Имагология и компаративистика. 2025. № 24. С. 186–208
Imagology and Comparative Studies. 2025. 24. pp. 186–208

Научная статья
УДК 821.161.1+519.6
doi: 10.17223/24099554/24/9

Машинная атрибуция почерка в решении источниковедческих проблем (на материале переписки Г.Н. Потанина)

Виталий Сергеевич Киселев[1]
Наталия Михайловна Пронина[2]

[1] *Томский государственный университет, Томск, Россия, kv-uliss@mail.ru*
[2] *Московский государственный университет имени М. В Ломоносова, Москва, Россия, natalka-pronina@mail.ru*

Аннотация. Статья посвящена возможностям применения в исследовательско-поисковой и архивной практике инструментов машинной атрибуции почерка. Предлагается оригинальная методика сопоставления почерков, подразумевающая использование сиамской нейронной сети. Описываются результаты ее применения в атрибуции писем неустановленных адресатов к Г.Н. Потанину (на материале писем, хранящихся в Красноярском краевом краеведческом музее).

Ключевые слова: русская литература, Г.Н. Потанин, рукопись, архивное описание, цифровая копия, сиамские нейронные сети, идентификация и верификация почерка автора

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

Для цитирования: Киселев В.С., Пронина Н.М. Машинная атрибуция почерка в решении источниковедческих проблем (на материале переписки Г.Н. Потанина) // Имагология и компаративистика. 2025. № 24. С. 186–208. doi: 10.17223/24099554/24/9

Original article
doi: 10.17223/24099554/24/9

Machine attribution of handwriting in solving source studies problems (based on Grigory Potanin's correspondence)

Vitaly S. Kiselev[1]
Natalia M. Pronina[2]

[1] *Tomsk State University, Tomsk, Russian Federation, kv-uliss@mail.ru*
[2] *Lomonosov Moscow State University, Moscow, Russian Federation, natalka-pronina@mail.ru*

Abstract. This article examines the potential applications of automated handwriting detection tools for research, archival retrieval, and archival practice. It proposes an original methodology for handwriting comparison based on a Siamese neural network. The article describes the results of applying this method to attribute letters from unidentified correspondents to Grigory N. Potanin, using correspondence held in the Krasnoyarsk Regional Museum of Local Lore.

Keywords: Russian literature, Grigory N. Potanin, manuscript, archival description, digital copy, Siamese neural networks, identification and verification of the author's handwriting

Financial support: The research was conducted at Tomsk State University and supported by the Russian Science Foundation, Project No. 22-68-00066: Cultural Heritage of Russia: Intellectual Analysis and Thematic Modeling of the Corpus of Handwritten Texts.

For citation: Kiselev, V.S. & Pronina, N.M. (2025) Machine attribution of handwriting in solving source studies problems (based on Grigory Potanin's correspondence). *Imagologiya i komparativistika – Imagology and Comparative Studies*. 24. pp. 186–208. (In Russian). doi: 10.17223/24099554/24/9

Введение

Рукописное наследие, хранящееся в архивах различных стран мира, огромно и разноаспектно. Совокупный его объем исчисляется многими миллиардами листов – и даже в тех случаях, когда перед нами источники, важные для того или иного национального канона, в

научный и культурный оборот введена малая их часть. Возможности ученых здесь ограничены с ресурсной стороны: при самой широкой эрудиции и знании архивных источников «ручная» их обработка требует много времени и сил без гарантии того, что самые значимые документы не окажутся пропущенными, затерявшись в массе других рукописей. Это сказывается и на деятельности архивистов: фонды многих лиц и институций (кружков, объединений, организаций и т.п.) долгое время остаются не разобранными и не описанными по причине большой трудоемкости подобной работы.

Машинный анализ данных предоставляет в данной сфере новые инструменты, способные радикально изменить подходы к обработке рукописных источников. Архивные фонды и хранящиеся в них рукописи все более массированно переводятся в цифровой формат (сканируются) и предстают перед исследователем как изображения, а это открывает возможности графической, синтаксической и семантической классификации последних с целью разработки алгоритмов, настроенных на тип документа и позволяющих автоматически его анализировать.

Одной из важнейших задач в этой сфере выступает машинная атрибуция почерка. Ее решение открывает новые перспективы в поиске и классификации рукописных документов. С одной стороны, программа машинной атрибуции значительно расширяет возможности исследователей, в частности, литературоведов или историков, в плане выявления в архивных собраниях текстов того или иного лица. С другой стороны, она дает архивным работникам удобный инструмент для автоматической классификации: выявленный программой реестр соответствий почерка может служить основой для описания рукописи (или ее состава) и внесения соответствующей информации в каталог. При использовании программы автоматизированной атрибуции поиск нужной информации в растровом массиве нераспознанного текста будет занимать не месяцы и годы, а несколько минут. Более полное описание контекста данной проблемы дано в нашей статье [1].

Материал

Репрезентативный материал здесь предлагает творческое наследие Г.Н. Потанина, литератора, журналиста, ученого, путешественника и

общественного деятеля второй половины XIX – начала XX в., инициатора областнического движения в Сибири. Выявленные на сегодняшний день творческие рукописи и эгодокументы самого поэта и лиц, с ним связанных через переписку или по другим каналам коммуникации, оказались разбросаны по всей России (поиск подобных документов в зарубежных архивах только начинается, среди перспективных направлений – Казахстан, Украина, Китай, Монголия, Франция, Англия, Германия и Польша). Если говорить только о российских архивах, то среди них фактически все центральные (ГАРФ, АВПРИ, РГАЛИ, НИОР РГБ, ОРФ ГЛМ, АРАН, СПБФ АРАН, ОР РНБ, РО ИРЛИ, РГИА, СПБИИ РАН, АВ ИВР РАН, НА РГО, обследование продолжается) и многие региональные (Отдел рукописей и книжных памятников Научной библиотеки (а также Музей археологии и этнографии Сибири) Томского государственного университета, Государственный музей истории литературы, искусства и культуры Алтая, Государственные архивы Иркутской и Новосибирской областей, Государственный архив Красноярского края, Исторический архив Омской области, Иркутский областной художественный музей, Красноярский краевой краеведческий музей, Омский государственный историко-краеведческий музей, Томский областной краеведческий музей, Минусинский краеведческий музей имени Н.М. Мартьянова, Хабаровский краевой музей, Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН, Научный архив Калмыкского института гуманитарных исследований РАН, Отдел рукописей и редких книг научной библиотеки им. Н.И. Лобачевского Казанского федерального университета; обследование продолжается).

Как и в случае с многими другими российскими историческими персонажами, степень введения в научный и культурный оборот рукописного наследия Потанина – при всей его значимости и безусловной актуальности – остается невысокой. Из крупных публикаций здесь можно назвать, пожалуй, только два тома «Литературного наследства Сибири» [2, 3], в которых была собрана часть биографических материалов родоначальника областничества, и пятитомник его «Писем» [4]. К сожалению, эти книги были подготовлены на низком источниковедческом и текстологическом уровне.

Помимо обширного научного, публицистического и литературно-критического наследия, Потанин оставил богатейший эпистолярий. В ходе архивных и источниковедческих разысканий было установлено, что его объем составляет около 1 400 писем самого ученого и около 7 000 писем, адресованных ему. Из писем Потанина на сегодняшний день опубликовано около 1 000, из писем его корреспондентов – менее 400 [5].

Совокупный список адресатов и корреспондентов Потанина велик (1 559 позиций) и включает в себя большинство представителей культурной элиты Сибири и значительную часть ученого сообщества Европейской России второй половины XIX – начала XX веков. Его разработка, составление справок об адресатах / корреспондентах, иногда малоизвестных, уточнение биографических деталей, установление датировок составляют отдельную исследовательскую задачу, решение которой позволит реконструировать персоносферу эпохи, включая итинерарий разных лиц, весьма важный для активно передвигавшихся по России и Сибири ученых, купцов, инженеров, учителей и врачей, земских деятелей, чиновников и администраторов, революционеров, священников, да и просто грамотных крестьян или мещан.

В решении источниковедческих проблем, связанных с потанинским эпистолярием, важную роль могут сыграть машинные методы. В хорошо разобранном и атрибутированном архивистами материале его переписки содержится тем не менее небольшой комплекс писем сомнительного авторства. Среди писем Потанина таких посланий неустановленным адресатам совсем мало (6 документов). Однако писем с неустановленным авторством среди адресованных Потанину несравненно больше (всего 89), и они составляют серьезную проблему в связи с большим кругом возможных претендентов. Вручную сопоставить почерк этих писем с образцами почерка почти 1 500 точно установленных корреспондентов ученого – почти неразрешимая задача. Найти подходы к ней, однако, представляется крайне важным.

Постановка задачи

С этой целью было решено использовать методику, разработанную для машинной атрибуции рукописей В.А. Жуковского в подбор-

ках графических файлов (сканов) с помощью сиамских нейронных сетей. Ее особенности мы уже описывали в отдельной статье [6]. Для адаптации и проверки методики был сформирован большой массив графических изображений, содержащий сканы автографов писем с образцами почерка корреспондентов Потанина. На этих образцах должно было произойти научение нейронной сети. Основой выступил корпус отсканированных документов, хранящихся в Красноярском краевом краеведческом музее (далее – КККМ). Этот блок потанинского архива в начале 1920-х гг. достался историку и этнографу Н.Н. Козьмину и впоследствии был передан в музей его семьей. В состав подборки вошло 2 604 документа, принадлежащих перу 811 корреспондентов.

Второй массив был сформирован из сканов автографов неустановленных лиц. Таких неатрибутированных писем в собрании КККМ оказалось 40. Обученная на основе 811 достоверно установленных образцов почерка нейронная сеть должна была определить, с какими из них имеет максимальный процент близости каждый из сканов второй подборки.

Основные понятия и подходы, определяющие постановку математического аспекта задачи (сиамские нейронные сети, contrastive loss function и triplet loss function), были определены нами в уже упомянутой статье [6. С. 160–163].

Предварительная обработка

Для эффективного обучения сетей изображения, на которых оно происходило, подверглись предварительной обработке.

Чтобы предотвратить переобучение модели, т.е. плохую обобщающую способность модели по причине того, что в данных присутствует дополнительная информация – фон, от которой не должен зависеть результат, мы бинаризировали изображения с помощью предобученной нейронной сети DocEnTr [7], состоящей из трансформерного энкодера и декодера. Входное изображение разбивается на фрагменты, которые преобразуются в эмбеддинги, к ним добавляется информация о местоположении фрагмента. Результирующая последовательность векторов подается в энкодер для получения скрытых пред-

ставлений, далее полученные скрытые представления подаются в декодер для получения вектора, который линейно проецируется на векторы пикселей, представляющих участки выходного изображения. Для предотвращения попадания «пустых» изображений в обучающую выборку аугментация производилась таким образом, чтобы доля белых пикселей была не более 95%.

Для обучения строчной модели необходимо нарезать на строки каждую страницу. Выделение строк производилось с помощью python-библиотеки kraken [8].

Обучение сиамской сети происходило по двум моделям – картиночной и строчной.

Картиночная модель

Применим идею обучения сиамской сети на бинарных изображениях, фрагментах размера 300×300 пикселей. Возьмём предобученный ResNet18. Последний слой имеет размерность 1 000 на выходе, поэтому сиамская сеть будет выучивать 1 000-размерный эмбеддинг изображения. Будем обучать последние два слоя, 513 000 обучаемых параметров (рис. 1).

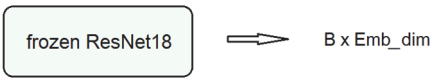


Рис. 1. Архитектура картиночной модели

При обучении в сеть будут подаваться тройки изображений, триплеты: anchor, positive, negative. В качестве anchor подаётся очередная строка, в качестве positive – случайная строка того же автора, в качестве negative – случайная строка другого автора. При оценке качества сиамской сети точность считается как доля триплетов, для которых евклидово расстояние между эмбеддингами anchor и positive меньше, чем между anchor и negative. На полученных эмбеддингах обучим классификатор – двухслойную полносвязную нейронную сеть с 513 538 параметрами.

Итоговое предсказание для письма производилось с помощью агрегации выходов сети для нескольких случайных фрагментов данного письма. В качестве функции агрегации было выбрано усреднение.

Строчная модель

В качестве строчной была выбрана гибридная модель, сочетающая свёрточные и рекуррентный слои, для обработки последовательностей данных (строк) она будет учиться со случайной инициализацией (рис. 2).

Свёрточный блок состоит из трёх слоёв по типу: свёртка, нормализация по батчу, активация, max-пуллинг с асимметричным шагом, для того чтобы по высоте строки сжимать меньше, чем по ширине. В качестве рекуррентного слоя выбрана однонаправленная LSTM (205 507 обучаемых параметров).

Рис. 2. Архитекттура строчной модели

Классификатор тот же самый – 513 538 параметров.

Итоговое предсказание для письма производилось с помощью усреднения выходов сети для всех строк данного письма.

Эксперименты

Были проведены эксперименты на разном числе классов и с разным способом обучения. Всюду далее определение balanced (сбалансированный) указывает на то, что при обучении модели объекты подавались равновероятно для разных классов. По умолчанию же используется обычный способ – каждый объект подаётся равновероятно, поэтому классы с большим числом примеров при обучении встречаются чаще остальных. В таком смысле обучение является несбалансированным.

Картиночная модель

Картиночная модель учится классифицировать фрагменты рукописных документов размера 300×300 пикселей, но итоговое предсказание для изображения производится на основе нескольких случайных фрагментов. Исследуем, какое количество изображений оптимально для качественного определения автора данного письма.

В табл. 1 представлена зависимость качества картиночной модели от числа фрагментов в задачах классификации на 3, 10, 20, 100 классов. Можно заметить, что увеличение количества фрагментов при агрегации практически всегда приводит к увеличению и accuracy, и F1-меры. Поэтому всюду далее будем брать максимальное число фрагментов для классификации одного документа – 20 штук.

Таблица 1

Accuracy и F1-мера для картиночной модели

Metric / Windows num	1	5	10	20
3 classes, Hits@	0.9209	0.9379	0.9322	0.9492
3 classes, Macro-F1@1	0.9203	0.9385	0.9338	0.9510
10 classes, Hits@1	0.7073	0.7878	0.7854	0.8171
10 classes, Macro-F1@1	0.6856	0.7766	0.7768	0.8091
20 classes, Hits@1	0.6416	0.7621	0.7731	0.7731
20 classes, Macro-F1@1	0.6198	0.7363	0.7598	0.7566
100 classes, Hits@1	0.4415	0.5549	0.5650	0.5882
100 classes, Macro-F1@1	0.3208	0.4239	0.4286	0.4624

Таблица 2

Accuracy и F1-мера для картиночной модели при сбалансированном обучении

Metric / Windows num	1	5	10	20
3 classes, Hits@	0.9209	0.9209	**0.9435**	0.9209
3 classes, Macro-F1@1	0.9217	0.9217	**0.9469**	0.9220
10 classes, Hits@1	0.6854	0.7488	0.7732	0.7634
10 classes, Macro-F1@1	0.6856	0.7408	0.7656	0.7567
20 classes, Hits@1	0.5696	0.6682	0.6854	0.6964
20 classes, Macro-F1@1	0.5496	0.6331	0.6584	0.6705
100 classes, Hits@1	0.3642	0.4783	0.4993	0.5094
100 classes, Macro-F1@1	0.3146	**0.4251**	**0.4457**	0.4609

Проведём такой же эксперимент со сбалансированным обучением сети, исследуем, какой эффект это окажет на F1-меру. Сравнение табл. 1 и 2 показывает, что balanced-обучение даёт незначительное улучшение в редких случаях (такие случаи выделены жирным шрифтом), в основном результат становится немного хуже. Вероятно, это объясняется тем, что малые классы и так легко детектируются, поэтому увеличение частоты их встречаемости при обучении не даёт большого прироста качества.

Строчная модель

Проведём эксперименты для строчной модели и сравним с картиночной. В табл. 3 для картиночной модели приведены лучшие результаты с 20 окнами при несбалансированном обучении. Для строчной модели уже виден прирост в качестве при переходе к сбалансированному обучению (выделено жирным шрифтом), особенно заметен прирост в F1-мере в многоклассовой классификации. В большинстве случаев строчная модель справляется лучше, но по качеству классификации на 100 классов всё же уступает.

Таблица 3

Сравнение метрик для картиночной и строчной модели

Metric / Model	10 classes			20 classes			100 classes		
	win	str	bal	win	str	bal	win	str	bal
Hits@1	0.817	0.657	**0.694**	0.773	**0.552**	0.515	0.588	0.526	**0.582**
Hits@2	0.915	0.813	**0.851**	0.889	**0.761**	0.727	0.692	0.657	**0.712**
Hits@3	0.946	0.888	**0.908**	0.930	**0.832**	0.815	0.758	0.741	**0.771**
Hits@4	0.968	0.935	**0.943**	0.955	**0.880**	0.859	0.808	0.796	**0.812**
Hits@5	0.980	**0.965**	0.960	0.969	**0.919**	0.905	0.852	0.835	**0.847**
Macro-F1@1	0.809	0.585	**0.676**	0.757	0.484	**0.497**	0.462	0.395	**0.579**
Macro-F1@2	0.913	0.786	**0.843**	0.884	**0.724**	0.722	0.594	0.564	**0.708**
Macro-F1@3	0.948	0.884	**0.908**	0.927	0.801	**0.812**	0.681	0.674	**0.765**
Macro-F1@4	0.969	0.938	**0.946**	0.954	0.861	**0.861**	0.745	0.743	**0.804**
Macro-F1@5	0.981	**0.968**	0.962	0.968	0.904	**0.904**	0.796	0.793	**0.840**

Для картиночной модели используется предобученная сеть с большим числом параметров (205 507 против 513 000), попробуем улучшить строчную модель. Проведём эксперименты ещё с двумя архитектурами строчной модели: предобученный ResNet18 и изменённая архитектура (10) с 4 слоями в энкодере вместо 3, с двунаправленным LSTM-блоком (770 936 обучаемых параметров).

Таблица 4

Сравнение метрик для разных архитектур строчных моделей

Metric / Model	10 classes			20 classes			100 classes		
	win	str	bal	win	str	bal	win	str	bal
Hits@1	0.694	0.968	0.973	0.515	0.862	0.910	0.582	0.715	0.782
Hits@2	0.851	0.990	0.990	0.727	0.943	0.957	0.712	0.818	0.876
Hits@3	0.908	0.990	0.990	0.815	0.965	0.971	0.771	0.855	0.921
Hits@4	0.943	0.998	0.998	0.859	0.975	0.981	0.812	0.879	0.943
Hits@5	0.960	0.998	0.998	0.905	0.983	0.987	0.847	0.901	0.965
Macro-F1@1	0.676	0.967	0.972	0.497	0.860	0.894	0.579	0.739	0.727
Macro-F1@2	0.843	0.991	0.990	0.722	0.945	0.950	0.708	0.836	0.849
Macro-F1@3	0.908	0.991	0.990	0.812	0.968	0.968	0.765	0.871	0.902
Macro-F1@4	0.946	0.997	0.997	0.861	0.976	0.980	0.804	0.893	0.935
Macro-F1@5	0.962	0.997	0.997	0.904	0.983	0.988	0.840	0.911	0.961

В табл. 4 представлены лучшие результаты строчной модели при сбалансированном обучении и лучшие результаты двух новых строчных моделей. Для ResNet качество лучше при сбалансированном обучении, для большой LSTM-модели – при несбалансированном, но в обоих случаях качество заметно выше, чем у картиночной.

Определение неизвестного класса

После обучения модели на задачу классификации с функцией потерь перекрёстной энтропии модель не в состоянии «отказаться от классификации», если встречает объект класса, который не был представлен в обучающих данных, но в текущей постановке задачи обяза-

тельно нужно иметь такую возможность. Такая классификация называется классификацией в открытом мире (open-world classification) или открытой классификацией (open classification).

Идея перехода к открытой классификации заключается в том, чтобы заменить стандартный softmax-слой, так как функция softmax по умолчанию распределяет 100% вероятности между известными классами, не оставляя места для неизвестных. Заменим softmax на функцию сигмоиды, по одной на каждый класс, функцию потерь заменим на бинарную перекрёстную энтропию [9]. Таким образом, задача мультиклассовой классификации, где softmax даёт взаимоисключающие вероятности, переходит в задачу множественной бинарной классификации, где сигмоиды позволяют объекту принадлежать к нескольким классам или ни к одному.

Итоговое предсказание модели строится следующим образом: если сигмоиды всех классов меньше наперёд заданного порога (использовался порог, равный 0.5), то объект классифицируется как неизвестный, иначе выдаётся класс с наибольшим значением сигмоиды.

Эксперименты показали, что подобная замена последнего softmax-слоя и функции потерь практически никак не влияет на метрики качества, полученные на обучающей и валидационных выборках.

Решение

Был предложен подход к задаче атрибуции рукописных писем, основанный на использовании сиамской нейронной сети для сравнения и анализа уникальных характеристик почерка. Предложенное решение позволяет не только сравнить рукописи и определить принадлежность их известным авторам, но и выделить документы, чьи создатели отсутствуют в архиве.

Основным результатом является алгоритм, входом которого является сканированный неатрибутированный документ, а на выходе выдаётся ранжированный по убыванию вероятности список возможных его авторов, а также вероятность того, что автор документа неизвестен, т.е. не упоминался в архиве. Полученное решение задачи атрибуции включает предобработку сканированных документов и качественное обучение сиамских нейронных сетей.

Исследованы два варианта анализа сканированного текста: анализ фрагментов изображения и анализ каждой строки рукописного текста от-

дельно. В первом случае (для картиночной модели) в качестве архитектуры сиамской сети использовался ResNet18; во втором случае (для строчной модели) для обработки последовательностей (строк) использовалась гибридная модель, сочетающая свёрточные и рекуррентный слои.

Эксперименты показали, что модели устойчивы к дисбалансу данных, способны работать даже с малым числом образцов почерка, показывая высокое значение F1-меры, подтвердили практическую применимость метода для задач атрибуции.

Результаты машинной атрибуции

В источниковедческом плане применение разработанного алгоритма позволило получить следующие результаты. Для каждого из 40 неатрибутированных писем нейронная сеть подбирала соответствия из 811 образцов почерка и определяла в процентах степень близости. Для удобства последующей «ручной» проверки были отобраны лишь наиболее схожие почерки (пять образцов с максимальными показателями сходства). Эти прогнозы представлены в табл. 5.

Таблица 5

Результаты машинного определения почерка

Музейный шифр письма	Прогноз авторства (%)
КККМ ОФ 7928/957	23.09 Ржавская Надежда Федоровна 17.18 Проскурякова Е. Павловна 16.49 Фарафонтова Таисия Михайловна 15.86 Свентицкая Мария Хрисанфовна 5.66 Крутовский Владимир Михайлович
КККМ ОФ 7928/1391	42.17 Ошурков Василий Александрович 10.84 Ивановский Алексей Осипович 8.48 Адрианов Александр Васильевич 6.68 Зобнин Филипп Козмич 6.57 Багашев Иван Васильевич
КККМ ОФ 7928/1392	24.22 Вагнер (Дзвонкевич) Екатерина Николаевна 22.13 Талько-Гринцевич Юлиан Доминикович 5.90 Штильке Василий Константинович 5.53 Сапожников Василий Васильевич 5.14 Лаврский Константин Викторович
КККМ ОФ 7928/1395	16.30 Гуркин Григорий Иванович 15.45 Анучин Дмитрий Николаевич 14.05 Ошурков Василий Александрович

Музейный шифр письма	Прогноз авторства (%)
	9.96 Лаврский Валериан Викторович 8.06 Серошевский Вацлав Леопольдович
КККМ ОФ 7928/1513	32.02 Фарафонтова Таисия Михайловна 18.73 Круговская Лидия Симоновна 10.63 Попов Иван Иванович 7.05 Вагнер (Дзвонкевич) Екатерина Николаевна 6.18 Карпова Наталия Петровна
КККМ ОФ 7928/1894	31.88 Мендельсон Николай Михайлович 8.91 Головачев Петр Михайлович 5.20 Руднев Андрей Дмитриевич 5.12 Васильева Агния Евгеньевна 5.04 Сибиряков Иннокентий Михайлович
КККМ ОФ 7928/2335	38.07 Семидалов Вениамин Иванович 12.59 Никифоров Николай Яковлевич 11.82 Лаврская Екатерина Васильевна 7.76 Адрианов Александр Васильевич 4.63 Потанина Александра Викторовна
КККМ ОФ 7928/2336	39.29 Свентицкая Мария Хрисанфовна 36.66 Белоголовая Надежда Александровна 3.89 Ивановский Алексей Осипович 3.21 Ядринцева (Злобина) Лидия Николаевна 2.80 Фарафонтова Таисия Михайловна
КККМ ОФ 7928/2340	23.30 Серошевский Вацлав Леопольдович 9.17 Фарафонтова Таисия Михайловна 9.11 Баландина Вера Арсеньевна 9.07 Педашенко-Третьякова Мария Ивановна 7.43 Семидалов Вениамин Иванович
КККМ ОФ 7928/2480	86.99 Вагнер (Дзвонкевич) Екатерина Николаевна 5.31 Карпова Наталия Петровна 2.52 Белоголовая Надежда Александровна 1.03 Муромов И.Г. 0.41 Леонович (Ангарский) Василий Викторович
КККМ ОФ 7928/2486	23.45 Муромов И.Г. 11.95 Пантелеев Лонгин Федорович 6.66 Антонова Валентина Константиновна 6.00 Лаврский Константин Викторович 4.92 Костюрина Мария Николаевна
КККМ ОФ 7928/2651	12.20 Сорокин Павел 8.36 Проскурякова Е. Павловна 8.29 Зобнин Филипп Козмич 7.69 Костюрина Мария Николаевна

Музейный шифр письма	Прогноз авторства (%)
	7.40 Фарафонтова Таисия Михайловна
КККМ ОФ 7928/2701	9.52 Третьякова Анна Иосифовна (Осиповна) 8.02 Солдатов Владимир Константинович 7.81 Штильке Василий Константинович 5.94 Антонова Валентина Константиновна 5.40 Вагнер (Дзвонкевич) Екатерина Николаевна
КККМ ОФ 7928/2702	27.18 Никифоров Николай Яковлевич 13.96 Ивановский Алексей Осипович 9.36 Проскурякова Е. Павловна 4.99 Обручев Владимир Афанасьевич 4.07 Зобнин Филипп Козмич
КККМ ОФ 7928/2704	21.99 Антонова Валентина Константиновна 9.79 Проскурякова Е. Павловна 7.99 Серошевский Вацлав Леопольдович 7.42 Свентицкая Мария Хрисанфовна 7.18 Крутовский Владимир Михайлович
КККМ ОФ 7928/2705	19.07 Леонович (Ангарский) Василий Викторович 9.57 Козлова Елизавета Митрофановна 6.19 Григорьев Александр Васильевич 6.08 Круковский Михаил Антонович 5.73 Курский Михаил Онисифорович
КККМ ОФ 7928/2706	12.49 Серошевский Вацлав Леопольдович 10.57 Мендельсон Николай Михайлович 8.53 Фарафонтова Таисия Михайловна 7.54 Антонова Валентина Константиновна 6.70 Раппопорт (Ан-ский) Семен Акимович
КККМ ОФ 7928/2707	34.36 Проскурякова Е. Павловна 14.30 Потанина Александра Викторовна 13.33 Ржавская Надежда Федоровна 5.34 Веселовский Александр Николаевич 3.25 Лонгиновский Карп Дмитриевич
КККМ ОФ 7928/2708	22.95 Миклашевская (Лаврская) Софья Львовна 15.86 Потанина Александра Викторовна 13.70 Попов Иван Иванович 12.34 Козлова Елизавета Митрофановна 10.09 Козьмин Николай Николаевич
КККМ ОФ 7928/2711	22.99 Миклашевская (Лаврская) Софья Львовна 10.91 Базанова Лидия Павловна 8.66 Леонович (Ангарский) Василий Викторович 8.52 Крутовский Владимир Михайлович 6.78 Козлова Елизавета Митрофановна

Музейный шифр письма	Прогноз авторства (%)
КККМ ОФ 7928/2712	24.47 Сибирякова Анна Михайловна 16.41 Крутовский Владимир Михайлович 8.73 Свентицкая Мария Хрисанфовна 7.50 Ватсон Л. 7.33 Фарафонтова Таисия Михайловна
КККМ ОФ 7928/2713	26.83 Свентицкая Мария Хрисанфовна 11.28 Семидалов Вениамин Иванович 8.77 Сибирякова Анна Михайловна 7.18 Проскурякова Е. Павловна 6.86 Ватсон Л.
КККМ ОФ 7928/2714	18.26 Миклашевская (Лаврская) Софья Львовна 13.24 Белоголовая Надежда Александровна 8.86 Потанина Александра Викторовна 8.66 Проскурякова Е. Павловна 8.28 Сибирякова Анна Михайловна
КККМ ОФ 7928/2715	6.84 Серошевский Вацлав Леопольдович 5.85 Лаврский Константин Викторович 5.17 Фарафонтова Таисия Михайловна 4.86 Ошурков Василий Александрович 4.85 Антонова Валентина Константиновна
КККМ ОФ 7928/2716	27.55 Мендельсон Николай Михайлович 14.41 Потанина Александра Викторовна 9.41 Наумов Николай Николаевич 6.22 Анучин Дмитрий Николаевич 6.06 Базанова Лидия Павловна
КККМ ОФ 7928/2717	23.25 Фарафонтова Таисия Михайловна 19.23 Карпова Наталия Петровна 15.74 Тиблен Ольга Николаевна 5.31 Попов Иван Иванович 3.12 Красноженова Мария Васильевна
КККМ ОФ 7928/2719	88.71 Фарафонтова Таисия Михайловна 4.93 Свентицкая Мария Хрисанфовна 3.35 Крутовская Лидия Симоновна 1.08 Ватсон Л. 0.45 Сибирякова Анна Михайловна
КККМ ОФ 7928/2720	29.83 Григорьев Александр Васильевич 21.13 Киселева Ольга 12.37 Сукачев Владимир Платонович 12.03 Попов Иван Иванович 4.26 Сибирякова Анна Михайловна
КККМ ОФ 7928/2721	20.07 Григорьев Александр Васильевич

Музейный шифр письма	Прогноз авторства (%)
	12.29 Круковский Михаил Антонович 11.59 Козлова Елизавета Митрофановна 6.73 Костюрина Мария Николаевна 5.83 Попов Иван Иванович
КККМ ОФ 7928/2722	24.28 Мендельсон Николай Михайлович 9.50 Попов Иван Иванович 8.30 Солдатов Владимир Константинович 5.37 Гуркин Григорий Иванович 4.84 Скалозубов Николай Лукич
КККМ ОФ 7928/2723	31.68 Мендельсон Николай Михайлович 8.65 Адрианов Александр Васильевич 7.78 Солдатов Владимир Константинович 5.65 Попов Иван Иванович 5.59 Головачев Петр Михайлович
КККМ ОФ 7928/2724	16.60 Антонова Валентина Константиновна 10.26 Руднев Андрей Дмитриевич 10.21 Базанова Лидия Павловна 7.49 Круковский Михаил Антонович 6.41 Костюрина Мария Николаевна
КККМ ОФ 7928/2725	15.72 Антонова Валентина Константиновна 12.26 Козлова Елизавета Митрофановна 7.63 Попов Иван Иванович 7.61 Миклашевская (Лаврская) Софья Львовна 7.44 Балакшин Александр Николаевич
КККМ ОФ 7928/2726	18.94 Фарафонтова Таисия Михайловна 9.70 Крутовская Лидия Симоновна 7.91 Руднев Андрей Дмитриевич 7.13 Семидалов Вениамин Иванович 6.72 Серошевский Вацлав Леопольдович
КККМ ОФ 7928/2727	17.14 Гуркин Григорий Иванович 9.56 Педашенко-Третьякова Мария Ивановна 7.33 Серошевский Вацлав Леопольдович 7.27 Тихонравова Раиса 5.74 Анучин Дмитрий Николаевич
КККМ ОФ 7928/2728	20.84 Ивановский Алексей Осипович 11.46 Адрианов Александр Васильевич 8.99 Серошевский Вацлав Леопольдович 6.91 Гуркин Григорий Иванович 4.97 Макеров Яков Антонович
КККМ ОФ 7928/2830	12.39 Муромов И.Г. 10.52 Головачев Петр Михайлович

Музейный шифр письма	Прогноз авторства (%)
	10.09 Мендельсон Николай Михайлович 7.80 Руднев Андрей Дмитриевич 7.02 Досекина Зинаида
КККМ ОФ 7928/2831	21.04 Антонова Валентина Константиновна 15.46 Потанина Александра Викторовна 14.51 Наумов Николай Николаевич 3.98 Костюрин Виктор Федорович 3.88 Мендельсон Николай Михайлович
КККМ ОФ 7928/2832	22.59 Ржавская Надежда Федоровна 19.90 Проскурякова Е. Павловна 18.53 Лонгиновский Карп Дмитриевич 7.47 Свентицкая Мария Хрисанфовна 5.96 Фарафонтова Таисия Михайловна
КККМ ОФ 7928/2833	16.97 Серошевский Вацлав Леопольдович 12.34 Семевский Василий Иванович 10.90 Фарафонтова Таисия Михайловна 7.36 Гуркин Григорий Иванович 5.43 Бейлин Соломон Хаимович

Верификация полученных результатов

Полученные прогнозы авторства подверглись «ручной» проверке.

Во-первых, грамматические показатели (окончания глаголов) рукописи позволили определить гендерную принадлежность автора и, соответственно, убрать из прогноза несоответствующие ей варианты. Для дальнейшей доработки алгоритма было предложено ввести дополнительный аспект в виде инструментов машинного определения гендера (по грамматическим признакам), что позволит ограничить проводимые сопоставления только одним классом (авторы-мужчины / авторы-женщины).

Во-вторых, для тех из 40 писем, где присутствовали обозначения времени (число, месяц, год) и места написания, было проведено сопоставление с тем, чтобы установить, мог ли этот человек в данный момент находиться в указанном месте.

Например, в письме с музейным шифром КККМ ОФ 7928/1392 по грамматическим признакам автор был мужчиной, а само оно было написано 13 июня 1891 г. во Владивостоке. Тем самым из числа возможных авторов сразу выпадала Е.Н. Вагнер (Дзвонкевич), а остальные 4 персоны (польский антрополог Ю.Д. Талько-Гринцевич, обще-

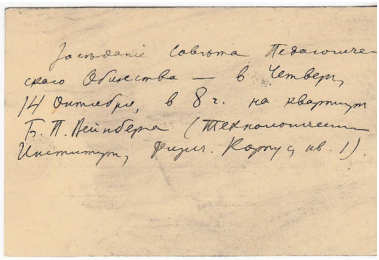
ственный деятель В.К. Штильке, биолог и одно время ректор Томского университета В.В. Сапожников и шурин Потанина журналист К.В. Лаврский) не могли быть в данное время во Владивостоке. Тем самым для указанного письма выяснилось, что в собрании КККМ не присутствовали другие письма с идентичным почерком и его автора нужно искать за пределами списка из 811 корреспондентов.

Третья процедура верификации результатов включала сопоставление реалий, упоминаемых в неатрибутированном письме, с биографическими обстоятельствами лиц, фигурирующих в машинном прогнозе. Здесь важны были любые указания на статус, профессию, семейное положение автора, характер его связи с Потаниным (коллега, родственник, знакомый и т.п.) и другие приметы. Иногда они достаточно очевидны: например, в коротенькой записке с музейным шифром КККМ ОФ 7928/2486 автор подписался как «ректор». Музейные работники не смогли прочесть неразборчивый почерк дальнейшей подписи и отнесли записку к неустановленному лицу. Машинная атрибуция в качестве возможных авторов предложила И.Г. Муромова, Л.Ф. Пантелеева, В.К. Антонова, К.В. Лаврского и М.И. Костюрину. Однако никто из них ректором не являлся. Кроме того, в записке речь шла о поздравлении Потанина с юбилеем и упоминалась жизнь автора в Самаре. Анализ текста позволил вначале определить, что ректорство относилось к Томской духовной семинарии и приходилось на какой-то из двух широко отмечавшихся юбилеев Потанина (1905 или 1915 г.). Кроме того, в Самаре родился, учился и преподавал до 1903 г. только один ректор семинарии этого периода – Иоанн Александрович Панормов. Ректорствовал в Томске он с 1903 по 1907 г., что позволило записку датировать 21 сентября (день рождения Потанина) 1905 г. Приводим полный текст записки с учетом всех реалий: «Достопочтенный Григорий Николаевич! Присоединяю и свое поздравление к Вашим почитателям. Я знал Вас и уважал Вас, не видя Вас, живя в г<ороде> Самаре. Лично я не дошел до Вас за недосугами… Будьте здоровы! Ректор Т<омской> Д<уховной> С<еминарии> прот<оиерей> И. Панор<мов>. 21-ого». Заметим, что в известном на сегодняшний день архиве Потанина других писем, принадлежащих руке И.А. Панормова, не было выявлено, это не позволило машинным способом правильно определить авторство.

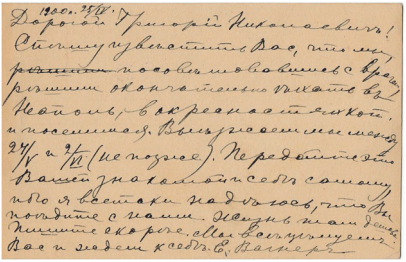
Наконец, в-четвертых, как самая проблематичная процедура проводилось визуальное сличение почерка неатрибутированного письма

и почерка авторов из машинного прогноза. Тотальная визуальная проверка позволила прийти к выводу, что достоверность прогнозов оказывается достаточной только после порога в 85% (все остальные сопоставления оказались далекими по почерку). Таким критериям соответствовали только два: для письма с шифром КККМ ОФ 7928/2480 – Е.Н. Вагнер (Дзвонкевич) (86,99%) и для письма с шифром КККМ ОФ 7928/2719 – Т.М. Фарафонтова (88,71%). После проведения проверок, упомянутых в первых трех пунктах, оба прогноза были признаны реалистичными. Тем не менее визуальное сличение почерков оставляет сомнение в обоих случаях, хотя авторство Е.Н. Вагнер (Дзвонкевич) выглядит более вероятным, чем авторство Т.М. Фарафонтовой.

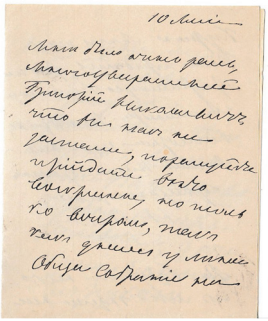
Приводим сканированные изображения данных неатрибутированных писем в сопровождении образцов почерка из записок, имеющих ясную авторскую подпись.



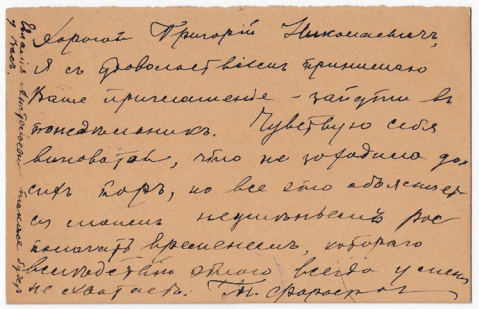
Почерк письма
КККМ ОФ 7928/2480



Почерк Е.Н. Вагнер (Дзвонкевич)
КККМ ОФ 7928/308



Почерк письма
КККМ ОФ 7928/2719



Почерк Т.М. Фарафонтовой
КККМ ОФ 7928/1273

Выводы

Проведенный машинный анализ обнаружил свою эффективность. Он позволил значительно оптимизировать поиск возможных соответствий почерка, т.е. не перебирать вручную все возможные варианты (811 образцов, учитывая, что в разных письмах почерк даже одного автора может варьировать в зависимости от условий написания, возраста, скорости письма и т.п.), а быстро провести ранжирование и выделить несколько наиболее близких – с последующей оценкой по грамматическим категориям (гендеру), биографическим реалиям и визуальному сходству. Малый процент подтверждающихся прогнозов (2 письма из 40) объясняется ограниченностью базы образцов. Скорее всего, в нее не попали реальные авторы из установленных на сегодняшний день полутора тысяч корреспондентов Потанина. Это подводит к необходимости расширить материал эксперимента за счет образцов почерка еще 700 авторов. Составление этой базы сканированных изображений и проведение нового обследования, в том числе с попыткой установления авторства еще 49 неатрибутированных писем из фондов ОР НБ ТГУ, РО ИРЛИ, РГАЛИ, определяет ближайшую перспективу проводимого исследования.

Список источников

1. *Киселев В.С., Лебедева О.Б., Третьяков Е.О.* Проблема машинного выявления текстов с почерком определенного автора в составе больших баз данных растровых изображений рукописных документов (на основе опыта выявления писем В.А. Жуковского в делопроизводственных конволютах РГИА) // Имагология и компаративистика. 2023. № 20. С. 247–262.

2. Литературное наследство Сибири. Новосибирск : Зап.-Сиб. кн. изд-во, 1983. Т. 6. 338 с.

3. Литературное наследство Сибири. Новосибирск : Зап.-Сиб. кн. изд-во, 1986. Т. 7. 344 с.

4. Письма Г.Н. Потанина : в 5 т. Иркутск : Изд-во Иркут. ун-та, 1987–1992.

5. *Киселев В.С.* Переписка Г.Н. Потанина: проблемы источниковедческого описания и машинной атрибуции // Вестник Томского государственного университета. 2025 (в печати).

6. *Киселев В.С., Кропотов Д.А., Пронина Н.М.* Сиамская сеть, машинная атрибуция почерка и неизвестный Жуковский // Имагология и компаративистика. 2024. № 22. С. 156–179.

7. *Souibgui M.A., Biswas S., Jemni S.K., Kessentini Y., Forn´es A., Llad´os J., Pal U.* Docentr: An end-to-end document image enhancement transformer // 2022 26th International Conference on Pattern Recognition (ICPR). Montreal: IEEE, 2022. P. 1699–1705.

8. *Wood D.E., Salzberg S.L.* Kraken: ultrafast metagenomic sequence classification using exact alignments // Genome biology. 2014. № 15. P. 1–12.

9. *Lei Shu, Hu Xu, Bing Liu*. Doc: Deep open classification of text documents // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. P. 2911–2916.

References

1. Kiselev, V.S., Lebedeva, O.B. & Tretyakov, E.O. (2023) The problem of machine identification of texts with the handwriting of a specific author as part of large databases of raster images of handwritten documents (based on the experience of identifying letters from Vasily Zhukovsky in office convolutes of the Russian State Historical Archive). *Imagologiya i komparativistika – Imagology and Comparative Studies*. 20. pp. 247–262. (In Russian). doi: 10.17223/24099554/20/13

2. Yanovsky, N.N. (1983) *Literaturnoe nasledstvo Sibiri* [The Literary Heritage of Siberia]. Vol. 6. Novosibirsk: Zap.-Sib. kn. izd-vo.

3. Yanovsky, N.N. (1983) *Literaturnoe nasledstvo Sibiri* [The Literary Heritage of Siberia]. Vol. 7. Novosibirsk: Zap.-Sib. kn. izd-vo.

4. Potanin, G.N. (1987–1992) *Pis'ma G.N. Potanina: v 5 t.* [Letters of G.N. Potanin: in 5 vols]. Irkutsk: Irkutsk University.

5. Kiselev, V.S. (2025) Perepiska G.N. Potanina: problemy istochnikovedcheskogo opisaniya i mashinnoy atributsii [The correspondence of G.N. Potanin: Problems of source study description and machine attribution]. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. (In print).

6. Kiselev, V.S., Kropotov, D.A. & Pronina, N.M. (2024) Siamese network, machine attribution of the handwriting, and unknown Zhukovsky. *Imagologiya i komparativistika – Imagology and Comparative Studies.* 22. pp. 156–179. (In Russian). doi: 10.17223/24099554/22/10

7. Souibgui, M.A., Biswas, S., Jemni, S.K., Kessentini, Y., Forn´es, A., Llad´os, J. & Pal, U. (2022) Docentr: An end-to-end document image enhancement transformer. *26th International Conference on Pattern Recognition (ICPR)*. Montreal: IEEE. pp. 1699–1705.

8. Wood, D.E. & Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 15. pp. 1–12.

9. Lei Shu, Hu Xu & Bing Liu. (2017) Doc: Deep open classification of text documents. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics. pp. 2911–2916.

Информация об авторах:
Киселев В.С. – д-р филол. наук, заведующий кафедрой русской и зарубежной литературы Томского государственного университета (Томск, Россия). E-mail: kv-uliss@mail.ru
Пронина Н.М. – магистр кафедры математических методов прогнозирования Московского государственного университета имени М.В. Ломоносова (Москва, Россия). E-mail: natalka-pronina@mail.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:
V.S. Kiselev, Dr. Sci. (Philology), head of the Department of Russian and Foreign Literature, Tomsk State University (Tomsk, Russian Federation). E-mail: kv-uliss@mail.ru
N.M. Pronina, student, Lomonosov Moscow State University (Moscow, Russian Federation). E-mail: natalka-pronina@mail.ru

The authors declare no conflicts of interests.

Статья принята к публикации 16.09.2025.

The article was accepted for publication 16.09.2025.