

А.М. Либман

д.э.н., PhD (Economics), профессор,

Берлинский Свободный университет (Берлин)

ТЕКСТ-КАК-ДАННЫЕ: ЭКОНОМИЧЕСКАЯ ПЕРСПЕКТИВА

Аннотация. Формальные количественные методы анализа текстов – подходы, объединённые общим названием «текст-как-данные» («text-as-data») – играют всё более важную роль в современных общественных науках. В то же время, если в политологии и отчасти в социологии их внедрение является естественным процессом, во многом соответствующим логике развития этих дисциплин, то экономической науке место «text-as-data» неоднозначно. Соответствуя общим тенденциям на освоение постоянно растущего спектра количественных методов, «text-as-data» в то же время противоречат некоторым важным принципам экономической науки — в частности, вниманию к выявленным предпочтениям, а не риторике, и принципам «революции достоверности». Растущая роль «text-as-data» показывает сложный и нелинейный характер прогресса количественных методов в экономике (и в общественных науках в целом).

Ключевые слова: текст-как-данные, революция достоверности, выявленные предпочтения.

JEL: B40, B41, N10, Z13

УДК: 303.101.5, 303.43

DOI: 10.52342/2587-7666VTE_2025_3_38_48

© А.М. Либман, 2025

© ФГБУН Институт экономики РАН «Вопросы теоретической экономики», 2025

ДЛЯ ЦИТИРОВАНИЯ: Либман А.М. Текст-как-данные: экономическая перспектива // Вопросы теоретической экономики. 2025. №3. С. 38–48. DOI: 10.52342/2587-7666VTE_2025_3_38_48.

FOR CITATION: Libman A. Text-as-Data: An Economics Perspective // Voprosy teoreticheskoy ekonomiki. 2025. No. 3. Pp. 38–48. DOI: 10.52342/2587-7666VTE_2025_3_38_48.

Введение

Количественные методы и подходы, использующиеся в общественно-научных дисциплинах, сильно меняются во времени. В некоторых случаях это отражает объективный прогресс науки (скажем, появление новых статистических инструментов анализа), в некоторых — смену «научных мод», а в некоторых — «исчерпание» тех или иных исследовательских повесток дня, на смену которым приходят другие (требующие других инструментов). Тем не менее на протяжении последних десятилетий почти неизменным оставалось то обстоятельство, что экономическая наука являлась ведущей дисциплиной среди общественных наук с точки зрения внедрения новых количественных методов. Многие методы, ставшие сегодня стандартом в политологии или социологии, исходно были «позаимствованы» у экономистов. Так, «революция достоверности» — фокусирование на исследовании причинно-следственных связей как на основном ориентире количественных методов — исходно началась в экономике и лишь впоследствии «перекинулась» на социологию и политологию. Для многих методов количественного анализа — от регрессионного анализа до панельных данных — можно «проследить» их распространение от экономики

к другим дисциплинам (а сама экономика, в свою очередь, опирается на достижения статистики и математики).

Конечно, существуют примеры количественных методов, популярных за пределами экономической науки, но практически не получивших распространение в экономике. Экономисты редко используют многоуровневые модели [Oshchepkov, Shirokanova, 2022] или направленные нециклические графы (directed acyclical graphs, DAGs) как способ моделирования причинно-следственных взаимосвязей [Imbens, 2020]. Однако в общем и целом экономическая наука стабильно остаётся лидером по использованию формальных методов. Это связано с самим пониманием научности в экономике (если в политологии и социологии сторонники качественных подходов составляют влиятельные фракции в дисциплинах или даже доминируют, в экономике они маргинализированы в рамках отдельных областей гетеродоксии)¹; с особенностями программ подготовки экономистов (в которых основное внимание уделяется количественным методам и методам как таковым); и с самим характером экономических данных, которые легче поддаются квантификации. Действительно, если, скажем, цены или показатели государственной статистики «естественным образом» носят количественный характер, то объекты исследования других дисциплин — политические системы, программы политических партий, правовые или общественные нормы — существуют в форме текстов или даже неформальных практик поведения, т.е. для их количественного анализа необходимо подвергнуть их трансформации [Schumpeter, 1933].

В этой связи особенно интересными представляются примеры, когда экономисты скорее отстают во внедрении принципиально новых количественных исследовательских парадигм по сравнению с неэкономистами. Распространение методов «text-as-data» (текст-как-данные) — формальных подходов к анализу текстов — является таким необычным и в то же время важным примером. «Text-as-data», насколько можно судить, пользуется гораздо большей популярностью в политологии, чем в экономической науке. И хотя в обеих дисциплинах можно найти примеры применения «text-as-data», политологи более активно обращаются к этому инструменту. По некоторым оценкам, методы «text-as-data» становятся «мэйнстримом политической науки» [Wilkerson, Casas, 2017], стандартным элементом подготовки аспирантов и политологического анализа в целом.

В настоящей статье мы попытаемся понять, что означает распространение метода (точнее, целого комплекса быстро развивающихся методов) «text-as-data» для экономической науки, и как эти методы (как минимум, потенциально) могут изменить саму экономическую науку. Эта статья не является обзором самих методов «text-as-data» и областей их применения. Нас интересует, скорее, обсуждение «text-as-data» как примера необычных траекторий развития количественных методов в общественных науках, порой противоречащих типичным описаниям их истории и динамики. Помимо факта большого успеха других общественных дисциплин (а не экономической науки) во внедрении нового метода количественного анализа, «text-as-data» показывает и то, что сама по себе логика развития количественных методов часто сильно отличается от упрощённого представления о прогрессе методов, в рамках которого менее совершенные (с точки зрения того или иного стандарта научности) методы замещаются более совершенными.

Текст-как-данные: базовые концепции

Сама по себе идея, лежащая в основе «text-as-data», не нова. Исследование и анализ текстов является, наверное, одним из универсальных элементов общественных и гуманитарных наук, и в нём интуитивно можно выделить два возможных подхода. С одной стороны, тексты могут быть подвергнуты качественному анализу — в этом случае исследователь

¹ Хотя и существуют оптимистические оценки, предполагающие рост их влияния [Starr, 2014].

в буквальном смысле читает и интерпретирует текст, пытаясь понять его содержание, логику и аргументацию. С другой стороны, тексты могут быть предметом количественного исследования — в этом случае выделяются те или иные формальные характеристики текстов (например, использование тех или иных слов и выражений, лексических конструкций и т.д.), которые затем сравниваются с другими текстами. Преимущества количественного подхода достаточно очевидны — он, во-первых, позволяет анализировать большие массивы текстов, не читая их (что потребовало бы очень больших затрат времени и усилий), а, во-вторых, свободен от влияния субъективных факторов, ведущих к тому, что различные исследователи на основе одного и того же текста могут сделать разные выводы. В то же время количественный анализ обладает очевидным недостатком — любые формальные характеристики по определению упрощают текст, игнорируя часть его содержания. «Чтение» позволяет гораздо лучше понять текст, чем сравнение его количественных характеристик. Иначе говоря, количественные подходы всегда основаны на достаточно сильных допущениях, которые могут привести к искажению анализа (см., например: [Baden, Pipal, Schoonvelde, van der Velden, 2022]).

Расцвет методов «text-as-data» связан с двумя обстоятельствами. Во-первых, последние десятилетия стали периодом резкого роста доступных для формального анализа текстов — как за счёт дигитализации печатных материалов, ставшей основой больших корпусов текстов, так и постоянной генерации текстов в Интернете (в частности, в социальных сетях), наверное, являющейся беспрецедентной в человеческой истории. Не случайно ключевым аспектом революции «text-as-data» является именно анализ цифровых данных [Brady, 2019; Lazer, Radford, 2017], хотя, конечно, в центре исследования могут находиться и традиционные типы текстов (и их анализ может быть даже более интересным). Во-вторых, растущие компьютерные мощности и разработка методов машинного обучения позволяют использовать более сложные методы для анализа этих текстовых данных, в частности, сокращая влияние субъективных оценок исследователей. Речь может идти об анализе сотен тысяч или даже миллионов текстов — ещё совсем недавно такие проекты были бы технически невозможны.

Можно говорить о своего рода иерархии методов «text-as-data» от чисто качественного «понимания» текстов к безусловно количественным подходам. В работе [Benoit, 2020] выделяются два качественных и два количественных подхода к анализу текстов. Первый — анализ дискурсов (discourse analysis) уже упоминался нами и представляет собой «простое» чтение и понимание текстов (слово «простое», безусловно, следует поставить в кавычки, поскольку ничего «простого» в процессе чтения текстов нет — как показывают ограничения, с которыми сталкивается, например, искусственный интеллект в этом отношении). Второй — контент-анализ — основан на выделении исследователем смысловых единиц в тексте и их систематизации. Ключевую роль в данном случае всё равно играет чтение текста, но чтение, не просто ориентированное на понимание содержания текста, а систематически идентифицирующее фрагменты текста, содержащие те или иные аргументы или риторические приёмы.

Третий подход — гибридный количественный — основан на анализе формальных характеристик текста, которые, однако же, исходно определяются исследователями. Самым ярким примером этой группы методов является словарный подход (dictionary approach) — выделение ключевых слов и поиск частоты их употребления и использования в тексте. Например, исследователи политических программ пытаются понять, как часто в тексте упоминаются те или иные ключевые понятия, являющиеся основами определённых идеологий (в 1990-е гг. в России примером могло бы, например, быть словосочетание «радикальная рыночная реформа») или политических позиций. Естественно, выбор слов во многом субъективен и опирается на понимание исследователем типичных структур риторики акторов, которых он изучает, — т. е. или на качественный анализ, или просто на

знание «кейса». Если то или иное понятие находится вне внимания исследователя, то словарный подход никак не поможет выявить соответствующие характеристики риторики.

Наконец, четвёртый подход основан на методах машинного обучения. Его преимущество состоит в полном освобождении от вмешательства исследователя в процесс изучения текстов. Например, так называемый «topic modeling» позволяет выделить типичные темы, встречающиеся в корпусе текстов (на практике речь идёт о сочетаниях слов, типичных для тех или иных тем). «Sentiment analysis» позволяет выделить, какие эмоции обычно сопровождают использование тех или иных слов в тексте (скажем, названия стран или политических групп), упоминаются ли они скорее в позитивном или в негативном ключе. Методы этой четвёртой группы включают в себя, в частности, подходы, позволяющие выявлять латентные характеристики текстов — например, классифицировать тексты по их политическим позициям (см., например: [Proksch, Slapin, 2009; Ash, Gauthier, Widmer, 2024]).

Задачи работ, использующих методы «text-as-data», можно свести к трём группам вопросов. Во-первых, «text-as-data» пытается *описать* те или иные группы текстов с точки зрения их лингвистических и содержательных особенностей. Речь идёт не только о типичных нарративах или эмоциональной окраске, но и, например, о сложности языка — тема, весьма важная как для речей политиков, так и для стиля научных статей. Во-вторых, исследователи пытаются *классифицировать* доступные тексты, например, исходя из того, какие нарративы (topics) в них господствуют. Как уже было указано выше, современные методы позволяют реализовывать такого рода проекты с минимальными допущениями, вводящимися исследователем. В-третьих, естественным приложением «text-as-data» является *сравнение* текстов (речей, статей, постов в социальных сетях), произведённых теми или иными акторами — например, сравнение риторики стран в Генеральной ассамблее ООН, риторики политических партий или риторики исследователей, принадлежащих к различным научным школам. Собственно говоря, именно последняя задача наиболее важна для общественных наук, ориентированных на тестирование гипотез, — например, как те или иные характеристики «говорящего» влияют на содержание производимых им текстов?

Общей характеристикой всех методов «text-as-data» является, однако же, весьма специфическое восприятие текста: как совокупности слов (bag-of-words). Иначе говоря, первым шагом любого исследования «text-as-data» становится разбивка текста на минимальные единицы (например, слова) и сопоставление текстов, исходя из использующихся в них слов и их сочетаний. Это, вне всякого сомнения, представляет собой весьма упрощённый способ чтения текстов по сравнению с тем, как их читают люди. Однако он позволяет использовать формальные методы для анализа больших корпусов текстов.

В этой связи целесообразно упомянуть и новейшие тенденции в формальном анализе текстов — использование в них искусственного интеллекта (больших лингвистических моделей, LLM). В этом случае, например, LLM используется для классификации текстов: исследователь даёт соответствующее задание (промпт) модели, а она уже распределяет тексты по категориям и выделяет их основные темы. Некоторые исследования показывают, что LLM лучше справляются с этой задачей, чем люди, нанимаемые для аннотации текстов (и, безусловно, использование LLM обходится гораздо дешевле) [Gilardi, Alizadeh, Kubli, 2023]. В то же время искусственный интеллект сегодня сам генерирует большие массивы текстов, и такие искусственные массивы также могут представлять интерес для исследователей [Linegar, Kocielnik, Alvarez, 2023].

Легко понять, почему методы «text-as-data» так привлекают, например, специалистов в области политической науки [Grimmer, Stewart, 2013; Wikerson, Casas, 2017]. Для политологии изучение текстов (скажем, программ политических партий, законов, выступлений политиков или пропагандистских документов) — нормальная часть

исследовательской повестки дисциплины. Политологов интересует, например, как политическая риторика сама по себе (её содержание, типичные приёмы, эволюция во времени), так и её способность влиять на поведение политических игроков (например, как те или иные стратегии избирательной кампании содействуют успеху или провалу партий). К тому же во многих случаях у политологов нет выбора, кроме как обратиться к изучению текстов — скажем, опросы или полевые исследования могут быть нереализуемы в тех или иных политических контекстах или связаны с большими издержками. В отсутствие доступных опросных данных привлекательным становится изучение социальных сетей — несмотря на то, что, конечно, посты, которые публикуются в них, не могут считаться корректным отображением предпочтений акторов. Политологи и в прошлом были склонны к генерации количественных данных на основе трансформации текстов: например, существовали (и существуют) большие проекты, посвящённые описанию программ политических партий [Gemenis, 2013], пусть и генерировавшиеся «вручную» кодировщиками, читавшими тексты и классифицировавшими их. Сегодня методы «text-as-data» дополняют и совершенствуют такие проекты.

Для экономической науки, однако, ситуация более сложна. Как будет показано в следующем разделе, если для политологии введение методов «text-as-data» в анализ было во многом естественным, то для экономики оно означало довольно существенную трансформацию целого ряда важных направлений развития дисциплины. Логика развития экономической науки (и экономической методологии) во многом направлена в противоположном направлении по сравнению с тем, к которому исследователей «подталкивают» практики «text-as-data».

Текст-как-данные: вызов для экономистов?

Конечно, экономическая наука не может пройти мимо появления столь интересного набора количественных подходов — тем более, что масштабы текстовой информации, которую производит современное цифровое общество и которая может быть использована для исследования, постоянно растут. Огромные массивы данных не могут не привлекать экономистов, и, действительно, в экономической науке «text-as-data» тоже используется всё более и более активно [Gentzkow, Kelly, Taddy, 2019; Loughran, MacDonald, 2020; Ash, Hansen, 2023]. В каком-то смысле интерес к «text-as-data» является частью интереса к большим данным [Einav, Levin, 2014], которые в экономике также получают всё большее и большее распространение. В то же время «text-as-data» отличается по своей логике от многих других предшествующих этапов эволюции количественных методов дисциплины.

Экономика как позитивистская дисциплина описывает себя на языке накопления знаний — каждое следующее поколение экономистов, по мнению большинства исследователей, знает об обществе, политике и экономике больше, чем предшествующее. Справедливость такого самоописания для эмпирической экономики может быть оспорена [Либман, 2024]; для настоящей статьи, однако же, важно, что подобное описание экономисты применяют и тогда, когда говорят об эволюции *методов* экономической науки. Например, революция достоверности описывает себя как (на настоящий момент) высший этап в развитии эмпирических методов, направленных на выявление причинно-следственных взаимосвязей, т.е. на адекватное тестирование экономических теорий, — от простых корреляций и регрессий к панельным данным и инструментальным переменным к идее эксперимента как «золотого стандарта» экономической науки. Иначе говоря, господствующий нарратив предполагает, что работы экономистов, написанные десятилетием ранее, уступают в качестве современным; методология постоянно совершенствуется. «Text-as-data» во многом выбивается из такого нарратива.

Текст-как-данные и выявленные предпочтения

Общей чертой практически всех общественных наук, кажущейся, на первый взгляд, самоочевидной, является представление о возможности получения информации об объектах исследования — индивидах, их объединениях и институтах — за счёт коммуникации с ними. Для социолога или политолога очевидным представляется начало исследования той или иной темы с попытки понять, что именно *сами участники* того ли иного экономического или общественного явления говорят о причинах и логике своих действий (в некоторых случаях даже определение ключевых понятий следует тому, как их понимают реальные социальные игроки). Конечно же, политологи и экономисты далеки от наивного отношения к этой теме — им прекрасно известно, что люди часто искажают свои предпочтения и мотивы, а их коммуникации (как с исследователями, так и с другими людьми) могут носить стратегический характер. Значительная часть общественных (и, кстати, гуманитарных наук — если вспомнить о концепции «исторической критики» в истории) посвящена именно тому, каким образом можно справиться с различного рода искажениями в доступных нам высказываниях людей — письменных или устных².

Экономическая наука представляет собой исключение среди других общественно-научных дисциплин. Её ключевой подход — принципиальное недоверие к любой информации, которую люди сами сообщают о себе. Концепция «cheap talk» предполагает, что слова экономических агентов не могут восприниматься как значимая информация — они игнорируются другими экономическими агентами и, соответственно, должны игнорироваться исследователями ([Farrell, Rabin 1996]. Многие экономические подходы основываются на том, что агенты будут в общем случае рассматривать риторику других игроков как недостоверный (non-credible) сигнал: «человек экономический» принципиально не верит обещаниям других людей³. Для того чтобы сигнал стал достоверным, он должен быть связан с издержками для отправляющего его агента (costly signal), а такие издержки, как полагают экономисты, у простой риторики практически отсутствуют. Иначе говоря, любые слова должны быть подкреплены действиями и лишь в этой ситуации рассматриваться как значимые.

Именно поэтому экономисты традиционно крайне скептически относятся к таким инструментам, как опросы общественного мнения или анализ текстов; вместо этого они предпочитают опираться на методы определения выявленных предпочтений (revealed preferences), т.е. считывать цели агентов, исходя из их поведения и тех решений, которые они принимают (предпочтительно связанных с существенными издержками для агентов) ([Crawford, De Rock, 2014]. По мнению экономистов, нет смысла спрашивать агентов об их предпочтениях — последние можно выявить исключительно исходя из того, какие решения они будут принимать⁴. Игроки будут или стратегически манипулировать своими ответами для достижения тех или иных целей, или просто выдавать случайные ответы⁵. Эта же причина объясняет, почему в экспериментальной экономике основное внимание уделяется конкретным

² Некоторые направления социологии в принципе считают эту цель недостижимой, полагая главной целью исследования анализ того, каким образом люди «конструируют» социальную реальность, а не выявление «подлинных» фактов.

³ Теория рациональных ожиданий является примером применения этого типа аргументов к макроэкономическим проблемам.

⁴ Сказанное не отрицает, что экономисты порой пользуются основанной на опросах информацией [Boulrier, Goldfarb, 1998], но всё же внимание к опросам в экономике в разы меньше, чем в социологии или в политологии.

⁵ Кстати сказать, именно это представление о человеческом поведении лежит в основе некоторых аргументов, критикующих плановое хозяйство: начиная с Ф. Хайека, экономисты полагают, что единственным способом обработать большие массивы информации в экономической системе является спонтанная координация игроков в рыночной экономике; любые попытки централизованного сбора информации обречены на провал.

действиям, которые игроки предпринимают в рамках экспериментов, причём игроки должны обязательно получать реальное денежное вознаграждение по итогам своих решений.

Подход «text-as-data» в этом отношении представляет собой принципиальный разрыв с традициями экономической науки — ведь в центре его внимания находится по определению как раз то, что экономисты традиционно отказывались изучать — тексты, т.е. акты коммуникации экономических агентов. С точки зрения более традиционных подходов к экономической науке «text-as-data» представляет собой изучение огромных массивов «cheap talk». Собственно говоря, в этом и состоит проблема для экономистов — с одной стороны, постоянно возрастающий объём текстовой информации делает её изучение для дисциплины, традиционно воспринимающей себя как «самая количественная», весьма привлекательным, а с другой — не очень понятно, как интерпретировать полученные выводы с позиций строгих критериев подходов, основанных на выявленных предпочтениях.

Это противоречие может быть разрешено разными способами. С одной стороны, существует набор тем, в которых экономистов действительно интересуют слова — просто в силу того, что в центре их внимания находится как раз вопрос о реакции экономических агентов на слова других агентов (и если есть такая реакция, то каким образом она проявляется). Для исследователей финансовых рынков может быть интересным влияние коммуникации компаний (их пресс-релизов и отчётов) на изменение котировок; для исследователей макроэкономической политики — влияние словесных интервенций центральных банков на поведение рынков. Для поведенческой экономики исследование потенциальной значимости «cheap talk» и склонности экономических агентов всё же реагировать на казалось бы не подкреплённую никакими действиями коммуникацию в принципе является достаточно важной темой [Charness, Grosskopf, 2004]. В этих условиях использование методов «text-as-data» вполне приемлемо для большинства экономистов.

С другой стороны, «text-as-data» могут использоваться как способ генерации новых массивов данных, описывающих экономическую систему как таковую, т.е. не как инструмент анализа, а как способ получения количественных данных там, где они до этого отсутствовали. Например, в работе [Baker, Bloom, Davis, 2016] авторы разрабатывают новый индекс экономической неопределённости, основанный на анализе новостей в СМИ. В работе [Hoberg, Lewis, 2017] показывается, что риторика текстовой части финансовой отчётности фирм, публикующих недостоверную информацию, отличается от риторики «добросовестных» фирм. «Text-as-data» могут применяться для оценки масштабов эпидемий в отсутствие адекватных данных статистики, например на основе поисковых запросов в Интернете. В этой ситуации «text-as-data» используется как источник информации об аспектах функционирования экономической системы, для которых в противном случае количественные данные в принципе отсутствовали бы.

Тем не менее, несмотря на то, что, вне всякого сомнения, существуют примеры областей, в которых экономисты вполне открыты к применению методов «text-as-data», проблема соотношения этого подхода с традиционным акцентом на «дела», а не «слова» экономических агентов [Libman, 2015] остаётся. «Text-as-data» ставит экономистов перед выбором: или пересмотреть одно из традиционных положений их дисциплины, которое они считали её преимуществом, или отказаться от использования больших массивов данных, в принципе доступных для анализа.

Текст-как-данные и революция достоверности

Как уже упоминалось выше, наверное, ключевым трендом последних лет в эмпирической экономической науке является так называемая революция достоверности — представление об эксперименте как золотом стандарте экономических исследований,

позволяющем выявить причинно-следственные связи между переменными. Такое представление о критериях качества экономики существенным образом повлияло на практику работы экономистов. Большинство из них вместо больших межстрановых исследований, столь типичных для экономики прошлого, основное внимание уделяет поиску и исследованию конкретных, порой весьма специфических ситуаций, в которых или сам экономист может реализовать экспериментальное исследование (полевой эксперимент), или доступные данные могут интерпретироваться как результат эксперимента (естественный эксперимент). Такая тенденция нередко подвергалась критике: результатом становится крайне фрагментированная дисциплина, в которой экономисты часто изучают весьма специфические кейсы, единственным преимуществом которых является их соответствие строгим критериям «революции достоверности».

Конечно, методы «text-as-data» могут применяться в самых различных ситуациях, в том числе и в экспериментах. Однако в общем и целом сам подход, лежащий в основе «text-as-data», противоречит логике революции достоверности — фокус лежит на исследовании доступных массивов данных (и на выявлении закономерностей в них), а не специфических ситуаций, в которых корреляция переменных может интерпретироваться как причинно-следственная связь. Как уже упоминалось, «text-as-data» концентрируется на описании больших массивов текстов, т.е. на чисто дескриптивных задачах, а также на выявлении корреляции между теми или иными характеристиками текстов и их авторов или читателей⁶. Естественно, в политологии и в социологии существует большое число экспериментальных работ, посвящённых анализу влияния тех или иных текстов (риторики, пропаганды или рекламы) на поведение людей — речь идёт о так называемых «опросных экспериментах» (survey experiments)⁷. Но для их проведения не требуется применение методов «text-as-data»: ведь исследователям не нужно сводить к набору переменных большие корпуса текстов (достаточно понять, как конкретные тексты, выбранные ими, влияют на поведение или ответы участников опросных экспериментов).

Таким образом, «text-as-data» представляет собой с точки зрения развития экономической науки отход от принципа использования экспериментов как «критериев качества» количественных исследований. Наоборот, количественные исследования в каком-то смысле возвращаются к своим корням, когда главным фактором их реализации была доступность данных (например, статистической информации), а не наличие интересной и убедительной стратегии идентификации (т.е. выявления причинно-следственных связей). Огромные массивы текстовой информации (например, материалы социальных сетей) содействуют расцвету скорее дескриптивных исследований, чем работ, в центре внимания которых находятся причинно-следственные связи. Интересно, впрочем, что эта проблематика не вполне осознаётся самими экономистами — так, в статье [Currie, Kleven, Zwiars, 2020] развитие методов формального анализа текстов и «революция достоверности» рассматриваются как неразрывно связанные явления⁸.

Противоречия между исследованиями, опирающимися на «text-as-data», и революцией достоверности носят даже более фундаментальный характер, чем противоречия, связанные с отказом от акцента на изучении выявленных предпочтений. Последние десятилетия стали

⁶ К тому же сами по себе методы «text-as-data», сводящие текст к набору переменных, могут порождать специфические проблемы для причинно-следственного анализа [Egami et al., 2022].

⁷ В отличие от типичных для экономики лабораторных или полевых экспериментов, в опросном эксперименте участники сталкиваются с *гипотетическим выбором* — например, какой вариант они предпочли бы в той или иной ситуации. Конечно, с точки зрения ортодоксального экономиста речь идёт о типичном «cheap talk».

⁸ Интересно, что методы «text-as-data» находят своё применение и в исследовании масштабов самой революции достоверности — насколько её элементы получили распространение в научной литературе [Goldsmith-Pinkham, 2024], — а также других аспектов риторики экономистов (скажем, роли политических предпочтений исследователей [Jevleh, Kogut, Naidu, 2024]).

периодом фрагментации теоретических основ экономического мейнстрима, неотъемлемой частью которого стала, например, поведенческая экономика [Colander, Holt, Rosser, 2004]. С этой точки зрения представление о незначимости словесной коммуникации в экономической науке далеко не столь универсально, как ранее. Те же опросные эксперименты постепенно становятся всё более значимым инструментом и для экономистов, а не только для политологов и социологов. Р. Шиллер считает исследование нарративов (т.е., по сути дела, словесной коммуникации и текстов, которые как раз и являются предметом «text-as-data») важным аспектом современного экономического анализа [Shiller, 2017; Shiller, 2020]. Однако в ситуации, когда теоретические основы экономической науки делаются более плюралистичными и расплывчатыми, методология экономики, основанная на революции достоверности, становится ещё более важной общей характеристикой дисциплины, в каком-то смысле определяющей её границы. Отказ от экспериментов (и схожих методов) в пользу количественного описания (quantitative description) оказывается вызовом для такого самоопределения экономики — и пока не понятно, каким образом она будет с этим вызовом справляться.

Заключение

Растущая роль методов «text-as-data» в современных общественных науках (в том числе в экономике), вне всякого сомнения, обогащает эти дисциплины как за счёт введения в научный оборот новых данных, так и в связи с появлением возможностей формального анализа данных, ранее доступных исключительно для качественного исследования. В то же время для нас «text-as-data» интересны не просто как метод, а как пример гораздо более сложного характера прогресса общественных наук (и экономической науки), чем тот, который часто постулируется в литературе (и имплицитно признаётся самими экономистами).

Методы «text-as-data», с одной стороны, представляют собой *количественную*, статистическую методологию, обеспечивающую осмысление ещё одного аспекта социальной реальности с использованием количественных данных (и в этом отношении они полностью соответствуют амбициям экономической науки). Но с другой стороны, с точки зрения основных направлений прогресса экономической науки «text-as-data» представляют собой разрыв с господствовавшими на протяжении десятилетий представлениями: вместо фокусирования на выявленных предпочтениях и экспериментальных методах они предполагают акцент на словесной коммуникации, текстах и количественном описании. Распространение «text-as-data», более успешных в политологических исследованиях, чем в экономических, также отчасти меняет традиционную динамику диффузии количественных методов в общественных науках, обычно начинающейся с экономики и позднее распространяющейся на другие дисциплины.

Пример «text-as-data» показывает, что кажущиеся неизменными и очевидными тенденции развития методологии часто не являются таковыми. Они меняются под воздействием или изменений в социальной реальности (в данном случае — появления новых массивов данных), или новых методов, которые оказываются привлекательными для исследователей. Возможно, впоследствии эти трансформации методологии будут рационализироваться в рамках того или иного исторического нарратива, представляющего эволюцию дисциплины как логичный прогресс от более примитивных к более сложным и достоверным методам анализа. Однако речь идёт именно о рационализации *ex-post*, а не о самой динамике развития дисциплины. Собственно говоря, «text-as-data» не являются уникальным примером таких изменений. Так, сама революция достоверности была основана на отказе от ряда предшествовавших тенденций развития экономической методологии⁹. Скорее всего, такие неожиданные повороты будут встречаться и в будущем.

⁹ Речь идёт, в частности, о методах отбора переменных на основе статистических методов [Leamer, 2010].

ЛИТЕРАТУРА/REFERENCES

- Либман А. (2024). «Зоопарк» эмпирических результатов: количественные исследования и прирост знания об обществе [Libman A. (2024). Zoo of Empirical Results: Quantitative Research and Accumulation of Knowledge in Social Sciences] // *Журнал Новой экономической ассоциации*. №4. С. 178–194. DOI: 10.31737/22212264_2024_4_178-194.
- Ash E., Hansen S. (2023). Text algorithms in economics // *Annual Review of Economics*. Vol. 15. No. 1. Pp. 659–688. DOI: 10.1146/annurev-economics-082222-074352.
- Ash E., Gauthier G., Widmer P. (2024). Relatio: Text semantics capture political and economic narratives // *Political Analysis*. Vol. 32. No. 1. Pp. 115–132. DOI: 10.1017/pan.2023.8.
- 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*. Vol. 16. No. 1. Pp. 1–18. DOI: 10.1080/19312458.2021.2015574.
- Baker S.R., Bloom N., Davis S.J. (2016). Measuring economic policy uncertainty // *Quarterly Journal of Economics*. Vol. 131. No. 4. Pp. 1593–1636. DOI: 10.1093/qje/qjw024.
- Benoit K. (2020). Text as data: An overview // *The SAGE handbook of research methods in political science and international relations* / L. Curini, R. Franzese (eds.) Pp. 461–497. DOI: 10.4135/9781526486387.n29.
- Boulier B.L., Goldfarb R.S. (1998). On the use and nonuse of surveys in economics // *Journal of Economic Methodology*. Vol. 5. No. 1. Pp. 1–21. DOI: 10.1080/13501789800000001.
- Brady H.E. (2019). The challenge of big data and data science // *Annual Review of Political Science*. Vol. 22. No. 1. Pp. 297–323. DOI: 10.1146/annurev-polisci-090216-023229.
- Charness G., Grosskopf B. (2004). What makes cheap talk effective? Experimental evidence // *Economics Letters*. Vol. 83. No. 3. Pp. 383–389. DOI: 10.1016/j.econlet.2003.12.007.
- Colander D., Holt R., Rosser Jr. B. (2004). The changing face of mainstream economics // *Review of Political Economy*. Vol. 16. No. 4. Pp. 485–499. DOI: 10.1080/0953825042000256702.
- Crawford I., De Rock B. (2014). Empirical revealed preference // *Annual Review of Economics*. Vol. 6. No. 1. Pp. 503–524. DOI: 10.1146/annurev-economics-080213-041238.
- Currie J., Kleven H., Zwieters E. (2020). Technology and big data are changing economics: Mining text to track methods // *AEA Papers and Proceedings*. Vol. 110. Pp. 42–48.
- Egami N., Fong C.J., Grimmer J., Roberts M.E., Stewart B.M. (2022). How to make causal inferences using texts // *Science Advances*. Vol. 8. No. 42. P. eabg2652. DOI: 10.1126/sciadv.abg2652.
- Einav L., Levin J. (2014). Economics in the age of big data // *Science*. Vol. 346(6210), 1243089. DOI: 10.1126/science.1243089.
- Farrell J., Rabin M. (1996). Cheap talk // *Journal of Economic Perspectives*. Vol. 10. No. 3. Pp. 103–118. DOI: 10.1257/jep.10.3.103.
- Gemenis K. (2013). What to do (and not to do) with the comparative manifestos project data // *Political Studies*. Vol. 61(1_suppl). Pp. 3–23. DOI: 10.1111/1467-9248.12015.
- Gentzkow M., Kelly B., Taddy M. (2019). Text as data // *Journal of Economic Literature*. Vol. 57. No. 3. Pp. 535–574. DOI: 10.1257/jel.20181020.
- Gilardi F., Alizadeh M., Kubli M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks // *Proceedings of the National Academy of Sciences*. Vol. 120(30). e2305016120. DOI: 10.1073/pnas.2305016120.
- Goldsmith-Pinkham P. (2024). *Tracking the credibility revolution across fields*. arXiv preprint arXiv:2405.20604. DOI: 10.48550/arXiv.2405.20604.
- Grimmer J., Stewart B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts // *Political Analysis*. Vol. 21. No. 3. Pp. 267–297. DOI: 10.1093/pan/mps028.
- Hoberg G., Lewis C. (2017). Do fraudulent firms produce abnormal disclosure? // *Journal of Corporate Finance*. Vol. 43. Pp. 58–85. DOI: 10.1016/j.jcorpfin.2016.12.007.
- Imbens G.W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics // *Journal of Economic Literature*. Vol. 58. No. 4. Pp. 1129–1179. DOI: 10.1257/jel.20191597.
- Jelveh Z., Kogut B., Naidu S. (2024). Political language in economics // *Economic Journal*. Vol. 134(662). Pp. 2439–2469. DOI: 10.1093/ej/ueae026.
- Lazer D., Radford J. (2017). Data ex machina: introduction to big data // *Annual Review of Sociology*. Vol. 43. No. 1. Pp. 19–39. DOI: 10.1146/annurev-soc-060116-053457.
- Leamer E.E. (2010). Tantalus on the Road to Asymptopia // *Journal of Economic Perspectives*. Vol. 24. No. 2. Pp. 31–46. DOI: 10.1257/jep.24.2.31.
- Libman A. (2015). Words or deeds: what matters? On the role of symbolic action in political decentralization // *Empirical Economics*. Vol. 49. Pp. 801–838. DOI: 10.1007/s00181-014-0893-8.
- Linegar M., Kocielnik R., Alvarez R.M. (2023). Large language models and political science // *Frontiers in Political Science*. Vol. 5. 1257092. DOI: 10.3389/fpos.2023.1257092.
- Loughran T., McDonald B. (2020). Textual analysis in finance // *Annual Review of Financial Economics*. Vol. 12. No. 1. Pp. 357–375. DOI: 10.1146/annurev-financial-012820-032249.

- Oshchepkov A., Shirokanova A. (2022). Bridging the gap between multilevel modeling and economic methods // *Social Science Research*. Vol. 104. 102689. DOI: 10.1016/j.ssresearch.2021.102689.
- Proksch S.O., Slapin J.B. (2009). How to avoid pitfalls in statistical analysis of political texts: The case of Germany // *German Politics*. Vol. 18. No. 3. Pp. 323–344. DOI: 10.1080/09644000903055799.
- Schumpeter J. (1933). The common sense of econometrics // *Econometrica*. Vol. 1. No. 1. Pp. 5–12. DOI: 10.2307/1912225.
- Shiller R.J. (2017). Narrative economics // *American Economic Review*. Vol. 107. No. 4. Pp. 967–1004. DOI: 10.1257/aer.107.4.967.
- Shiller R.J. (2020). *Narrative economics: How stories go viral and drive major economic events*. — Princeton (N.J.): Princeton University Press.
- Starr M.A. (2014). Qualitative and mixed-methods research in economics: surprising growth, promising future // *Journal of Economic Surveys*. Vol. 28. No. 2. Pp. 238–264. DOI: 10.1111/joes.12004.
- Wilkerson J., Casas A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges // *Annual Review of Political Science*. Vol. 20. No. 1. Pp. 529–544. DOI: 10.1146/annurev-polisci-052615-025542.

Либман Александр Михайлович

alexander.libman@fu-berlin.de

Alexander Libman

PhD (Economics), Professor of Russian and East European Politics, Freie Universität, Berlin

alexander.libman@fu-berlin.de

TEXT-AS-DATA: AN ECONOMICS PERSPECTIVE

Abstract. Formal quantitative methods of text analysis — approaches collectively called “text-as-data” — are playing an increasingly important role in contemporary social sciences. At the same time, while their introduction in political science and partly in sociology is a natural process, largely consistent with the logic of the development of these disciplines, the place of “text-as-data” in economics is ambiguous; while corresponding to the general trends towards mastering an ever-growing range of quantitative methods, “text-as-data” at the same time contradicts some important principles of economics — in particular, attention to revealed preferences rather than rhetoric, and the principles of the “credibility revolution”. The growing role of “text-as-data” demonstrates the complex and nonlinear nature of the progress of quantitative methods in economics (and in the social sciences in general).

Keywords: *text-as-data, reliability revolution, revealed preferences.*

JEL: B40, B41, N10, Z13.