М. Н. Архипова, Д. Ш. Калхиташвили, О. А. Чагадаева

# ОПЫТ ПРИМЕНЕНИЯ ПЛАТФОРМЫ *POLYANALYST* ДЛЯ АНАЛИЗА БАЗЫ ДАННЫХ АНТРОПОЛОГИЧЕСКИХ ИНТЕРВЬЮ*

doi: 10.30759/1728-9718-2024-3(84)-142-152 УДК 39:004.42 ББК 63.5+32.972

Статья посвящена исследовательскому опыту применения инструментов аналитической платформы *PolyAnalyst* в области истории и социальной антропологии. *PolyAnalyst* — первая российская аналитическая платформа с интуитивно-понятным интерфейсом, которая предоставляет доступ к машинной аналитике данных пользователям, не владеющим навыками программирования. Такой функционал, по мнению авторов статьи, позволяет рассматривать *PolyAnalyst* в качестве перспективного вспомогательного метода в гуманитаристике. Авторы протестировали алгоритмы текстовой аналитики, заложенные в программу *PolyAnalyst*, для решения прикладных и теоретических задач в области социо-гуманитарных наук на материале собственной базы данных «Антропологические интервью с жителями Москвы и Подмосковья: “Повседневность 1990-х”». База данных, включающая в себя 50 глубинных полуструктурированных интервью, была создана авторами статьи в рамках работы над исследованием «Социально-экономическая трансформация России в 1987–1999 гг. Между проектами реформ и социальной реальностью». Промежуточные выводы этого исследования легли в основу гипотезы, которую авторы попытались проверить при помощи машинной аналитики. Авторы разъясняют методы компьютерной обработки естественных языков, анализируют конкретные методы и инструменты используемой программы. Рассматриваются плюсы и минусы работы на платформе, предлагаются пути улучшения алгоритмов и методов работы с источниками по социальной антропологии и истории. Делаются выводы о целесообразности и перспективах применения данной программы для решения исследовательских задач социально-гуманитарных наук.

Ключевые слова: *цифровые методы в гуманитарных науках, платформа PolyAnalyst, история повседневности, социальная антропология, 1990-е гг.*

В современной гуманитарной науке все чаще встает вопрос о целесообразности и необходимости использования компьютерных технологий. Многие ученые скептически относятся к перспективе анализа собственных документальных и полевых материалов с помощью программ, предпочитая классические методы, получившие одобрение академического сообщества.

Однако в связи с повсеместной компьютеризацией и цифровизацией всех сфер повседневной жизни гуманитарии, в частности историки и социальные антропологи, не могут игнорировать появление новых источников информации — дигитальных (цифровых). Значение этих источников для изучения современного общества невозможно переоценить. Каждый пользователь сети Интернет сознательно или неосознанно оставляет исследователям особый тип нарративного источника — «цифровой след», будь то информационные блоки (так называемые посты), видео в популярных пабликах (информационно-развлекательных сообществах) и в социальных сетях или комментарии к ним. Большой объем данных — тысячи таких комментариев к интернет-постам, а также и классические расшифровки интервью, которые могут занимать несколько сотен страниц печатного текста, — заставляют исследователей-гуманитариев все чаще смотреть в сторону специальных компьютерных

*Архипова Марьяна Николаевна* — к.и.н., старший преподаватель, Московский государственный университет; с.н.с. Центра прикладной истории Института общественных наук, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации (г. Москва)
E-mail: *marta_ko@mail.ru*

*Калхиташвили Давид Шалвович* — ассистент кафедры системного анализа и анализа данных, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации (г. Москва)
E-mail: *davidkalkhitashvili@gmail.com*

*Чагадаева Ольга Александровна* — к.и.н., доцент кафедры истории, ведущий эксперт Лаборатории цифровых методов в социально-гуманитарных науках Института фундаментальных проблем социо-гуманитарных наук, Национальный исследовательский ядерный университет «МИФИ» (г. Москва)
E-mail: *olenushk@mail.ru*

* *Исследование проводится в рамках НИР «Социально-экономические реформы в контексте эволюции социальной структуры российского общества (1996–2004 гг.)»*

технологий. Однако на сегодняшний день среди исследователей нет единого мнения относительно не только возможности и целесообразности изучения дигитальных источников, но и применения компьютерных программ в решении социо-гуманитарных исследовательских задач. Это во многом связано с междисциплинарными различиями: специалисту-гуманитарию без специальной подготовки трудно разобраться в информационных технологиях и терминологии специалистов в сфере IT. К счастью, сегодня появляются программы с понятным пользовательским интерфейсом, которые освобождают гуманитариев от необходимости постигать азы программирования. Однако у этих программ есть существенный недостаток: они чрезвычайно дороги, что значительно сокращает круг потенциальных пользователей. Покупка лицензий заграничных программ сейчас недоступна российским учреждениям, однако и отечественные аналоги слишком дорого обходятся вузам и научно-исследовательским центрам, что заставляет ежегодно ставить ребром вопрос о целесообразности продления лицензионной подписки.

Однако интерес к новым компьютерным технологиям в гуманитаристике, несомненно, заметен. В зарубежной и отечественной гуманитарной науке все чаще появляются работы о перспективах *Digital Humanities*.[1] Так, в 2023 г. вышло учебное пособие «Практики анализа качественных данных в социальных науках».[2] В нем помимо многочисленных методов качественного исследования, описывается работа с помощью программного обеспечения, например *ATLAS.ri*, *NVivo*, *Phyton* и др. — эти программы зарубежного производства, то есть имеют понятные ограничения в использовании их на русскоязычном материале. Авторы делают неутешительный вывод о том, что каждая из описываемых программ имеет свои подводные камни и самостоятельно гуманитариям в них разобраться трудно, а потому рекомендуют обратиться к коллегам-программистам, которые могут стать «своеобразным проводником»[3] в работе с конкретной программой.

Российская программа *PolyAnalyst* пока не широко применяется отечественными научными учреждениями и вузами не в последнюю очередь из-за дороговизны, поэтому материалов по работе с ней в историографии достаточно мало.[4] Наша статья отчасти призвана восполнить этот пробел.

Авторы статьи проанализировали собственный опыт работы с программой *PolyAnalyst* на материалах исследовательского проекта, посвященного социально-экономической трансформации России в 1987–1999 гг. в столичном регионе. В ходе полевой антропологической работы было собрано 50 глубинных интервью о повседневной жизни жителей московской агломерации в указанный период, расшифровки которых представляют более 500 страниц печатного текста. Указанные интервью были зарегистрированы как база данных под названием «Антропологические интервью с жителями Москвы и Подмосковья: “Повседневность 1990-х”».[5] Исследователи также провели большую работу по ручному сбору данных «цифрового следа», имеющих отношение к указанной проблематике — более 2 000 интернет-постов и комментариев к ним в популярных сообществах и социальных сетях. Однако для того, чтобы составить представление, насколько цифровая платформа способна решать задачи классической социальной антропологии, было принято решение в данной статье ограничиться анализом 50 глубинных полуструктурированных интервью.[6]

Несколько слов о дизайне исследования. Было решено выделить определенную группу респондентов для оптимизации решения исследовательской задачи. Это граждане РФ, рожденные в РСФСР (18 мужчин и 32 женщины 1947–1973 г. р.), проживающие ныне и проживавшие

[1] См.: Zhang Q., Segall R. S. Review of Data, Text and Web Mining Software // Kybernetes. 2010. Vol. 39, no. 4. P. 625–655; Cosgrave M. Digital Humanities Methods as a Gateway to Inter and Transdisciplinarity // Global Intellectual History. 2021. Vol. 6, iss. 1. P. 24–33; Антопольский А. Б. Европейский опыт организации цифровой инфраструктуры для социальных и гуманитарных наук // Информационные ресурсы России. 2021. № 4 (182). С. 12–19; Лаптева М. А., Гордеева Е. А. Digital Humanities в России: перспективы развития // Прикладная информатика. 2018. Т. 13, № 1 (73). С. 44–51; Попова С. М. Анализ отечественного и зарубежного опыта развития цифровой инфраструктуры социально-гуманитарных исследований // Genesis: исторические исследования. 2015. № 1. С. 208–251.

[2] Практики анализа качественных данных в социальных науках: учеб. пособие. М., 2023.

[3] Там же. С. 25.

[4] См.: Петров Е. Ю., Саркисова А. Ю. Ресурс аналитической платформы PolyAnalyst в социогуманитарных научных исследованиях // Открытые данные — 2021: материалы форума. Томск, 2021. С. 94–104; Kiselev M. V. PolyAnalyst — A Machine Discovery System Inferring Functional Programs // AAAI Technical Report. AAAI. AAAI-94 Workshop on Knowledge Discovery in Databases (WS-94-03). Retrieved 15 March. 2021. P. 237–249.

[5] Что не вполне точно отражает временные рамки исследования (1987–1999 гг.).

[6] Антропологические интервью с жителями Москвы и Подмосковья: «Повседневность 1990-х». База данных. Сост. Архипова М. Н., Головина А. В., Голечкова О. Ю., Чагадаева О. А.

в 1990-е гг. в Москве и Подмосковье. То есть это экономически и политически активная в 1990-е гг. часть населения, сформировавшаяся в позднем СССР; носители достаточно конкретных воспоминаний о быте и повседневных практиках столичного региона. Треть информантов к началу 1990-х гг. училась в вузах, остальные получили высшее образование и начали трудовую деятельность еще до распада СССР. 33 респондента получили техническое образование, 17 — гуманитарное или художественное. В силу меняющейся экономической ситуации многие собеседники меняли сферы деятельности, порой кардинально.

Такое ограничение круга респондентов позволило проанализировать повседневность более-менее однородной социальной категории и прийти к следующим промежуточным выводам. Реперными точками в воспоминаниях респондентов становились события частной жизни — рождение ребенка, смена работы, свадьба и т. д. Кардинальные социально-экономические, структурные перемены, происходившие в стране, фактически оставались лишь фоном течения повседневной жизни. В воспоминаниях отсутствует четкая привязка к ключевым реформам правительства и важнейшим политическим событиям. Общим местом всех интервью выступает разве что ваучерная приватизация, в то время как, например, силовые противостояния 1991 и 1993 гг. по прошествии лет сливаются у многих в одно событие. Период перестройки и «лихих 1990-х» респонденты практически не разграничивают и оценивают как единый период больших перемен: изменения социальной структуры общества и поисков себя в этой структуре, потери социально-профессиональной идентичности, появления серьезных материальных трудностей и вместе с тем чувства ответственности за свое материальное благополучие, свободы слова, открытия границ и появления зарубежных товаров и т. д. Для того чтобы проверить эту гипотезу, мы воспользовались инструментарием аналитической платформы *PolyAnalyst*, предполагая, что программа позволит более глубоко проанализировать имеющиеся тексты.

### *Функционал платформы PolyAnalyst*

Итак, рассмотрим аналитическую платформу *PolyAnalyst* как вспомогательный метод для решения социо-гуманитарных задач. Платформа (программа) представляет собой метод компьютерной обработки естественных языков, в англоязычных источниках известный как *Natural Language Processing* (далее — *NLP*). Сама по себе технология *NLP* позволяет обрабатывать большие массивы данных для выявления тех или иных закономерностей в тексте и включает в себя несколько инструментов, позволяющих упростить процесс анализа корпусов текста: 1) предварительную обработку текста; 2) статистический анализ текстового корпуса; 3) создание семантического ядра текстового корпуса; 4) анализ тональности слов или словосочетаний; 5) выделение ключевых фраз и слов из текста; 6) визуализацию данных.

Для понимания принципов работы программы *PolyAnalyst* с текстовыми данными кратко рассмотрим каждый из них.

### *Предварительная обработка текста*

Вне зависимости от поставленной задачи основным этапом при анализе больших массивов текстовых данных является предварительная обработка текста. Она делится на несколько процессов, называемых «токенизация», «лемматизация» и «векторизация», а также удаление так называемых стоп-слов. Эта обработка необходима для приведения текста в пригодный для машинного анализа вид.

На этапе так называемой токенизации текст разбивается на уникальные единицы текста — токены. Это первый этап анализа текста, который позволяет преобразовать непрерывный текст в дискретные элементы, с которыми будет работать программа.

Каждое отдельное слово, предлоги, а также знаки препинания получают свой уникальный счетчик токенов. Этот счетчик необходим для того, чтобы к текстовому массиву данных можно было применить методы статистических исследований: для создания семантического ядра слов, выделения ключевых слов и фраз, а также выделения фактов из текста. Так, предложение «*Сейчас я думаю о том, что в 90-е годы были у меня, конечно, трудности, но их не сравнить с теми проблемами, которые были у других людей*»,[7] разбивается на отдельные токены: «*сейчас*», «*я*», «*думаю*», «*о*», «*том*», «*,*», «*что*», «*в*», «*90-е*», «*годы*», «*были*», «*у*», «*меня*», «*,*», «*конечно*», «*,*», «*трудности*» и т. д.

Следующим этапом предварительной обработки текста является лемматизация, представляющая собой сокращение слов до основы

[7] Антропологические интервью с жителями Москвы и Подмосковья: «Повседневность 1990-х». База данных...

смысловой конструкции — леммы. Таким образом, к примеру, слова «*перестройка*» и «*стройка*» будут иметь общую лемму «*строй*». Так программа сможет находить повторяющиеся слова, которые отличаются только словоформой. Метод лемматизации, по опыту, может приводить как к повышению, так и к ухудшению показателей результатов анализа текста. Эти изменения связаны с особенностями грамматического строения русского языка. Так, прибавление суффиксов и предлогов может значительно изменять смысловое значение исходного слова, и при сведении данных слова до его леммы происходит смещение смыслов, что может сказываться на качестве анализа больших текстовых массивов данных в социально-гуманитарных науках. Поэтому метод лемматизации имеет смысл применять для определения контекста или тематики текстового массива. Определение тематики текстового массива можно отнести к более высокоуровневой задаче, то есть более поверхностной, но и более обширной по отношению к выявлению закономерностей в тексте на основе лемм слов. Вместе с тем при анализе интервью применение метода приведения слов к их леммам может позволить определить контекст, основные события, факты и процессы, на которых информант делал акцент и какие слова он использовал при описании данных процессов.

Метод векторизации схож с лемматизацией, но в процессе векторизации слова сокращаются до суффикса, то есть сохраняется грамматическая основа слова, урезается только окончание. Затем строится некоторая матрица со значениями, и каждое отдельное слово (токен) получает своё представление на координатной плоскости в виде комплексных чисел.

Наконец, последний из перечисленных методов предварительной обработки текстовых массивов данных — выделение стоп-слов. Данный метод применяется с целью повышения качества статистического анализа текстового массива данных путем очищения текста от предлогов, междометий и прочих частей речи, не несущих смысловой нагрузки, но статистическое значение которых в связи с частотой использования велико. В противном случае предлоги и междометия попадут в семантическое ядро, что приведет к изменению ключевых слов. Так, в процитированном выше фрагменте одного из интервью «*Сейчас я думаю о том, что в 90-е годы были у меня, конечно, трудности, но их не сравнить с теми проблемами, которые были у других людей*», алгоритм очистит предложение от токенов «*я*», «*о*», «*,*», «*в*», «*у*», «*но*», «*их*», «*не*», «*с*».

*Статистический анализ текстового корпуса* позволяет выделить статистическое распределение тех или иных слов в самом текстовом массиве данных. Иными словами, настоящий метод позволяет рассчитать количественное значение того или иного токена в тексте. Полученные количественные значения позволяют построить семантическое ядро текста, выделить ключевые слова и фразы, выделить сущности, определить факты и тональность слов. Обычно в тексте статистически превалируют токены-местоимения, однако последующие алгоритмы, о которых речь пойдет ниже, позволяют очистить статистику от подобных слов, лишенных смысловой нагрузки.

*Создание семантического ядра* — это процесс определения и выделения главных слов в тексте, то есть таких слов, которые наполняют текстовый массив данных смысловой нагрузкой. Семантическое ядро имеет общие характеристики с ключевыми словами, но существует основное отличие: семантическое ядро включает в себя, помимо существительных и глаголов, причастия и прилагательные. То есть кроме самих объектов и их действий в семантическом ядре можно увидеть ключевое описание объектов и действий. Так, уже не раз упомянутое в качестве примера предложение «*Сейчас я думаю о том, что в 90-е годы были у меня, конечно, трудности, но их не сравнить с теми проблемами, которые были у других людей*» после очищения от стоп-слов и слов, лишенных смысловой нагрузки, применения алгоритмов выделения ключевых слов и создания семантического ядра будет представлять собой следующие токены: «*сейчас*», «*думаю*», «*90-е годы*», «*были*», «*трудности*», «*не сравнить*», «*проблемами*», «*других людей*».

*Анализ тональности слов или словосочетаний позволяет* определять эмоциональную окраску слов, то есть выделять положительные или отрицательные характеристики того или иного объекта/действия (прилагательного к существительному). При помощи тональности слов можно определить отношение субъекта к конкретной ситуации, событию, личности или предмету. Можно достаточно точно определить негативное или положительное, смешанное отношение субъекта к тому или иному внешнему фактору. По своей сути настоящий метод является классификацией слов.

Условно лексику можно поделить на три группы: положительная, нейтральная и отрицательная. Задача определения тональности слов фактически сводится к распределению слов и словосочетаний в соответствии с заранее разработанным словарем на обладающие положительной, нейтральной или отрицательной тональностью. То есть если то или иное слово не внесено в словарь положительных или отрицательных слов, оно автоматически окажется в категории «нейтральные». Поэтому бывают случаи, когда к нейтральным словам могут быть отнесены либо ярко негативные, либо ярко положительные слова, не учтенные в словаре.

Более сложной задачей является отнесение предложения, абзаца или всего текста к положительно или негативно окрашенной единице. Для решения данной задачи каждому слову в словаре присвоено свое весовое значение, показывающее насколько оно негативное или положительное. После все значения складываются: положительные слова имеют под собой натуральные числа, а негативные — отрицательное значение. После сложения данных значений определяется, какой знак стоит перед суммой. Если значение больше нуля, данное предложение (или словосочетание, абзац, текст) определяется положительным, если меньше нуля — отрицательным. Если значение нулевое, то данное предложение определяется нейтральным.

Трудность адекватного анализа тональности при работе программы с антропологическими источниками заключается в том, что программа не умеет считывать скрытые смыслы, иронию, сарказм, игру слов, а потому достаточно часто определяет тональность сложных, эмоционально окрашенных текстов (коими, в частности, являются расшифровки интервью) некорректно. С анализом тональности связана и еще более глубокая проблема, речь о которой пойдет позднее.

Алгоритм программы «*Выделение ключевых слов, словосочетаний, сущностей и фактов*» позволяет выделять слова и пр., которые либо указывают на объект или субъект, либо выделяют действия: слова, которые определяют действующее лицо в тексте, либо лицо, на которое действие направлено. Ключевыми словами могут быть события, имена собственные, личности и иные сущности. Ключевые слова включают в себя сущности и факты, то есть задают основной смысл предложения, чаще всего это имена существительные в именительном падеже или глаголы несовершенного вида. При этом сущность — это некий объект в тексте, слово, имеющее свой уникальный смысл. Данная сущность может выступать в роли как субъекта, так и объекта, а факт — это действие объекта, либо свершенное, либо в настоящий момент происходящее.

Под *визуализацией текстового массива данных* подразумевается построение либо столбчатых, либо круговых диаграмм на основе статистического распределения слов. Основным элементом визуализации текстового массива данных является построение графа на основе связи терминов или статистического распределения слов. При помощи визуализации можно выделить новые закономерности, которые невозможно выстроить человеческим взглядом. Этот прием достаточно информативен и может помочь в решении задач социально-гуманитарных дисциплин. В частности, можно выявить словарный оборот каждого респондента в отдельности и всех собранных интервью в целом, визуализировать статистическое распределение слов внутри текста. Исследователь может увидеть своими глазами, какие слова и словосочетания наиболее часто употребляются респондентами при описании того или иного исторического периода, той или иной проблемы и явления.

Все эти методы обработки текстовых данных выглядят крайне привлекательными для использования при решении задач социально-гуманитарных дисциплин. Возможность в короткий срок обработать огромный массив данных, автоматическое определение ключевых слов, анализ эмоциональной окраски текста и прочие полезные функции программы вплоть до визуализации обещают гуманитариям вывести глубину исследования на новый уровень. Однако на практике мы столкнулись с существенными трудностями, связанными прежде всего с тем, что программа в настоящий момент «заточена» под определенные области, и история и социальная антропология пока не входят в их число.

### *Применение аналитической платформы PolyAnalyst в работе над исследовательским проектом по социальной антропологии*

Вышеописанные методы в аналитической платформе *PolyAnalyst* представлены в виде узлов, или инструментов, представляющих собой запрограммированные блоки.

Как было указано выше, платформа *PolyAnalyst* позиционируется как воплощение концепции демократизации работы с данными и не требует от пользователя даже начальных навыков программирования и *Data Science*. Программа имеет интуитивно понятный интерфейс, что чрезвычайно важно для пользователя с гуманитарным бэкграундом. Система предоставляет экспертам в любой предметной области упрощенные и интуитивно понятные пользовательские интерфейсы к серьезным математическим инструментам, автоматизирующим проведение глубокого и разностороннего анализа данных. То есть исследователи, специализирующиеся на социально-гуманитарных науках, получили возможность использовать алгоритмы, которыми до недавнего времени могли пользоваться только технические специалисты, владеющие языками программирования.

В программу загружаются текстовые материалы в формате *Word* или таблицы *Excel*. Для обработки данных нашего антропологического исследования база данных «Повседневность 1990-х» была переведена в формат *Word*. После загрузки файла пользователь переходит к текстовому анализу, задавая программе алгоритмы при помощи выбора тех или иных узлов.

На рис. 1 представлена рабочая зона с готовым проектом для анализа текстового массива данных. То, что мы видим на изображении, — скрипт, то есть алгоритм, составленный из узлов (инструментов).

Метод обработки текстового массива данных требует настройки тематики, превалирующей в пакете данных. Для корректности исследования перед каждой задачей необходимо самостоятельно составлять словарь либо использовать предварительно установленный. На платформе отсутствует специальный социо-гуманитарный словарь, и мы не имели технической возможности и достаточных временных ресурсов, чтобы составить свой, поэтому в исследовании использовался словарь, предварительно установленный разработчиком. Это, с одной стороны, существенно сократило затраченное время, с другой — использование встроенного словаря заметно снизило качество результатов исследования. Так, программа определила, что загруженные тексты относятся к тематике «экономика», поскольк-

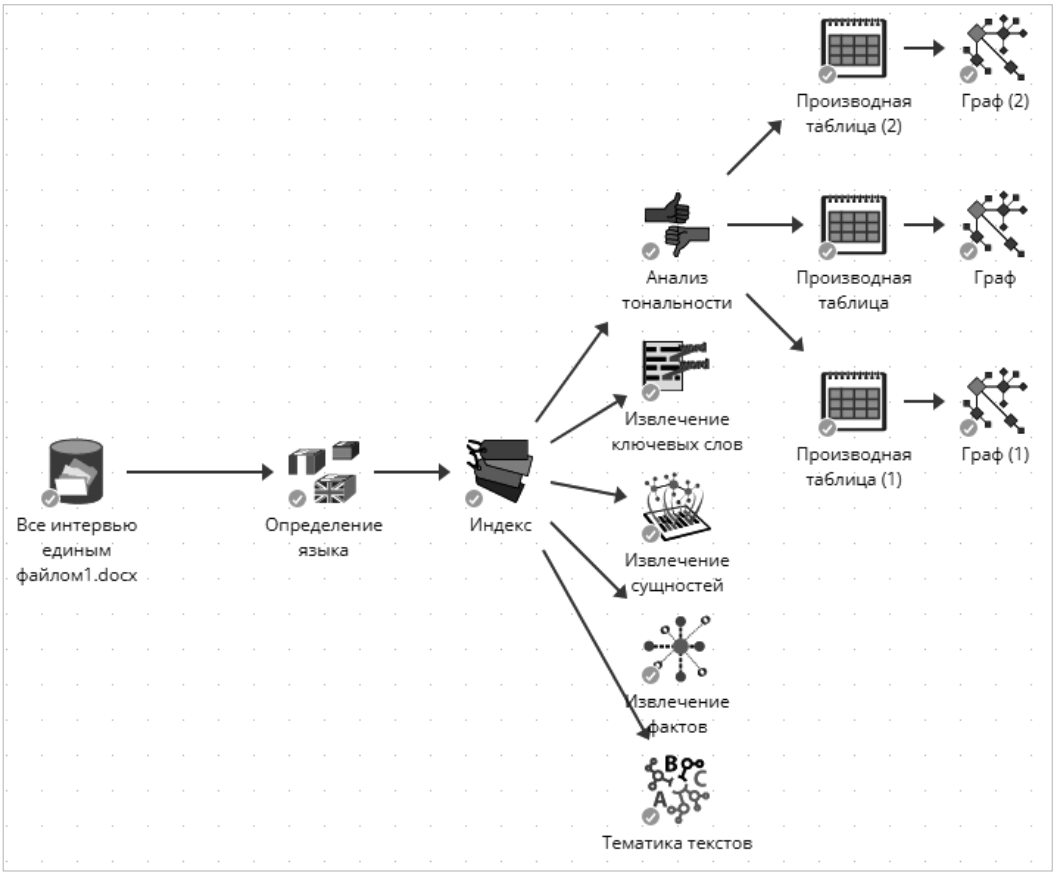



Рис. 1. Готовый проект для реализации задачи социально-гуманитарной направленности

ку, при отсутствии словарей по антропологии и истории, она обнаружила в тексте большое количество лексики, связанной с экономической тематикой. Употребление этой лексики совершенно естественно при исследовании повседневности в условиях глубокой социально-экономической трансформации. В то же время смысловой контекст выявленных программой слов часто далек от экономики и связан исключительно с повседневными практиками. Так, к примеру, слово «магазин», определенное программой как «экономическое» в рамках данного исследования, носило сугубо бытовой характер: респонденты вспоминали, как искали продукты по более низким ценам, как менялось их материальное положение. Так же и курс доллара упоминался респондентами исключительно в контексте непосредственного влияния на повседневные практики (рис. 2).

На этапе выделения ключевых слов и семантического ядра программа выделила из корпуса представленных текстов самые употребляемые и самые значимые в тексте слова. Как видно из рис. 3, в текстовом массиве интервью самым часто употребляемым словом оказался «развал» — в контексте распада Советского Союза. Антропологи осознанно избегали использования слова «развал» применительно к событиям 1991 г., то есть слово употребляется только респондентами, что может многое сказать исследователю об эмоциональном отношении к данному событию и времени.

На втором месте совершенно неожиданно для исследователей оказалось слово «священник». Объяснение можно искать в частоте употребления этого слова рядом респондентов, оказавшихся напрямую связанных с РПЦ в годы социально-экономического транзита. Несмотря на то что многие информанты отмечали повышенный интерес общества к духовным практикам в изучаемый период, на основании этой статистики было бы ошибочно делать выводы об особой важности Православной церкви в повседневной жизни москвичей и жителей Подмосковья в 1990-е гг. То есть частота употребления определенной лексической единицы в одном-двух интервью экстраполируется программой на весь корпус текстов. Этот кейс показывает, что результаты машинной обработки данных и машинного анализа данных должны быть скрупулезно проанализированы исследователями вручную.

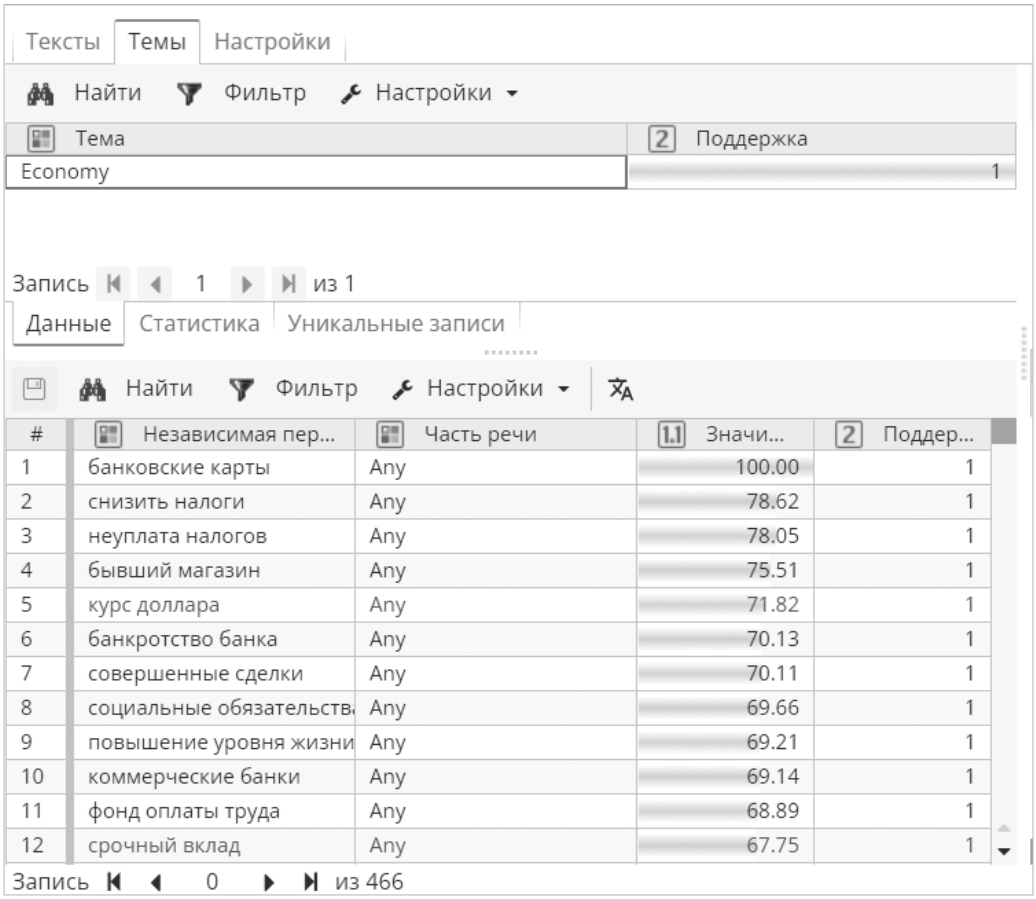
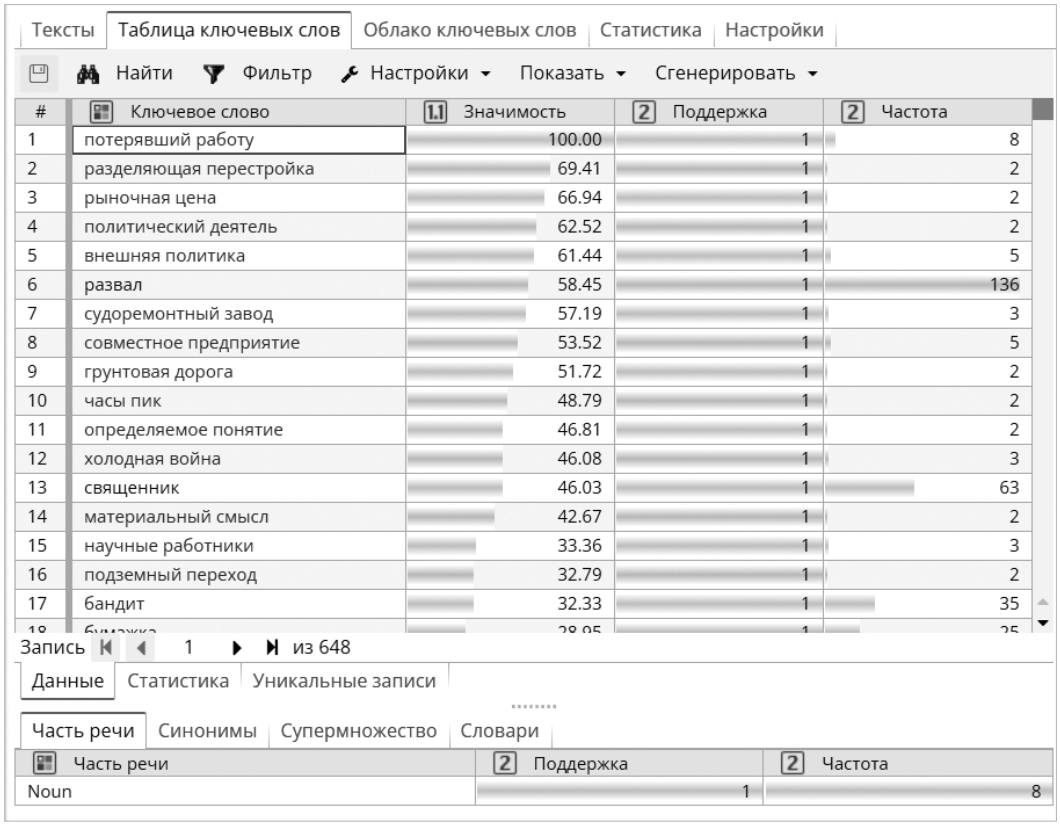



Рис. 2. Определение тематики текста при помощи аналитической платформы *PolyAnalyst*




| # | Ключевое слово | Значимость | Поддержка | Частота |
|---|---|---|---|---|
| 1 | потерявший работу | 100.00 | 1 | 8 |
| 2 | разделяющая перестройка | 69.41 | 1 | 2 |
| 3 | рыночная цена | 66.94 | 1 | 2 |
| 4 | политический деятель | 62.52 | 1 | 2 |
| 5 | внешняя политика | 61.44 | 1 | 5 |
| 6 | развал | 58.45 | 1 | 136 |
| 7 | судоремонтный завод | 57.19 | 1 | 3 |
| 8 | совместное предприятие | 53.52 | 1 | 5 |
| 9 | грунтовая дорога | 51.72 | 1 | 2 |
| 10 | часы пик | 48.79 | 1 | 2 |
| 11 | определяемое понятие | 46.81 | 1 | 2 |
| 12 | холодная война | 46.08 | 1 | 3 |
| 13 | священник | 46.03 | 1 | 63 |
| 14 | материальный смысл | 42.67 | 1 | 2 |
| 15 | научные работники | 33.36 | 1 | 3 |
| 16 | подземный переход | 32.79 | 1 | 2 |
| 17 | бандит | 32.33 | 1 | 35 |

Рис. 3. Таблица ключевых слов.

Полезными для исследователей-гуманитариев оказались результаты анализа значимости слов. Под значимостью программа подразумевает поддержку словосочетания внутри предложения другими дополнительными словами, пояснениями, причастными оборотами, прилагательными, которые поясняют или раскрывают смысл ключевого понятия. Также программа обращает внимание на знаки препинания: так, восклицательный знак в конце предложения будет считываться как усилитель значимости понятия, и оно будет перенесено выше в таблице ранжирования.

По значимости слов в нашем исследовании, согласно программе, на первом месте оказалось словосочетание «потерявший работу». На протяжении практически 500-страничного текста оно употреблялось всего 8 раз, однако имело наибольшую поддержку, то есть программа видит это словосочетание как особо важное для текста. И здесь *PolyAnalyst* со своей задачей справился: выявил неочевидный, но достаточно важный тренд. Информанты тяжело переживали падение социального и материального статуса, депрофессионализацию. Даже если личный повседневный опыт 1990-х гг. оказывался более-менее удачным, все респонденты упоминали истории о маргинализации, суицидальном поведении и просто тяжелых испытаниях, через которые пришлось пройти людям их социального круга. Каждый подобный рассказ был эмоционально окрашен, сопровождался подробным описанием кейса, что обоснованно отразилось в результатах выделения ключевых слов программой.

Также казался достаточно перспективным следующий узел — «*Определение тональности*», то есть статистический анализ положительных и отрицательных слов. Эмоционально положительно окрашенные слова в тексте оказались превалирующими: по подсчетам программы, 652 из 1 033 слов окрашены позитивно, то есть 63 % позитивных к 37 % негативно окрашенных слов. Однако стоит рассказать о некоторых трудностях, которые удалось заметить при обработке полученного результата. Как было указано выше, *PolyAnalyst* выделяет слова, которые общепринято имеют позитивную или негативную окраску, согласно встроенному базовому словарю. При работе с интервью большую роль играют интонации, паузы, жестикуляция — при машинном анализе все эти нюансы теряются. Приведем показательный пример. Одна из информанток с нескрываемой иронией рассказывала о своей преподавательнице по истории КПСС,

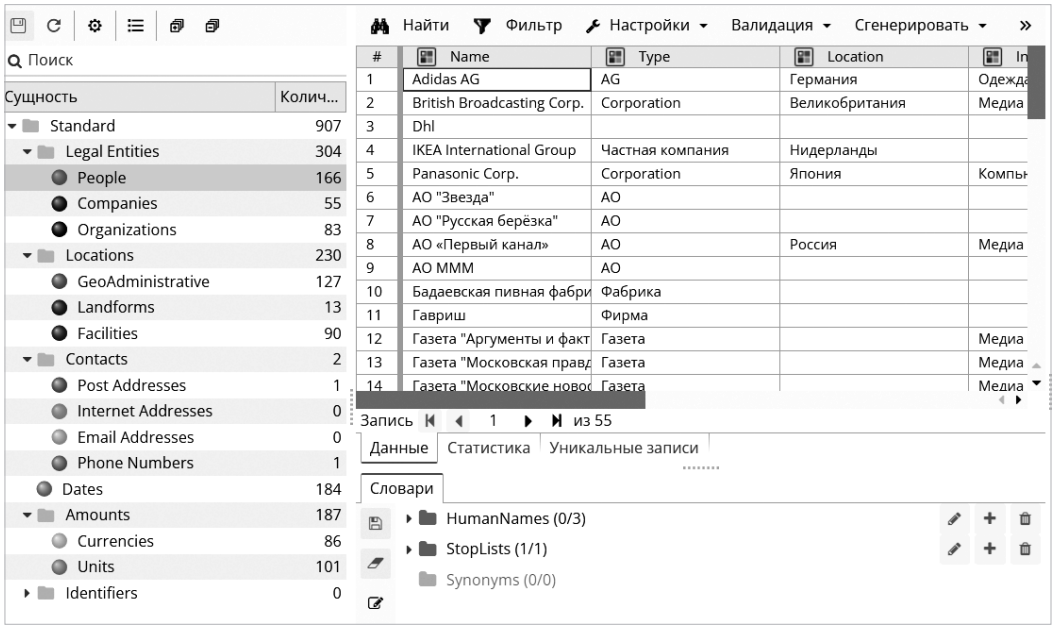



Рис. 4. Частота распределения сущностей

восхищавшейся В. И. Лениным, и произнесла фразу: «Ленин любил фиалки».[8] Сказано это было с сарказмом, однако программа воспринимает эту фразу всерьез и выделяет ее как положительную из-за сильно позитивно окрашенного слова «любить».

Таким образом, мы можем констатировать, что применение встроенного универсального словаря не позволяет достичь точности в исследовании, которая должна присутствовать в научной работе. Только составление и использование специального словаря тональностей для решения каждой конкретной задачи социально-гуманитарных наук позволит добиться большей точности в аналитике. Без этого теряет смысл и визуализация узла «Тональность слов». Попытка такой визуализации на нашем материале привела к построению перегруженного графа, состоящего из сотен слов и совершенно непригодного в качестве инфографики.

Выделение сущностей и фактов позволило получить более точные, с точки зрения количественных показателей, распределения тех или иных объектов. Были использованы два узла — «*Извлечение сущностей*» и «*Извлечение фактов*». С помощью узла «*Извлечение сущностей*» выделяются преимущественно объекты и субъекты, в то время как узел «*Извлечение фактов*» выделяет действия, процессы или состояния объектов и субъектов. Эти алгоритмы обещали существенное сокращение временных затрат на «ручную» обработку интервью. Узел «*Извлечение сущностей*» позволил выявить ключевые объекты и определить количественно их распределение в тексте.

В таблице на рис. 4 представлены виды сущностей, распределение по категориям, количество сущностей, внутренняя единица измерения «*Поддержка*» и последний столбец — количественный показатель, который указывает, сколько раз в тексте встречались сущности из этой категории.

Как видно из рисунка, самой популярной сущностью становится *Dates*, то есть временные указания. В эту сущность входят фразы наподобие «несколько дней назад», «вчера», «много лет назад» и т. д. В нашем историко-антропологическом исследовании такое лидерство абсолютно очевидно, ведь интервью были полны указаниями на конкретные периоды в прошлом. На втором месте оказалась сущность *People* (персоналии), что тоже легко объясняется биографической особенностью интервью: в них много имен известных, в том числе государственных деятелей, и неизвестных, много отсылок к личным историям с конкретными людьми. В связи с тем что глубинные интервью носят личный характер, большое количество употребляемых имен уникальны (имена членов семьи, друзей, клички животных). То же относится и к категории *Organizations* (организации): в текстах встречается большое количество наименований локальных магазинов, бюро, административных учреждений, которые не могли стать статистически значимыми. Однако эти данные,

[8] Антропологические интервью с жителями Москвы и Подмосковья: «Повседневность 1990-х». База данных...

безусловно, имеют самостоятельную ценность, и мы уверены, что эта информация еще ждет своего исследователя. По нашему мнению, в данный момент анализ сущностей может быть более информативным и полезным для анализа отдельного корпуса текстов: статей, книг, монографий. В то же время обилие имен собственных при недостаточно частотном упоминании публичных персон и известных институций подтверждает нашу гипотезу о том, что несмотря на глобальные изменения эпохи социально-экономического транзита, люди продолжали мерить повседневность в категориях частной жизни.

Что касается объединения различных наименований сущности в одну, здесь программа тоже не справляется. Так, токены «Борис Николаевич» и «Борис Ельцин» представлены в статистической таблице как отдельные элементы. Это еще раз говорит о критической необходимости составления тематического словаря если не под каждую исследовательскую задачу, то, во всяком случае, под каждую социально-гуманитарную дисциплину. Базовые словари недостаточно релевантны для научных социо-гуманитарных исследований. Для сравнения, географические названия программа распознает успешно. «США», «Америка» и «Соединенные штаты Америки» были скомпонованы в одну позицию. Так же и «СССР», «Советский Союз» и др. стали одним государством в графах географических названий.

Таким образом, на опыте анализа базы данных антропологических интервью с жителями Москвы и Подмосковья на тему повседневности в период социально-экономического транзита с помощью программы *PolyAnalyst* мы выявили как плюсы, так и минусы использования этой программы для решения академических задач. Несмотря на определенные достоинства программы, которая значительно экономит время исследователя за счет машинного анализа текстов, на сегодняшний день требуется обязательная ручная перепроверка показателей, чтобы избежать случайных ошибок в выводах, которые возможны при опоре на данные аналитики, проведенной программой. Попытка избежать субъективизации исследователем при слепом доверии к программе может привести к достаточно нелепым искажениям, что было показано в статье. Один из перспективных шагов к решению этой проблемы — создание тематических словарей, словарей позитивно, негативно и нейтрально окрашенных слов для каждой из академических дисциплин, а в идеальном случае — указанных словарей под каждую исследовательскую задачу социально-гуманитарного профиля. Однако подобная подготовительная работа к проведению машинного анализа окажется едва ли не более трудоемкой и затратной по времени, чем сам анализ. Сегодня цена годовой подписки на программу *PolyAnalyst* останавливает научные учреждения и тем более индивидуальных исследователей от ее использования, а небольшое число пользователей, в свою очередь, делает создание тематических словарей для ученых социально-гуманитарного профиля невыгодным для разработчика. Исходя из вышесказанного, целесообразность использования данной программы для решения задач социально-гуманитарных наук на данный момент кажется недостаточной. Хочется верить, что в обозримом будущем использование компьютерных технологий станет распространенным методом, который позволит раскрыть новые грани работы с социо-гуманитарными источниками.

***Maryana N. Arkhipova***
Candidate of Historical Sciences, Lomonosov Moscow State University; Russian Presidential Academy of National Economy and Public Administration (Russia, Moscow)
E-mail: *marta_ko@mail.ru*

***David Sh. Kalkhitashvili***
Aspirant, Russian Presidential Academy of National Economy and Public Administration (Russia, Moscow)
E-mail: *davidkalkhitashvili@gmail.com*

***Olga A. Chagadaeva***
Candidate of Historical Sciences, National Research Nuclear University (Russia, Moscow)
E-mail: *olenushk@mail.ru*

## USING THE POLYANALYST PLATFORM TO ANALYZE A DATABASE OF ANTHROPOLOGICAL INTERVIEWS

The article is devoted to the research experience of using the tools of the PolyAnalyst analytical platform in the field of history and social anthropology. PolyAnalyst is the first Russian analytical platform with a translucent interface that provides access to machine analytical data to users who do not have programming skills. This functionality, in our opinion, allows considering PolyAnalyst as a promising auxiliary method in the humanities. The authors tested text analytics algorithms embedded in the PolyAnalyst program to solve applied and theoretical problems in the field of social sciences and humanities using the material from their own database "Anthropological interviews with residents of Moscow and the Moscow region: "Everyday life in the 1990s." The database, which includes 50 in-depth semi-structured interviews, was created by the authors of the article as part of the work on the study "Socio-economic transformation of Russia in 1987–1999. Between reform projects and social reality". The interim findings of this study formed the basis of a hypothesis, which the authors tried to test using machine analytics. The authors explain methods of computer processing of natural languages, analyze specific methods and tools of the program used. The pros and cons of working on the platform are considered, ways to improve algorithmic rhythms and methods of working on the platform are proposed with sources on social anthropology and history. Conclusions are drawn about the feasibility and prospects for using this program to solve research problems in the social sciences and humanities.

Keywords: *digital humanities, PolyAnalyst, history of everyday life, social anthropology, 1990-s in Russia*

## REFERENCES

**A**ntopolsky A. B. [European Experience in Digital Infrastructure for the Social Sciences and Humanities]. *Informatsionnyye resursy Rossii* [Information Resources of Russia], 2021, no. 4 (182), pp. 12–19. DOI: 10.52815/0204-3653_2021_04182_12 (in Russ.).

**C**osgrave M. Digital Humanities Methods as a Gateway to Inter and Transdisciplinarity. *Global Intellectual History*, 2021, vol. 6, iss. 1, pp. 24–33. DOI: 10.1080/23801883.2019.1657639 (in English).

**L**apteva M. A., Gordeeva E. A. [Digital Humanities in Russia: Development Prospects]. *Prikladnaya informatika* [Journal of Applied Informatics], 2018, vol. 13, no. 1 (73), pp. 44–51. (in Russ.).

**P**etrov E. Yu., Sarkisova A. Yu. [Resource of the PolyAnalyst Analytical Platform in Socio-Humanitarian Scientific Research]. *Otkrytyye dannyye — 2021* [Open Data — 2021. Forum Materials]. Tomsk: Izd-vo Tom. gos. un-ta Publ., 2021, pp. 94–104. (in Russ.).

**P**opova S. M. [Analysis of Foreign and Russian Experience in the Development of Digital Infrastructure of Socio-Humanitarian Researches]. *Genesis: istoricheskiye issledovaniya* [Genesis: Historical Research], 2015, no. 1, pp. 208–251. DOI: 10.7256/2409-868X.2015.1.13820 (in Russ.).

***P**raktiki analiza kachestvennykh dannykh v sotsial'nykh naukakh: uchebnoye posobiye* [Qualitative Data Analysis Practices in the Social Sciences: A Textbook]. Moscow: Izd. dom VShE Publ., 2023. (in Russ.).

**Z**hang Q., Segall R. S. Review of Data, Text and Web Mining Software. *Kybernetes*, 2010, vol. 39, no. 4, pp. 625–655. DOI: 10.1108/03684921011036835 (in English).

*Для цитирования:* Архипова М. Н., Калхиташвили Д. Ш., Чагадаева О. А. Опыт применения платформы *PolyAnalyst* для анализа базы данных антропологических интервью // Уральский исторический вестник. 2024. № 3 (84). С. 142–152. DOI: 10.30759/1728-9718-2024-3 (84)-142-152.

*For citation:* Arkhipova M. N., Kalkhitashvili D. Sh., Chagadaeva O. A. Using the PolyAnalyst Platform to Analyze a Database of Anthropological Interviews // Ural Historical Journal, 2024, no. 3 (84), pp. 142–152. DOI: 10.30759/1728-9718-2024-3 (84)-142-152.