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999, *American Sociological Review, 69* (2), 213–238. DOI: 10.1177/000312240406900204.

Moody, J. (2006). Trends in Sociology Titles, *The American Sociologist, 37* (1), 77–80. DOI: 10.1007/s12108-006-1016-6.

Moody, J., Light, R. (2006). A View from Above: The Evolving Sociological Landscape, *The American Sociologist, 37* (2), 67–86. DOI: 10.1007/s12108-006-1006-8.

Parodi, G. (2014). Genre Organization in Specialized Discourse: Disciplinary Variation across University Textbooks, *Discourse Studies, 16* (1), 65–87. DOI: 10.1177/1461445613496355.

Radha, N. et al. (2024). AI-Driven Summarization of Academic Literature using Transformer Model, in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)* (pp. 359–364), IEEE. DOI: 10.1109/ICICI62254.2024.00065.

Rona-Tas, A. et al. (2019). Enlisting Supervised Machine Learning in Mapping Scientific Uncertainty Expressed in Food Risk Analysis, *Sociological Methods & Research, 48* (3), 608–641. DOI: 10.1177/00491241177297.

Safonova, M.A., Viner, B.E. (2013). Setevoy analiz sotsitirovaniy etnologicheskikh publikatsiy v rossiyskikh periodicheskikh izdaniyakh: predvaritel'nyye rezul'taty [Network analysis of ethnological publications in Russian periodicals: preliminary results], *Sotsiologiya: metodologiya, metody, matematicheskoye modelirovaniye*, no. 36, 140–176 (in Russian). EDN: RCFOWT.

Sbalchiero, S. et al. (2018). What's Old and New? Discovering Topics in the American Journal of Sociology, in *Proceedings of 14th International Conference on Statistical Analysis of Textual Data* (pp. 724–732), Rome: UniversItalia Editore. DOI: 10.1007/s11135-020-00976-w.

Schwemmer, C., Wieczorek, O. (2020). The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications, *Sociology, 54* (1), 3–21. DOI: 10.1177/0038038519853146.

Shen, S. et al. (2023). SsciBERT: A Pre-Trained Language Model for Social Science Texts, *Scientometrics, 128* (2), 1241–1263. DOI: 10.1007/s11192-022-04602-4.

Siler, K. (2013). Citation Choice and Innovation in Science Studies, *Scientometrics, 95* (1), 385–415. DOI: 10.1007/s11192-012-0881-8.

Silver, D. et al. (2022). The Rhetoric of the Canon: Functional, Historicist, and Humanist Justifications, *The American Sociologist, 53* (3), 287–313. DOI: 10.1007/s12108-022-09529-0.

Turner, J.H. (2006). Explaining the Social World: Historicism Versus Positivism, *The Sociological Quarterly, 47* (3), 451–463. DOI: 10.1111/j.1533-8525.2006.00053.x.

Turner, J.H. (2016). Academic Journals and Sociology's Big Divide: A Modest but Radical Proposal, *The American Sociologist, 47* (2), 289–301. DOI: 10.1007/s12108-015-9296-3.

Twine, R. (2020). Where Are the Nonhuman Animals in the Sociology of Climate Change?, *Society & Animals, 31* (1), 105–130. DOI: 10.1163/15685306-BJA10025.

Varga, A.V. (2011). Measuring the Semantic Integrity of Scientific Fields: a Method and a Study of Sociology, Economics and Biophysics, *Scientometrics, 88* (1), 163–177. DOI: 10.1007/s11192-011-0342-9.

Weber, M., Reichardt, M. (2023). Evaluation Is All You Need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer Using Open Models*, arXiv preprint arXiv:2401.00284.* DOI: 10.48550/arXiv.2401.00284.

Wu, L. et al. (2022). Metrics and Mechanisms: Measuring the Unmeasurable in the Science of Science, *Journal of Informetrics, 16* (2), 101290. DOI: 10.1016/j.joi.2022.101290.

Ziems, C. et al. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics, 50* (1), 237–291. DOI: 10.1162/coli_a_00502.

Zougris, K. (2019). Detecting Topical Divides and Topical "Bridges" across National Sociologies, *The American Sociologist, 50* (1), 63–84. DOI: 10.1007/s12108-018-9392-2.