

Научная статья

УДК 94+929+930.25+51-77

doi: 10.17223/24099554/23/17

Распознавание рукописного текста и интеллектуальный анализ: возможности нейронных технологий (на примере работы с «Дневником» Ф.П. Литке)

Екатерина Михайловна Болтунова¹

Антон Константинович Лаптев²

^{1, 2} *Национальный исследовательский университет «Высшая школа экономики», Москва, Россия*

¹ *ekboltunova@hse.ru*

² *aklaptev@hse.ru*

Аннотация. Авторы описывают проведенный практический эксперимент по интеллектуальному анализу дневника адмирала Ф.П. Литке, расшифрованного с помощью специально обученной нейросети. Эксперимент показывает практическую реализацию принципов автоматизированной работы исследователя с большими объемами сложного рукописного текста с минимальными затратами времени и человеческого ресурса.

Ключевые слова: нейронные сети, машинный анализ данных, Ф.П. Литке, дневник, Николай I, императорская семья, большие языковые модели

Источник финансирования: исследование проведено в Национальном исследовательском университете «Высшая школа экономики» в рамках проекта Российского научного фонда № 22-68-00066 «Культурное наследие России: интеллектуальный анализ и тематическое моделирование корпуса рукописных текстов».

Для цитирования: Болтунова Е.М., Лаптев А.К. Распознавание рукописного текста и интеллектуальный анализ: возможности нейронных технологий (на примере работы с «Дневником» Ф.П. Литке) // Имагология и компаративистика. 2025. № 23. С. 358–379. doi: 10.17223/24099554/23/17

Original article

doi: 10.17223/24099554/23/17

Handwriting recognition and data mining: Possibilities of neural network technologies (based on Admiral Fyodor Lütke's *Diary*)

*Ekaterina M. Boltunova*¹

*Anton K. Laptev*²

^{1, 2} *HSE University, Moscow, Russian Federation*

¹ *ekboltunova@hse.ru*

² *aklaptev@hse.ru*

Abstract. The main attention in the article is given to the description of a practical experiment conducted in order to search for methods of automated intellectual analysis of a historical manuscript document. On the example of Admiral Fyodor Lütke's diary from the collections of the State Archive of the Russian Federation, an integrated approach to the problem to be solved was demonstrated. The approach consisted in the use of two neural networks – a specially designed and trained neural network for deciphering the manuscript (obtaining machine-readable symbols) and a large language model of GPT type (YandexGPT) for labelling (collecting) the main metadata in the source. The metadata obtained (identified by the model) make it possible to evaluate the main themes of the diary and to obtain lists of keywords (for each text segment), personalities, geographical locations, and place names. With the help of ready-made tools (using the example of the web-application Voyant-Tools), which allow statistical analysis of the entire array of metadata, the researchers were able to carry out a near-automatic process of text mining and obtain information about the composition of the diary, the characteristics of personalities mentioned in the text (including members of the imperial family, representatives of court society, officials of the Military Department, etc.), as well as the distribution of geographical areas and toponyms in the text. The process was carried out without directly reading the historical source itself. The results obtained after the manuscripts were analysed by the YandexGPT model (taking into account the subsequent processing in Voyant-Tools) showed that the system correctly identified the main themes embedded in the text. The list of main themes correlated with the block of selected keywords (acting as a control group in relation to the list of themes), allowing to get a more objective picture. When compiling the register of personalities mentioned in the text and identifying a particular person, the system correctly recorded the patronymic,

rank or title in addition to the name. In some cases the system offered possibilities of interpretations concerning this or that name that appeared in the text. The obtained characteristic of the majority of personalities identified in the deciphered text was also quite accurate. It is significant that in addition to describing the status and position of a particular person, the system was able to determine with a high percentage of accuracy the content of the latter's relationship with the author (Lütke) and other persons. Thanks to the identified toponyms and geographical objects, the authors of the article managed to reconstruct the route of Lütke's European voyage in 1839, to record his assessments of this part of the world, which were reflected – directly or in a less explicit form – in the diary of the corresponding period. The authors also obtained a complete list of topographical and geographical objects mentioned by Lütke in his diary. As a result of this work, metadata were identified, which with high accuracy form an idea of the studied text and its author. The authors of the article managed to establish that the joint work of two neural networks can be carried out “seamlessly” (without human participation), based on the application of mathematical algorithms and technical means. The resulting metadata can be used as a navigation system, allowing the researcher to learn about the content of an originally unencrypted manuscript.

Keywords: neural networks, machine data analysis, Fyodor Lütke, diary, Nicholas I, imperial family, large language models

Financial Support: The research was conducted at the National Research University Higher School of Economics and supported by the Russian Science Foundation (RSF), Grant No.22-68-00066: Cultural Heritage of Russia: Intellectual Analysis and Thematic Modeling of the Corpus of Handwritten Texts.

For citation: Boltunova, E.M. & Laptev, A.K. (2025) Handwriting recognition and data mining: Possibilities of neural network technologies (based on Admiral Fyodor Lütke's *Diary*). *Imagologiya i komparativistika – Imagology and Comparative Studies*. 23. pp. 358–379. (In Russian). doi: 10.17223/24099554/23/17

«Стоит ли труда писать журнал? Если судить по сожалению, которое ощущаешь *после*, что не заметил того, что случилось *прежде*, то, конечно, стоит. Жизнь похожа на декорацию театральную; стоишь вблизи – видишь маранье; пятна разных цветов без связи и смысла; иногда что-то похожее на человеческие фигуры в уродливом виде; что-то подобное столбам, или деревьям; но все неопределено, не

окончено; – отодвинься на 50 шагов – и все сливается в гармоническое целое; кляксы и мазанье образуют оттенки и переливы цветов, света и тени; кривые, изломанные линии строятся в обманчивую перспективу. Так и жизнь: всякий день настоящий кажется обыкновенным, скучным, незначущим, люди уродами, происшествия кляксами; а отодвинется на год, и увидишь что все вместе составляет гармонический отрывок жизни, в котором все на своем месте, – но все контуры слабы, краски бледны, – отчего? оттого, что все это эскисировано только на полотне памяти, улетающими красками» [1. Л. 2].

Этот пассаж из «Дневника» известного мореплавателя Федора Петровича Литке, рассуждавшего в 1837 г. о призрачности памяти, значении временной дистанции и – что особенно примечательно – потребности своего рода объемного зрения для понимания истории, как нельзя лучше подходит в качестве отправной точки для разговора о возможностях, которые современная гуманитаристика открывает для себя с развитием технологий распознавания текста и интерпретации последнего посредством интеллектуального анализа текста (*Text Mining*) [2].

Граф Федор Петрович Литке (1797–1882) – адмирал, географ-гидрограф, исследователь Арктики, командовавший Кронштадтским портом в самый сложный период Крымской войны, когда у берегов русской Балтики действовала англо-французская эскадра, занимавший в конце жизни пост президента Императорской Академии наук, оставил огромное рукописное наследие. Помимо собственно описаний путешествий [3–11], этот комплекс документов включает в себя переписку [12–13], исследовательские тексты, очерки по истории русского флота [14], а также обширную коллекцию дневников [1, 15].

Эго-документы авторства Ф.П. Литке, прежде всего материалы его дневников, представляют особый интерес. Адмирал фиксировал события, происходившие с ним и вокруг него, большую часть жизни – с конца 1810-х до начала 1860-х гг., т.е. в период правления Александра I, Николая I и Александра II. На страницах «журналов» Федор Петрович писал о морских путешествиях, событиях, случавшихся в его профессиональном, дружеском или семейном кругу. Важно отметить, что в 1832–1848 гг. Ф.П. Литке руководил воспитанием второго сына императора Николая I, великого князя Константина Николае-

вича, будущего генерал-адмирала и управляющего морским ведомством. Описания отношений с юным великим князем, императором Николаем I, императрицей Александрой Федоровной составляют большую часть информации в дневнике 1830–1840-х гг. Здесь также немало описаний происходящего при императорском дворе – за один только 1837 г. адмирал подробно описывает такие значимые для империи события, как смерть А.С. Пушкина, открытие первой железнодорожной ветки и пожар Зимнего дворца; упоминаются имена И.К. Айвазовского, Ф.П. Врангеля, В.А. Жуковского, И.Ф. Крузенштерна, И.А. Крылова и многих других [1]. Не менее интересны и более поздние дневники. Так, в начале 1850-х гг. адмирал описывает события Крымской войны [15], а в «журнале» за 1861 г. – известное заседание Государственного совета 28 января по освобождению крестьян [16. С. 188]. Отметим, что «журналы» Федора Петровича отличаются высокая эмоциональность в передаче информации. Так, описания отношений с воспитанником носят характер постоянного эмоционального перехода от отчаяния и ощущения невозможности изменить характер юного великого князя к радости и гордости за успехи Константина Николаевича. Примечательны и страницы, посвященные началу Крымской войны – свое назначение в Кронштадт Литке описывает с нескрываемым раздражением и вне каких бы то ни было отсылок в парадигме долга перед империей [3; 15. Л. 2 об.].

При всем значении дневникового комплекса Ф.П. Литке, в котором содержится уникальная информация и неожиданные трактовки, и широком современном интересе к эго-документам, эти материалы практически не опубликованы (за исключением [17–20]) и существуют только в виде рукописного текста. Проекты издания дневника, анонсировавшиеся в академическом пространстве в разное время, реализованы не были [21. С. 164; 22]. Причин тому несколько. Прежде всего дневники Литке находятся сейчас в разных архивохранилищах – Государственном архиве Российской Федерации (ГА РФ), Санкт-Петербургском филиале архива Российской академии наук и Российском государственном архиве Военно-морского флота. При этом сам дневниковый комплекс исключительно объемный – он насчитывает более 2 100 архивных листов (более 4 200 страниц) [21. С. 164].

Очевидно, впрочем, что существуют и другие обстоятельства, объясняющие, почему столь ценное историческое свидетельство оказалось на периферии внимания публикаторов и, как следствие, значительного числа исследователей. Речь идет о специфике источника, а именно об особенностях текстуальной составляющей и неверной оценке содержания материала. Отметим, что прочтение дневников Ф.П. Литке серьезно затруднено из-за плохо читаемого почерка, сокращений и помет, а также многочисленных вставок на иностранных языках. Кроме того, можно предположить, что низкая востребованность комплекса дневников вызвана некорректным представлением о содержании текстов, которые принято соотносить исключительно с военно-морской тематикой. Анализ историографии показывает, что дневники оказались востребованными среди биографов самого Ф.П. Литке, историков российского флота и исследователей истории Латинской Америки [23–28]. При этом специалисты по истории имперской России, прежде всего, николаевского периода, на который приходится основной комплекс дневников, этот источник практически не используют.

В современной гуманитаристике такого рода проблемы достаточно часты – объемный сложно читаемый текст, разбор которого требует колоссальных затрат времени без гарантии получения нужной информации, часто откладывается в сторону. Вместе с тем развитие современных технологий (включая нейросети) и практик использования интеллектуального анализа способны радикально изменить ситуацию в этом отношении. Целью нашей статьи является подтверждение этой установки в рамках представления результатов работы по расшифровке и интерпретации части рукописного наследия Ф.П. Литке. Отметим сразу, что проект был проведен в два этапа, первым из которых стала расшифровка текста с использованием нейросетевой модели, а вторым – работа с большой языковой моделью для «базового» (или первоначального) интеллектуального анализа текста. Таким образом, в процессе исследования историческая рукопись первой половины XIX в. была обработана при помощи сразу двух нейронных сетей [29–31].

От архивной рукописи к постановке задач для интеллектуального анализа

Проект был реализован на материалах дневников Ф.П. Литке, датированных концом 1830-х – началом 1840-х гг. и хранящихся в ГА РФ. Выбор указанной части комплекса эго-документов определялся компактностью хранения этого сегмента информации и ее значимостью для специалистов по истории Российской империи – данные за этот отрезок времени наиболее подробны и относятся к периоду, когда Ф.П. Литке курировал воспитание великого князя Константина Николаевича, проводя большую часть времени при императорском дворе.

Этот массив рукописных текстов значителен по объему и скомпонован в 4 архивных дела (4 тома), 944 архивных листа (1 888 страниц) [1]. Каждый из томов дневника представляет собой объемную тетрадь в коричневом кожаном переплете с ленточкой-закладкой; на обложке золотым теснением представлена дата (например, «1837. 1838» на Т. 1). Сам текст написан на русском языке со вставками на французском и немецком (в большей степени) и английском языках. В тексте содержатся также вклейки с печатными текстами (схемы, приглашения и пр.) и позднейшие пометы и подчеркивания, которые, вероятнее всего, оставил один из первых читателей дневника. Отдельно отметим, что в дневнике Ф.П. Литке мы обнаружим устаревший синтаксис и словоформы, характерные для первой половины XIX столетия.

На первом этапе для расшифровки рукописного текста и перевода его в машиночитаемый формат была разработана и применена специализированная нейронная сеть¹. Особенностью этой части проекта стало использование комплекса математических алгоритмов, позволяющих распознавать историческую рукопись с высокой степенью точности на относительно малом объеме данных для предварительного обучения [32]. Используемая нейронная сеть была обучена на 80 листах архивного документа, распознанных «вручную», что с учетом расположения текста (примерно 25 строк на каждой странице) сформировало дата-сет из 2 000 строк. Для сравнения скажем, что из-

¹ Разработчики – Л.М. Местецкий, Н.А. Ломов, Д.А. Кропотов.

вестный проект «DigitalПетр», в задачи которого входила расшифровка рукописей Петра Великого, был реализован с обучением нейросети на дата-сете из 10 000 распознанных «вручную» строк [33].

Распознанный таким образом текст дневника Ф.П. Литке имел крайне низкий процент ошибок CER и WER¹, а в самом рукописном тексте верно было распознано более 96% символов (букв). Отметим, однако, что использованный алгоритм был разработан для рукописного текста на русском языке и не предполагал разбор вставок на иностранных языках, что в целом корректирует процент нераспознанной информации в сторону его уменьшения.

Расшифровка текста, полученная при помощи специализированной нейронной сети, позволила перейти к следующему этапу работы – проведению интеллектуального анализа текста, т.е. автоматическому извлечению сведений из корпуса данных.

Принято думать, что использование интеллектуального анализа применительно к историческому источнику имеет существенные ограничения, поскольку исходная информация, заключенная в таком тексте, часто не является формализованной [34], а значит, ее нельзя описать простой математической функцией. Это, в свою очередь, требует привлечения ресурсов для организации и структурирования данных, которые впоследствии будут подвергнуты анализу². Вместе с тем существенно, что сведения, необходимые для разработки математических алгоритмов, – это параметры, которые на первом этапе работы собирает исследователь. Речь идет о таких базовых позициях, как тип документа, дата, место создания, авторство и т.д., а также повторяющихся элементах (ключевые слова, включая упомянутые в тек-

¹ CER – частота ошибок в символах, WER – частота ошибок в слов. Показатели фиксируются за счет сравнения текстовых блоков, распознанных математическим алгоритмом с контрольной группой – этим же текстовым блоком, но распознанным человеком «вручную».

² В рамках работы с обучением нейронных сетей существует ряд проблем, которые определяются сложностью подготовки данных, корректных и подходящих для обучения нейронных моделей. Так, из общего времени, затраченного на решение конкретной задачи с помощью нейронной сети, до 90% может уйти на «вручную» подготовку («чистку») данных для их последующего использования в рамках реализации корректного процесса обучения.

сте персоналии и институции). Именно эти данные (их следует рассматривать как метаданные) могут стать основой структурирования материала, открывая возможности для интеллектуального анализа такого зачастую непростого текста, как исторический источник.

Существенно, однако, что исследователь с самого начала рассматривает такую информацию в рамках определенной системы координат или сложившейся иерархии, отмечая для себя главное и второстепенное, выстраивая хронологическую последовательность или фиксируя уровни социального пространства. Специализированный алгоритм, напротив, не видит разницы между типами данных без дополнительного обучения и донастройки: первоначально все текстовые данные воспринимаются им как равновесные и равнозначные. Это ставит вопрос о возможности обучения системы с целью идентификации машиной заложенного в тексте смысла и выявления расставленных в рукописи акцентов, например, в части семантики или геометрического расположения данных на странице рукописного текста.

В современном академическом пространстве возможность выявить метаданные, необходимые для идентификации в историческом источнике смысловых структур и их иерархий, может быть реализована за счет привлечения больших языковых моделей (*Large Language Model (LLM)*). Последние получили распространение после 2018 г., когда компанией OpenAI была представлена модель GPT (*Generative Pre-Trained Transformer*¹), способная к выполнению широкого спектра задач за счет обучения на больших объемах данных и использования десятков миллиардов (в последующих генерациях – сотен миллиардов) внутренних параметров². Такие нейронные языковые модели могут отражать или «схватывать» большую часть синтаксиса и семантики естественного языка.

¹ Генеративный предобученный трансформер (англ.).

² Параметры – это характеристики модели, которые определяют ее способность к обучению и выполнению задач. В контексте больших языковых моделей параметры представляют собой веса (весовые коэффициенты) связей между нейронами (словами) в нейронной сети (языковой модели). Эти веса настраиваются в процессе обучения модели на большом наборе данных [35].

Для реализации наших задач по интеллектуальному анализу дневника Ф.П. Литке наиболее подходящим из инструментов была признана YandexGPT – российская генеративная языковая модель, разработанная компанией Яндекс и представленная в публичном пространстве в 2023 г.¹ Ключевым критерием выбора стал фактор языка: основным источником для обучения YandexGPT было использование текстов на русском языке. Это сформировало ожидания, что в результате обработки текста будет получен меньший процент искажений (по сравнению с ChatGPT, где обязательным элементом является машинный перевод (английский-русский)), а сгенерированные метаданные будут более релевантны в отношении исследуемого исторического источника.

В рамках проведенного исследования для языковой модели YandexGPT был сформулирован пул запросов (инструкций). Они были составлены как ряд последовательных вопросов, ответы на которые, как предполагалось, могли дать набор сведений (метаданных), позволяющих структурировать содержание рукописи, предварительно расшифрованной с помощью нейронной сети. В нашем случае существенно, что распознанный текст дневника Ф.П. Литке не подвергался последующей «ручной» обработке и сверке, т.е. при переходе к интеллектуальному анализу все еще сохранял определенный процент ошибок (около 4% в основном тексте на русском языке) и сегменты нераспознанного текста на иностранных языках.

На этом этапе проекта было принято решение использовать для обработки YandexGPT один из томов дневника (Т. 2), наиболее значительный по объему (423 архивных листа, 846 страниц) [36]. В рамках исследования была поставлена задача сформировать пул данных с возможностью последующего объединения в три блока (словаря) – «Персоналии», «Места и топонимы», «Эмоциональный спектр».

В практическом отношении при работе с YandexGPT формат запроса к системе выглядел следующим образом:

«Шаг за шагом проанализируй представленный текст. Предоставь информацию о:

¹ YandexGPT основана на архитектуре языковой модели YaLM (Yet another Language Model), которая была обучена на широком наборе данных.

- 1) темах, характерных для текста (укажи их в порядке от наиболее к наименее точной);
- 2) ключевых словах, характерных для текста (укажи ключевые слова (не меньше 8 и не больше 14) в порядке от наиболее к наименее часто упоминаемому);
- 3) персоналиях, упомянутых в представленном тексте (максимально точно);
- 4) местах и топонимах, упомянутых в представленном тексте (максимально точно);
- 5) основных эмоциях, выражение которых можно выявить в тексте (укажи их в порядке от наиболее к наименее релевантной);
- 6) взаимосвязях между самыми упоминаемыми словами в тексте».

Отметим, что интерфейс YandexGPT базируется на загрузке текстового фрагмента определенного размера, поэтому весь массив расшифрованного текста второго тома дневника Ф.П. Литке был разделен на условные блоки размером около 4 500 токенов.¹ В нашем случае это соответствовало 2,5 архивным листам (5 страницам) рукописного текста. Такой объем был выбран исходя из конкретных практических условий: запрос и ответ на него в сумме не должен был превышать 8 000 токенов, что установлено правилами YandexGPT.

Для последующего статистического анализа результатов, полученных после обработки расшифрованного текста, выявления коллокаций² и визуализации полученной информации, был также задействован ресурс «Voyant-tools».

¹ Токены – это последовательности символов, на которые YandexGPT API разбивает текст. В контексте нейросетей термин «tokens» (токены) обычно относится к минимальным единицам, на которые разбивается входной текст или последовательность символов перед подачей на обработку модели. Токеном может быть одна буква, одно слово или даже целая фраза, в зависимости от типа и задачи модели. Слово может представлять как один токен, так и занимать размер в несколько токенов в зависимости от используемых алгоритмов анализа объема текста.

² Здесь под коллокациями мы понимаем совместно встречаемые слова, которые наиболее часто фиксируются рядом с ключевыми словами. Такая терминология характерна для инструментов сетевого анализа текста. См. Voyant Tools: веб-пространство для чтения и анализа цифровых текстов [37].

Результаты работы большой языковой модели

Результаты, полученные после анализа рукописей Ф.П. Литке моделью YandexGPT (с учетом последующей обработки в Voyant-Tools), позволяют зафиксировать несколько выводов. Прежде всего, система вполне корректно идентифицировала основные темы, заложенные в тексте. При этом список последних хорошо соотносится с блоком идентифицированных ключевых слов, которые в нашем случае выступали как контрольная группа по отношению к списку тем, позволяя получить более объективную картину.

Частотный анализ ключевых слов из дневника Ф.П. Литке показывает, что самыми востребованными словами являются здоровье (27), семья (19), путешествие (18), К.Н. (Константин Николаевич) (16), болезнь (15), погода (14), корабли (14), уроки (13), Имп. (Императрица Мария Федоровна) (12), дети (12), Г-рь (Государь) (12), обед (11), князь (11), праздник (10), Эмс¹ (9), дворец (9), бал (9), церковь (8).

Система достаточно точно определила и связи между позициями в списке выше. Так, соотнесенными оказались слова «К.Н.» и «Г-рь», т.е. указания на юного великого князя Константина Николаевича, воспитанника Ф.П. Литке, и императора Николая I. Действительно, в дневнике Литке описание воспитания одного из сыновей императора передано как процесс непрерывной коммуникации между Федором Петровичем, воспитанником и монархом-отцом.

Обращает на себя внимание и указание системы на связь между словом «путешествие» и состоянием здоровья автора. Это соответствует содержанию дневника – европейская поездка Ф.П. Литке в 1839 г. должна была послужить восстановлению здоровья адмирала, страдавшего от «судорог в желудке», постоянной тошноты, истощения и испытывавшего сильнейшее беспокойство относительно своего физического состояния [36. Л. 75 об.–76]. Приезд в Бад-Эмс (Эмс), известный бальнеологический курорт, сопровождается в «журнале» следующими рассуждениями: «Июня 8 (20). Я уже в Эмсе. Это цель, к которой странствование мое до сих пор служило только введением. В Эмсе, кот. должен отвечать мне на вопрос: to be, or not to be? Каков то будет этот ответ, когда придет время сказать: Прощай Эмс? Будем ожидать с надеждою на Бога...» [36. Л. 75].

¹ Эмс (Бад-Эмс), город-курорт в Германии.

Система также отметила связь между категорияй «театр» и состоянием тревожности, которое время от времени описывает автор. При этом реакции Федора Петровича, отмеченные в дневнике применительно к театральным постановкам и вечерам классической музыки, интерпретированы YandexGPT как различные. Например, в качестве ответа на запрос было получено утверждение, что посещение театра зачастую не удовлетворяет Федора Петровича, однако он «часто упоминает музыку, музыкальные произведения и события, связанные с музыкой, что говорит о том, что музыка играет важную роль в его жизни». Отметим, что все четыре тома дневников наполнены выражениями неудовольствия Литке относительно необходимости сопровождать Константина Николаевича в театр и осуждением подобного времяпрепровождения, которое адмирал полагал забавой, которая только отвлекает великого князя от учебы и нарушает режим дня [36. Л. 13 об., 14, 178, 297, 300 об., 301; 38. Л. 63, 63 об., 120]. С другой стороны, в своем дневнике Литке предстает ценителем серьезной музыки, организатором частных музыкальных событий и поклонником Ф. Листа [38. Л. 84 об. – 85 об.].

Данные, полученные после обработки, позволяют составить исключительно точный реестр персоналий, упомянутых в тексте. При этом, идентифицируя человека, система корректно фиксировала помимо имени его отчество, звание или титул (например, граф Григорий Александрович Строганов, князь Александр Сергеевич Меншиков, генерал от кавалерии Николай Андреевич Рсад, Иван Федорович Крузенштерн). В отдельных случаях система предлагала возможности трактовок относительно того или иного имени, которое появлялось в тексте.

Анализ частотности демонстрирует, что самыми упоминаемыми в тексте оказались слова (сокращения слов) «К.Н.», «Г-рь», «Наследник», «В.К.М.Н.» и «Имп.»¹. Эти данные коррелируют с содержанием

¹ Ряд упоминаний были приведены «к единому знаменателю» на этапе обработки полученных от языковой модели данных. Например, упоминания Великой княгини Марии Николаевны по-разному отражались автором. Он мог использовать только инициалы «М.Н.» либо указывать ее статус «В.К.М.Н.», или просто прибегать к сокращениям «Вел. Кн. Мар. Ник.». Для удобства статистического анализа вся эта вариативность была сведена к единой форме представления конкретной личности – «В.К.М.Н.». Данное действие не представляет серьезной сложности для ее автоматизации. Вероятно, в будущем исследователь сможет

второго тома дневника Ф.П. Литке – центральное место здесь занимает наследник Константин Николаевич. Последний упоминается в дневнике существенно чаще других (47 упоминаний). Для сравнения: сокращение «Г-рь», относящееся к императору Николаю I, встречается в тексте всего 19 раз. Однако не все результаты обработки, которые предлагает YandexGPT, оказываются предсказуемыми. Так, великая княгиня Мария Николаевна («В.К.М.Н.») упоминается в тексте чаще своей матери, императрицы Александры Федоровны («Имп.») (13 против 10 упоминаний). При этом график коллокаций (сетевой график слов с более высокой частотой совместного употребления) показывает, что к ключевому слову «В.К.М.Н.» наиболее близко стоят «императрица», «государь» и «К.Н.», а коллокацией для «В.К.М.Н.» выступает «дворец». Это свидетельствует о роли, которую играла в эти годы при дворе и в императорской семье старшая дочь императора Николая I, принявшая решение после замужества остаться в России.

Достаточно точной стоит признать и характеристику большинства персоналий. В частности, система характеризует воспитанника Литке великого князя Константина Николаевича следующим образом: «К.Н. – ребенок, о котором идет речь в тексте. Он прилежен в учебе, но испытывает трудности в общении с товарищами. Автор текста беспокоится о его будущем и пытается найти баланс между учебой и общением со сверстниками». За вычетом указания на успехи в учебе – сведения о достижениях великого князя в учебе в дневнике разноплановы – данные переданы вполне корректно, а представленная детализированная характеристика дает не только описание характера конкретного человека, но и раскрывает суть взаимоотношений последнего с автором и другими лицами.

Отметим и некоторые ошибочные трактовки, предложенные системой. В целом обработка большого массива сведений, а также извлечение существенной номенклатуры метаданных до определенной степени нивелируют отдельные ошибки при работе автоматического алгоритма. Вместе с тем важно помнить, что неверная интерпретация даже небольшого сегмента может стать проблемой при последующем

объединять схожие имена и указания на людей в единые блоки метаданных на основе предложений со стороны нейронных сетей или более простых математических алгоритмов.

анализе текста на основе полученных метаданных. Так, языковая модель YandexGPT в ответе на один из запросов верно зафиксировала, что автор дневника – «наставник К.Н., который описывает свои наблюдения за жизнью подопечного и событиями при дворе; К.Н. – воспитанник автора, о котором он рассказывает; Имп. – человек, от которого зависит судьба К.Н. и его обучение». Однако алгоритм охарактеризовал императора Николая I («Г-рь») как коллегу автора, с которым последний обсуждает вопросы воспитания и обучения. Более того, в рамках работы алгоритма, словоформа «Г-рь» стала самой неверно интерпретируемой позицией. Очевидно, однако, что речь идет не столько о проблеме технического характера, сколько о сложностях с интерпретацией этого конкретного сокращения – другие краткие словоформы алгоритм распознавал достаточно точно, в том числе разделяя в смысловом отношении сокращения «Имп.» (Императрица) от «Г-рь» (Государь/Император).

Обработка текста YandexGPT позволила также составить перечень топографических и географических объектов, которые Ф.П. Литке упоминает в дневнике. Среди упомянутых мест и топонимов наибольшей частотой отличаются «дворец» (35), «Петербург» (26), «Царское Село» (18) и «церковь» (13). В последнем случае перед нами двойная проекция – отсылка к религиозной жизни самого адмирала и указание на значимый сегмент придворной жизни.

Система четко локализовала и территорию путешествия адмирала, о котором речь идет во втором томе дневников. Автор сообщает здесь о Варшаве, Берлине, Гамбурге, Дармштадте, Лондоне, Париже и пр. При этом самыми упоминаемыми странами (по плотности географических привязок) в порядке убывания выступают: Швейцария, Германия, Англия, Польша, а также Бельгия и Нидерланды, а ключевыми топонимами в анализируемом тексте оказываются Петербург и Париж. Франция, за редким исключением, сводится в дневнике к Парижу.

Сама по себе такая выкладка позволяет реконструировать маршрут движения Федора Петровича по Европе и определить его предпочтения и симпатии при взгляде на страны этой части мира. Отдельно отметим, что возможности Voyant-Tools помогают визуализировать путешествие адмирала в виде динамической онлайн карты, на которой

связи между конкретными локациями отображаются в виде линий между точками на карте.

Качественный анализ слов, маркирующих эмоции, демонстрирует, что для второго тома дневников Ф.П. Литке значимой частью спектра выступают «радость» (75 упоминаний), «удивление»/«сюрприз» (54 упоминания), «грусть» (54 упоминания) и «беспокойство» (38 упоминаний). В дневнике достаточно часто фигурирует указание на «гнев» (22 упоминания).

Интересным результатом при рассмотрении сетевых связей внутри эмоционального спектра оказывается соотношение таких ключевых слов, как «грусть» и «здоровье». При этом на графике коллокаций видна разнонаправленность сегментов текста с рассказами о здоровье и указанием на испытываемую радость: чем чаще в тексте упоминается здоровье, тем более нисходящий тренд у графика, описывающего частоту упоминания словоформы «радость», в заключительной части второго тома упоминания здоровья приобретают минимальные значения, а указания на радость достигают близких к пиковым значений. По сути, тема здоровья является основой для развития динамики эмоционального спектра в этом томе дневника. Этот вывод, полученный с помощью инструментов анализа текста, также вполне коррелирует с содержанием дневника.

Значимым для анализа содержания текста является и график коллокаций словоформ, соотнесенных с эмоциональной сферой и возникающих при описании взаимоотношений Ф.П. Литке с воспитанником. Наиболее часто здесь встречаются указания на беспокойство (12 упоминаний), грусть (11 упоминаний), гнев (6 упоминаний) и радость (6 упоминаний). По сути, алгоритмы предлагают нам картину, в которой Ф.П. Литке выступает как наставник, испытывающий по отношению к воспитаннику беспокойство, а сами по себе отношения с юным великим князем окрашены в тона грусти. Действительно, в конце второго тома Литке все чаще задумывается над эффективностью принятой системы обучения: «Все ведет к тому, чтобы развлечение, веселие, поставить на первую степень важности, а работы, порядок словом все, о чем мы (учителя. – *Авт.*) хлопочем представить вещью постороннюю, которую позволено заняться так, *faute de*

miex¹» [36. Л. 309 об – 310]. В начале следующего, третьего, тома Федор Петрович, составив своего рода реестр знаний великого князя, и вовсе задумывается над сложившемся положением дел и перспективой будущего: «Что ж из этого следует? Что мы со стороны умственного развития В.К. можем быть совершенно успокоены и щастливы? Отнюдь нет, с этою необыкновенною остротою и понятливостию ума, позволяющею ему схватывать, так сказать, на лету самые сложные и трудные вещи, и знать их, если памяти для того достаточно, соединена какая-то непостижимая инерция ума... В нем вкоренилось какое-то неодолимое отвращение от всякой работы, от всякого труда. Покуда дело идет гладко, легко, так сказать само собою, до тех пор и все хорошо; но едва встретится малейшее затруднение, для преодоления которого нужно соображение или размышление – кончено, пропала и охота, <...> начались капризы» [38. Л. 8 об. – 9].

Выводы

В целом в результате проведенного эксперимента, реализованного в два этапа (расшифровка рукописного текста и его анализ при помощи большой языковой модели YandexGPT) на материалах рукописных текстов первой половины XIX в., нам удалось выявить метаданные, которые с высокой точностью формируют представление об изучаемом тексте и его авторе. Эксперимент демонстрирует, что совместная работа двух нейронных сетей может быть осуществлена «бесшовно» (фактически без участия человека), на основе применения математических алгоритмов и технических средств. Полученные метаданные могут быть использованы как система навигации, позволяющая исследователю узнать о содержании изначально нерасшифрованной рукописи.

Существенно, что сопоставление метаданных, полученных в рамках работы большой языковой модели, с уже известными по другим источникам сведениями и/или трактовками высвечивает спектр несоответствий, искажений и отклонений. Это, в свою очередь, может вос-

¹ За неимением лучшего (*франц.*). В цитате сохранена пунктуация автора.

приниматься исследователем как точка входа и основа для формирования нового исследовательского вопроса в рамках работы с указанным источником.

Очевидно, что подобные принципы работы могут быть использованы при разборе сложных в прочтении рукописей, которые хранятся в настоящий момент в российских архивах и не введены в научный оборот.

Список источников

1. Государственный архив Российской Федерации (далее ГАРФ). Ф. 1463. Оп. 1. Д. 1111–1114.
2. *Guldi Jo*. The Dangerous Art of Text Mining: A Methodology for Digital History. Cambridge : Cambridge University Press, 2023. 496 p.
3. *Литке Ф.П.* Четырехкратное путешествие в Северный Ледовитый океан на военном бриге «Новая Земля» в 1821–1824 годах. М. : Географиз, 1948. 337 с.
4. *Литке Ф.П.* Путешествие вокруг света на военном шлюпе «Сенявин» в 1826–1829. М. : Географиз, 1948. 303 с.
5. Российский государственный исторический архив (далее РГИА). Ф. 472. Оп. 13. Д. 885.
6. РГИА. Ф. 560. Оп. 15. Д. 31.
7. РГИА. Ф. 733. Оп. 89. Д. 133.
8. РГИА. Ф. 789. Оп. 1. Ч. 2. Д. 1825.
9. Российский государственный архив Военно-морского флота (далее РГА ВМФ). Ф. 296. Оп. 1. Д. 151.
10. РГА ВМФ. Ф. 384. Оп. 1. Д. 81.
11. РГА ВМФ. Ф. 402. Оп. 1. Д. 337, 458.
12. РГИА. Ф. 537. Оп. 1. Д. 195, 202, 216, 225–226, 228–229, 233.
13. РГИА. Ф. 759. Оп. 7. Д. 777, 1009.
14. РГА ВМФ. Ф. 315. Оп. 1. Д. 99, 225, 530.
15. Санкт-Петербургский филиал архива Российской академии наук. Ф. 34. Оп. 1. Д. 8.
16. *Померанская Т.В., Сидорова А.Н., Сидорова М.В.* «Сохранить память о впечатлениях каждого дня...». Из дневника Ф.П. Литке за 1837 год // Российский архив. М. : Российский фонд культуры, 2016. Вып. 22. С. 186–308.
17. *Шур Л.А.* К берегам Нового Света. Из неопубликованных записок русских путешественников начала XIX века. М. : Наука, 1971. 300 с.
18. Елагинские чтения. СПб. : Остров, 2003. Вып. 1. 102 с.
19. Российский архив: история Отечества в свидетельствах и документах: XVIII–XX вв. М. : Российский фонд культуры, 2016. 728 с.
20. Российский архив: история Отечества в свидетельствах и документах: XVIII–XX вв. М. : Российский фонд культуры, 2017. 687 с.

21. Династия Романовых: 400 лет в истории России : материалы междунар. науч. конф. СПб. : ФГБОУВПО «СПГУТД», 2013. 562 с.
22. Пивоваров Е.Г., Скрыдлов А.Ю. К истории изучения бумаг Ф.П. Литке из собрания Российского государственного архива древних актов // Вестник архивиста. 2022. № 2. С. 335–337.
23. Безобразов В.П. Граф Федор Петрович Литке. СПб. : Тип. Имп. Акад. Наук, 1888. 239 с.
24. Алексеев А.И. Федор Петрович Литке. М. : Наука, 1970. 276 с.
25. Литке Н.Ф. Роль адмирала Ф.П. Литке в развитии русской географической науки : дис. ... д-ра геогр. наук. Л., 1951. 297 с.
26. Известия Всесоюзного Географического общества. Л. : Наука, 1964. Вып. 5. С. 414–419.
27. Копелев Д.Н. От мыса Головнина к Земле Александра I. Российские кругосветные экспедиции в первой половине XIX в. М. : РОССПЭН, 2021. 495 с.
28. Копелев Д.Н. На службе империи: немцы и российский флот в первой половине XIX в. СПб. : Изд-во Европейского ун-та, 2010. 338 с.
29. Бородин Л.И. Наука о данных и технологии искусственного интеллекта: возможности и ограничения в исследованиях историков // Труды Отделения историко-филологических наук. М. : РАН, 2022. Т. 11. 296 с.
30. Golubev A. Digitizing Archives in Russia: Epistemic Sovereignty and Its Challenges in the Digital Age // The Palgrave Handbook of Digital Russia Studies. Palgrave Macmillan, 2021. P. 353–369.
31. Болтунова Е.М. Региональная история России: исследовательское поле и архивная практика (1990-е – начало 2020-х гг.) // Новое литературное обозрение. 2022. № 6 (178). С. 241–243.
32. Laptev A., Lomov N., Stepochkin D., Kropotov D. Handwritten Text Recognition and Browsing in Archive of Prisoners' Letters from Smolensk Convict Prison // Analysis of Images, Social Networks and Texts: 11th International Conference, AIST 2023, Yerevan, Armenia, September 28–30, 2023, Revised Selected Papers. Yerevan : Springer, 2024. P. 227–240.
33. Базарова Т., Димитров Д., Потанин М., Проскуракова М. Распознать и транскрибировать. Автографы Петра Великого // Воронцово поле. 2020. № 4. С. 64–71.
34. Oiva M. Topic Modeling Russian History // The Palgrave Handbook of Digital Russia Studies. Palgrave Macmillan, 2021. P. 438–439.
35. OpenAI's GPT-3 Language Model: A Technical Overview. URL: <https://lambdalabs.com/blog/demystifying-gpt-3>
36. ГАРФ. Ф. 1463. Оп. 1. Д. 1112.
37. Voyant Tools. Веб-пространство для чтения и анализа цифровых текстов. URL: <https://voyant-tools.org/>
38. ГАРФ. Ф. 1463. Оп. 1. Д. 1113.

References

1. State Archive of the Russian Federation (GARF). Fund 1463. List 1. File 1111-1114.
2. Guldi, J. (2023) *The dangerous art of text mining: A methodology for digital history*. Cambridge: Cambridge University Press.
3. Lütke, F.P. (1948) *Chetirekhkratnoe puteshestvie v Severnyi Ledovityi okean na voennom brige "Novaya Zemlya" v 1821–1824 godakh* [Four-time journey to the Arctic Ocean on the military brig "Novaya Zemlya" in 1821–1824]. Moscow: Geografiz.
4. Lütke, F.P. (1948) *Puteshestvie vokrug sveta na voennom shlyupe "Senyavin" v 1826–1829* [Journey around the world on the military sloop "Senyavin" in 1826–1829]. Moscow: Geografiz.
5. Russian State Historical Archive (RGIA). Fund 472. List 13. File 885.
6. Russian State Historical Archive (RGIA). Fund 560. List 15. File 31.
7. Russian State Historical Archive (RGIA). Fund 733. List 89. File 133.
8. Russian State Historical Archive (RGIA). Fund 789. List 1. Part 2. File 1825.
9. Russian State Archive of the Navy (RGA VMF). Fund 296. List 1. File 151.
10. Russian State Archive of the Navy (RGA VMF). Fund 384. List 1. File 81.
11. Russian State Archive of the Navy (RGA VMF). Fund 402. List 1. Files 337, 458.
12. Russian State Historical Archive (RGIA). Fund 537. List 1. Files 195, 202, 216, 225–226, 228–229, 233.
13. Russian State Historical Archive (RGIA). Fund 759. List 7. Files 777, 1009.
14. Russian State Archive of the Navy (RGA VMF). Fund 315. List 1. Files 99, 225, 530.
15. St. Petersburg Branch of the Archive of the Russian Academy of Sciences (SPbF ARAN). Fund 34 List 1 File 8.
16. Pomeranskaia, T.V., Sidorova, A.N. & Sidorova, M.V. (2016) "Sokhranit' pamiat' o pechatleniiakh kazhdogo dnya...". Iz dnevnika F.P. Litke za 1837 god ["To preserve the memory of the impressions of each day...". From the diary of F.P. Lütke for the year 1837]. In: *Rossiiskii arkhiv* [Russian Archive]. Vol. MMXII. Moscow: Rossiiskiy fond kul'tury.
17. Shur, LA. (1971) *K beregam Novogo Sveta. Iz neopublikovannykh zapiskov russkikh puteshestvennikov nachala XIX veka* [To the shores of the New World. From the unpublished notes of Russian travelers of the early 19th century]. Moscow: Nauka.
18. Anon. (2003) *Elaginskie chteniya* [Elagin readings]. St. Petersburg: Ostrov.
19. Anon. (2016) *Rossiiskiy arkhiv: istoriya Otechestva v svidetel'stvakh i dokumentakh: XVIII–XX vv.* [Russian archive: History of the Fatherland in testimonies and documents: 18th–20th centuries]. Moscow: Rossiiskiy fond kul'tury.
20. Anon. (2017) *Rossiiskiy arkhiv: istoriya Otechestva v svidetel'stvakh i dokumentakh: XVIII–XX vv.* [Russian archive: History of the Fatherland in testimonies and documents: 18th–20th centuries]. Moscow: Rossiiskiy fond kul'tury.

21. Anon. (2013) *Dinastiya Romanovykh: 400 let v istorii Rossii: materialy mezhdunarodnoy nauchnoy konferentsii* [The Romanov dynasty: 400 years in the history of Russia: Proceedings of the International Conference]. St. Petersburg: FGBOU VPO "SPGUTD".
22. Pivovarov, E.G. & Skrydlov, A.Yu. (2022) K istorii izucheniya bumag F.P. Litke iz sobraniya Rossiiskogo gosudarstvennogo arkhiva drevnikh aktov [On the history of the study of the papers of F.P. Lütke from the collection of the Russian State Archive of Ancient Acts]. *Vestnik arkhivista*. 2. pp. 335–337.
23. Bezobrazov, V.P. (1888) *Graf Fedor Petrovich Litke* [Count Fyodor Petrovich Lütke]. St. Petersburg: Imperial Academy of Sciences.
24. Alekseev, A.I. (1970) *Fedor Petrovich Litke* [Fyodor Petrovich Lütke]. Moscow: Nauka.
25. Litke, N.F. (1951) *Rol' admirala F.P. Litke v razvitii russkoy geograficheskoy nauki* [The role of Admiral F.P. Lütke in the development of Russian geographical science]. Geography Dr. Diss. Leningrad.
26. *Izvestiya Vsesoyuznogo Geograficheskogo obshchestva*. (1964) 5. pp. 414–419.
27. Kopelev, D.N. (2021) *Ot mysa Golovnina k Zemle Aleksandra I: Rossiiskie krugosvetnye ekspeditsii v pervoy polovine XIX veka* [From Cape Golovnin to Alexander I's Land: Russian circumnavigation expeditions in the first half of the 19th century]. Moscow: ROSSPEN.
28. Kopelev, D.N. (2010) *Na sluzhbe imperii: nemtsy i rossiiskii flot v pervoy polovine XIX veka* [In the service of the empire: Germans and the Russian fleet in the first half of the 19th century]. St. Petersburg: European University.
29. Borodkin, L.I. (2022) Nauka o dannykh i tekhnologii iskusstvennogo intellekta: vozmozhnosti i ogranicheniya v issledovaniakh istorikov [Data science and artificial intelligence technologies: Opportunities and limitations in historians' research]. In: *Trudy Otdeleniya istoriko-filologicheskikh nauk* [Proceedings of the Department of Historical and Philological Sciences]. Vol. 11. Moscow: RAS.
30. Golubev, A. (2021) Digitizing archives in Russia: Epistemic sovereignty and its challenges in the digital age. In: *The Palgrave Handbook of Digital Russia Studies*. Palgrave Macmillan. pp. 353–369.
31. Bol'tunova, E.M. (2022) Regional'naya istoriya Rossii: issledovatel'skoe pole i arkhivnaya praktika (1990-e – nachalo 2020-kh gg.) [Regional history of Russia: Research field and archival practice (1990s – early 2020s)]. *Novoe literaturnoe obozrenie*. 6 (178). pp. 241–243.
32. Laptev, A., Lomov, N., Stepanchkin, D. & Kropotov, D. (2024) Handwritten text recognition and browsing in archive of prisoners' letters from Smolensk convict prison. *Analysis of Images, Social Networks and Texts: 11th International Conference, AIST 2023*. Yerevan, Armenia. 28–30 September 2023. Revised Selected Papers. Yerevan: Springer. pp. 227–240.
33. Bazarova, T., Dimitrov, D., Potanin, M. & Proskuryakova, M. (2020) Raspoznat' i transkribirovat'. Avtografy Petra Velikogo [Recognize and transcribe. Autographs of Peter the Great]. *Vorontsovo pole*. 4. pp. 64–71.

34. Oiva, M. (2021) Topic modeling Russian history. In: *The Palgrave Handbook of Digital Russia Studies*. Palgrave Macmillan. pp. 438–439.
35. Lambdalabs.com. (n.d.) *OpenAI's GPT-3 Language Model: A technical overview*. [Online] Available from: <https://lambdalabs.com/blog/demystifying-gpt-3>
36. State Archive of the Russian Federation (GARF). Fund 1463. List 1. File 1112.
37. Voyant Tools. (n.d.) [Online] Available from: <https://voyant-tools.org/>
38. State Archive of the Russian Federation (GARF). Fund 1463. List 1. File 1113.

Информация об авторах:

Болтунова Е.М. – канд. ист. наук, директор Института региональных исторических исследований факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва, Россия). E-mail: ekboltunova@hse.ru

Лаптев А.К. – канд. ист. наук, научный сотрудник Института региональных исторических исследований факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва, Россия). E-mail: aklaptev@hse.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

E.M. Boltunova, Cand. Sci. (History), director of the Institute of Regional Historical Studies, National Research University Higher School of Economics (HSE University) (Moscow, Russian Federation). E-mail: ekboltunova@hse.ru

A.K. Laptev, Cand. Sci. (History), research fellow, Institute of Regional Historical Studies, National Research University Higher School of Economics (HSE University) (Moscow, Russian Federation). E-mail: aklaptev@hse.ru

The authors declare no conflicts of interests.

Статья принята к публикации 10.12.2024.

The article was accepted for publication 10.12.2024.