Историческая информатика
Правильная ссылка на статью:
Меховский В.А., Кижнер И.А. Мир глазами образованного человека г. Минусинска конца XIX - начала XX веков: распределение частотности географических названий в книгах Минусинской общественной библиотеки // Историческая информатика. 2025. № 1. DOI: 10.7256/2585-7797.2025.1.72586 EDN: QCQWHG URL: https://nbpublish.com/library_read_article.php?id=72586

Мир глазами образованного человека г. Минусинска конца XIX - начала XX веков: распределение частотности географических названий в книгах Минусинской общественной библиотеки

Меховский Вадим Александрович

ORCID: 0009-0000-7786-0939

магистр; кафедра информационных технологий в креативных и культурных индустриях; Сибирский федеральный университет
Специалист лаборатории DHlab; Лаборатория DHlab; Сибирский федеральный университет

660130, Россия, Красноярский край, г. Красноярск, ул. Свободный, 82, оф. 402

mehovsky.zenit-champion@yandex.ru

Кижнер Инна Александровна

ORCID: 0000-0002-0775-9656

кандидат культурологии

доцент; кафедра информационных технологий в креативных и культурных индустриях; Сибирский федеральный университет
Старший научный сотрудник лаборатории DHlab; Сибирский федеральный университет

660074, Россия, Красноярский край, г. Красноярск, ул. Свободный, 82, ст1, оф. 440

inna.kizhner@gmail.com

Статья из рубрики "Искусственный интеллект и наука о данных"

DOI:

10.7256/2585-7797.2025.1.72586

EDN:

QCQWHG

Дата направления статьи в редакцию:

05-12-2024

Аннотация: Предметом исследования является корпус детской литературы из собрания

Минусинской общественной библиотеки конца XIX – начала XX века, состоящий из 121 произведения, написанных между 1719 и 1905 годами. Эти тексты представляют собой значимый источник для изучения формирования географического восприятия у жителей провинциального сибирского города через художественную литературу. Особое внимание уделено анализу географических названий (топонимов), встречающихся в текстах, с целью выявления их частоты и географического распределения. Это позволяет реконструировать картину мира, представленную в книгах того времени, и понять, как она воспринималась детской аудиторией, формируя их представление о странах, городах и культурных центрах. Работа направлена на изучение роли детской литературы как культурного инструмента, который отражает и формирует географические представления, а также на выявление методологических вызовов и ограничений при работе с историческими корпусами. Методологическая основа включает приведение дореформенных текстов к машиночитаемому виду с использованием инструментов оцифровки и геопарсинг для автоматического выявления географических сущностей. Для анализа применялась библиотека Spacy с последующей ручной проверкой и корректировкой данных. Результаты исследования включают выявление 668 городов и 97 стран, представленных в текстах, а также построение картографической визуализации частотного распределения упоминаний. Анализ выявил неравномерность распределения географических наименований в различных текстах, где преобладают упоминания России, Польши и Англии среди стран, а Киева, Москвы и Санкт-Петербурга среди городов. Область применения результатов включает исследования в области цифровых гуманитарных наук, библиотековедения и историко-культурных исследований. Новизна же работы заключается в использовании современных методов геопарсинга для обработки русскоязычных текстов дореформенной орфографии и в анализе ранее не изученного корпуса литературы Минусинской библиотеки. Выводы подчеркивают значимость картирования текстов для понимания формирования географического восприятия и необходимость дальнейшего развития инструментов NER для сложных корпусов. Несмотря на ограничения, исследование вносит вклад в развитие методов NLP для исторических текстов.

Ключевые слова:

Геопарсинг, Картирование, Выявление именованных сущностей, Историческая информатика, Сибирь, Минусинск, Карта мира, Детская литература, Минусинская общественная библиотека, Дореформенная орфография

Введение

В процессе обработки естественного языка (Natural Language Processing, NLP) важную роль играют задачи выявления именованных сущностей (Named Entity Recognition, NER) и их картирования (mapping). Эти задачи включают идентификацию и классификацию различных элементов текста, таких как имена людей, названия организаций и географические объекты, с последующей привязкой их к определенным категориям или базам знаний. В последние годы подходы к NER претерпели существенные изменения благодаря развитию методов глубокого обучения. Однако, несмотря на значительные достижения, большинство исследований продолжают фокусироваться на широко известных корпусах, таких как английский, китайский или арабский языки, в то время как русский язык остается относительно малоисследованным.

Настоящее исследование направлено на выявление именованных сущностей и их

картирование в корпусе детской литературы из собрания книг Минусинской общественной библиотеки конца XIX – начала XX века. В работе представлен алгоритм геопарсинга и картирования географических именованных сущностей в текстах, написанных дореформенной орфографией. Особое внимание уделено ограничениям примененных методов, таким как неравномерное распределение именованных сущностей в корпусе. Это приводит к тому, что значительная часть упоминаний одной и той же сущности может быть сосредоточена в одном источнике, что отрицательно сказывается на репрезентативности результатов картирования.

Важным фактором, влияющим на качество результатов, является состояние сканированных страниц. Автоматический геопарсинг плохо сохранившихся страниц часто сопровождается ошибками, требующими ручной корректировки для уменьшения числа неправильно идентифицированных или пропущенных географических имен.

Также исследование выявило ограничение используемой библиотеки для NER. В данном случае была применена библиотека Spacy, которая оказалась не способной обрабатывать корпуса объемом более одного миллиона символов. Это ограничение подчеркивает целесообразность использования Spacy для анализа относительно небольших корпусов. Кроме того, следует учитывать, что большинство библиотек для NER полагаются на существующие базы данных географических объектов, и при несовершенстве этих баз результаты могут быть неполными.

Нельзя утверждать, что полученные результаты полностью отражают объективную реальность, так как художественные тексты не всегда точно воспроизводят актуальные события или географические данные.

Таким образом, данное исследование представляет собой первую попытку выявления именованных сущностей и их картирования в корпусе детской литературы Минусинской общественной библиотеки конца XIX – начала XX века, что вносит значительный вклад в развитие геопарсинга и картирования для русского языка.

Объектом настоящего исследования является выявление именованных сущностей в корпусе детской литературы конца XIX – начала XX века и их картирование. Предметом же является распределение частотности географических названий в книгах Минусинской общественной библиотеки конца XIX - начала XX веков.

Научная новизна исследования состоит в первую очередь в том, что впервые было проведено выявление и картирование географических именованных сучщностей в корпусе рускоязычной детской литературы XIX - XX веков. Также подробно исследованы ограничения существующих методов геопарсинга применительно к дореформенным русским текстам. Впервые сформирован географический анализ восприятия мира через художественную литературу Минусинской общественной библиотеки, что может способствовать дальнейшему развитию гуманитарных цифровых исследований.

Обзор связанных работ

В условиях цифровизации и значительного увеличения объема оцифрованных документов геопарсинг стал пользоваться возрастающей популярностью среди исследователей гуманитарных дисциплин. Геопарсинг включает два последовательных этапа: распознавание топонимов и их картирование [1]. Процесс распознавания топонимов часто рассматривается как подзадача распознавания именованных сущностей (Named Entity Recognition, NER), или точнее, как задача классификации именованных объектов (NERC)[2]. Картирование предполагает привязку распознанных топонимов к

соответствующим географическим координатам и их визуализацию на карте.

Исторически методы распознавания именованных сущностей основывались на использовании правил, выведенных из лингвистических характеристик текста, и специализированных словарей [2]. Однако такие подходы требовали значительных трудозатрат на ручную настройку и не могли эффективно адаптироваться к новым областям или языкам. С развитием методов машинного обучения и, в частности, глубокого обучения, подходы к геопарсингу претерпели значительные изменения. Модели глубокого обучения, включая рекуррентные нейронные сети (RNN), сети долготочной кратковременной памяти (LSTM) и трансформеры, продемонстрировали заметное улучшение в распознавании именованных сущностей [3].

В зарубежной научной литературе на сегодняшний день уже сформировалась большая база знаний, связанных с геопарсингом. Проведено большое количество исследований, которые используют различные подходы. Условно их можно разделить на три крупных направления:

1) Описание исключительно алгоритма геопарсинга для выбранного корпуса и интерпретация результатов с гуманитарной точки зрения [5, 6, 7].

2) Сравнительный анализ эффективности нескольких геопарсеров на одном или нескольких корпусах [4, 8, 9].

3) Поиски решений по улучшению результатов геопарсинга [8, 10, 11].

Сравнительный анализ геопарсеров

В случае, когда проводится сравнительный анализ нескольких геопарсеров, один корпус анализируется несколькими методами выявления именованных сущностей, результаты анализируются и выделяется лучший геопарсер. В исследовании, проведенном учеными из США были проанализированы следующие геопарсеры: Spacy, NeuroTPR, Edinburgh Geoparser и CamCoder[4]. Для этого использовались такие англоязычные корпуса как: LGL, GeoVirus и WikToR. Исходя из результатов исследования, приводятся довольно распространенные для геопарсинга проблемы. В первую очередь, геопарсинг смещен в сторону более развитых регионов мира с большими лингвистическими корпусами, что в свою очередь отражается на репрезентативности результатов исследования. NER не совершенен и потому, что картирование может привести к сбою из-за топонимической двусмысленности [8]. В этом случае речь идет об одинаковых географических наименованиях, для правильного картирования которых, необходимо устранять географическую неопределенность с помощь использования дополнительного контекстного топонима. В большинстве случаев, для этого выбираются столицы и важные города.

В рамках исследования по распознаванию географических наименований в корпусе VIII века, переведенного на английский язык с армянского, авторами из Австрии приводится следующее распределение методов геопарсинга по качеству результатов. Лучше всего справился с задачей метод Flair, чуть хуже – TagMe, третье место занимает Spacy и на последнем месте оказался NLTK. Стоить отметить, что TagMe имеет лучшие показатели по нахождению устаревших названий из-за работы с корпусом Википедии [9].

В случае с поиском решений по улучшению результатов геопарсинга, исследователи часто создают собственные геопарсеры, в основу которых входят разного вида и

масштаба нейронные сети. Нейросети улучшают точность и эффективность алгоритмов NER, применяя методы глубокого обучения, случайного леса и наивного байесовского классификатора. Эти методы определяют принадлежность слов или словосочетаний к определённым категориям, таким как имена людей, географические названия, организации и даты. Чтобы подготовить нейросеть к работе, создается обучающий корпус, на котором она будет обучена, после чего проверяется ее работоспособность на тестовых корпусах. Ограничения в этом случае весьма тривиальны. Собственная модель долго обучается, что не позволяет проводить исследования быстро. Также велика роль обучающей выборки, чем она больше и качественнее, тем лучше результат работы геопарсера [8].

Имеют место попытки создания собственных методов для выявления именных сущностей в корпусах, специально созданных для проведения, определенного исследований. В этом случае показательна работа [10], в которой авторы используют собственные разработки на каждом этапе исследования (предварительная обработка, распознавание текста и последующая обработка). На этапе предварительной обработки оцифрованные страницы бинаризируются, то есть превращаются в черно-белые, при этом удаляются нежелательные искажения на страницах. После завершения этапа предварительной обработки на этапе распознавания берутся предварительно обработанные страницы, и выполняется их распознавание. По окончании этого процесса, на этапе постобработки, проверяется качество обработанных данных, и принимается решение о том, пригодна ли книга для дальнейшей работы или ее необходимо обработать заново. Безусловно, процесс обработки занимает достаточно продолжительное время (около трех часов на 500 страниц), однако это время можно значительно сократить путем увеличения количества серверов или повышения их производительности. Стоит отметить и тот факт, что для семантического анализа необходима точность распознавания не менее 90%.

Весьма интересны работы по созданию нейросетей для работы с вложенными именованными сущностями. Примером таких объектов является «Верховный суд Флориды», так как содержит две перекрывающиеся сущности «Верховный суд Флориды» и «Флорида». Исследователи из Чехии в своей работе предлагают две нейросетевые архитектуры для распознавания вложенных именованных объектов и анализируют их работоспособность на четырех корпусах вложенных объектов: английские ACE-2004, ACE-2005, GENIA и чешский CNEC [11].

В первой модели объединяются вложенные несколько меток объекта в одну мультиметку, которая затем прогнозируется с помощью стандартной LSTM-CRF модели. В этом случае под меткой понимается класс именованной сущности, например, географическое наименование. Это важно, так как во вложенных объектах могут присутствовать и объекты, относящиеся к другим классам.

Во второй модели вложенные объекты кодируются в последовательности, и затем задачу можно рассматривать как задачу от последовательности к последовательности (seq2seq), в которой входной последовательностью являются токены (формы), а выходной последовательностью являются метки. Декодер предсказывает метки для каждого токена, пока не дойдет до специальной метки: "" (конец слова), после чего декодер переходит к следующему токену [11].

Авторы приходят к выводу, что LSTM-CRF моделирование мультиметок лучше подходит для предположительно менее вложенных и плоских корпусов, в то время как архитектура от последовательности к последовательности фиксирует более сложные

взаимосвязи между вложенными и сложно именованными объектами.

Источник данных исследования

Наше исследование не направлено на создание принципиально новых методов геопарсинга, мы сконцентрировались на использовании уже готовых методов и интерпретации результатов с технической точки зрения. В исследовании был проведен географический анализ литературного корпуса. В центре внимания находился корпус раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков. В исследование было включено 121 произведение, написанное между 1719 и 1905 годами. Стоит отметить, что не было найдено ни одного научного труда, связанного с географическим распределением в корпусах собраний книг из библиотек Сибири. В работе будет показано, как формировалось пространственное представление о мире у жителя провинциального сибирского города. Предполагается, что это происходило с помощью постепенного введения географических названий, представленных в художественных произведениях, написанных для юного читателя. Цель исследования – получить географическое распределение частоты упоминаний локаций в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков.

Материалы и методы

Приведение к машиночитаемому виду

Первым этапом проведения географического анализа распределения локаций в исследуемом корпусе является преобразование текстов книжных изданий из детской коллекции Минусинской общественной библиотеки конца XIX – начала XX века в машиночитаемый формат. Этот процесс включает оцифровку страниц книг (фотографий или сканов) и преобразование дореформенной орфографии в современную. Для конвертации PDF-документов в текстовый формат, пригодный для компьютерного анализа, был использован редактор ABBYY FineReader 15. Преобразование дореформенной орфографии в современную было автоматизировано с использованием кода на языке Python и библиотеки prereform2modern. В результате было получено 119 текстовых файлов: 32 на английском (тексты изданий в переводе не удалось найти ни в одном агрегаторе) и 87 на русском языках. После завершения этого этапа корпус стал готовым к извлечению географических названий и их последующему картированию.

Важно отметить, что автоматическое преобразование текстов в машиночитаемый формат не исключает возможности ошибок. Эти ошибки могут быть обусловлены состоянием исходных книг, многие из которых, ввиду возраста, сохранились не в лучшем качестве. Это может привести к неправильному распознаванию слов или их полному пропуску. Чтобы минимизировать влияние этого фактора на результаты исследования, проблемные страницы были распознаны вручную. В случаях, когда даже ручное распознавание было невозможно, такие страницы исключались из анализа.

Геопарсинг

Следующим этапом исследования является проведение геопарсинга, который включает три ключевых шага: извлечение географических названий из корпуса, проверку и корректировку результатов вручную, а также объединение данных в единый корпус.

Для автоматического извлечения географических именованных сущностей из текстов был разработан код на языке Python с использованием библиотеки Spacy. Однако при

проверке результатов выяснилось, что алгоритм выделял также сущности, не являющиеся географическими объектами. В связи с этим все результаты были тщательно проверены вручную, и лишние, нерелевантные сущности были удалены.

Для обеспечения корректной работы на этапе картирования было необходимо объединить результаты геопарсинга для каждого издания в один общий файл и подсчитать частоту упоминаний конкретных географических объектов во всех книгах корпуса. Эта задача также была реализована с помощью Python. После завершения данного этапа геопарсинга корпус был готов к дальнейшему анализу и картированию.

Промежуточные результаты

В общей сложности в списке насчитывается 668 наименований городов и 97 названий стран. Для визуализации результатов были построены круговые диаграммы распределения для первых десяти локаций по количеству употреблений стран и городов (Рис. 2,3)

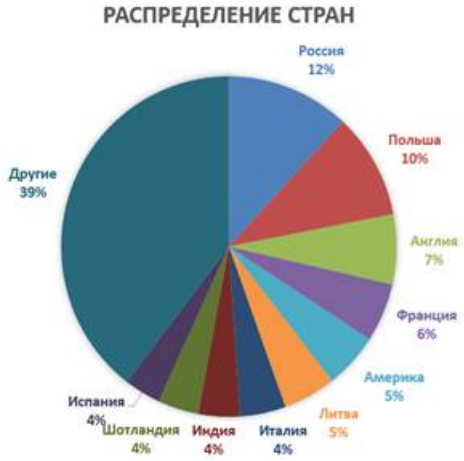

Рисунок 2 – Диаграмма распределения стран, представленных в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков

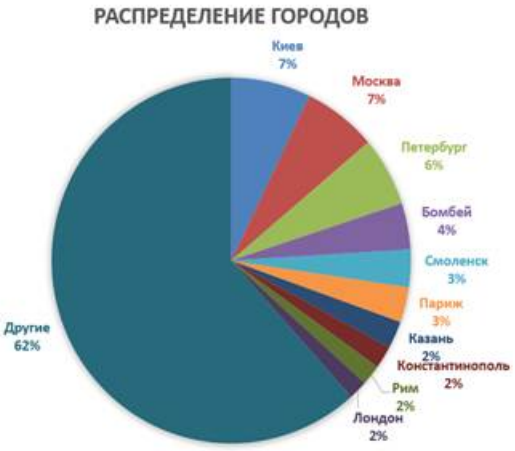

Рисунок 3 – Диаграмма распределения городов, представленных в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков

Первые десять стран по употреблению в текстах занимают 61% от общего употребления (рис. 2). Не удивительно, что на первом месте оказалась Россия. Примечательно, что на

втором месте оказалась Польша.

Данные отображенные на Рисунке 3 свидетельствуют о том, что употребление городов разнообразнее стран, очевидно из-за большего количества городов. Первая десятка занимает 38% от общего употребления, а на первом месте внезапно оказался Киев, вместо предполагаемых Москвы или Санкт-Петербурга. Также весьма популярным стал индийский город Бомбей (4 место).

Картирование

После того как геопарсинг корпуса был завершен и данные были откорректированы, можно приступить к выполнению географического анализа распределения локаций. В рамках нашего исследования были созданы карты распределения упоминаний стран и городов. Для построения карт использовался функционал Microsoft Excel.

При визуализации распределения стран использовался метод градиентной заливки: страны с наибольшим числом упоминаний выделялись более насыщенным цветом, а страны с минимальным количеством упоминаний – менее выраженным оттенком. Страны, которые не упоминались в текстах, оставались окрашенными в нейтральный серый цвет.

Для картирования выявленных городов использовалась другая методика – точечная тепловая карта, где каждое упоминание города отображалось в виде теплового пятна. Такая карта позволяет визуализировать концентрацию упоминаний городов по всему анализируемому корпусу.

Результаты и интерпретация

В рамках исследования были созданы карты распределения упоминаний стран и городов. На карте распределения стран (рис. 4) использована градиентная заливка: страны с наибольшим числом упоминаний выделяются наиболее яркими оттенками, в то время как страны, которые не были упомянуты в анализируемом корпусе, отмечены серым цветом. Среди наиболее часто упоминаемых стран выделяются Россия, Польша и Англия.

Рисунок 4 представляет особый интерес, поскольку позволяет определить, какие страны не упоминались в исследуемом корпусе. Среди них – Казахстан, Таджикистан, Узбекистан, Аргентина, Пакистан, Индонезия и некоторые африканские страны. Однако стоит отметить, что Казахстан (Казахское ханство), Таджикистан и Узбекистан, хотя и отображены на карте серым цветом, входили в состав Российской империи в конце XIX века, что объясняет их отсутствие как отдельных упоминаемых государств.

Экспансия Российской империи в Среднюю Азию привела к упразднению ханской власти в регионе. В частности, «Устав о Сибирских киргизах» 1822 года способствовал присоединению большей части Казахского ханства к России. Кокандское ханство, на территории которого располагались современные Узбекистан, Таджикистан, Кыргызстан и южный Казахстан, было аннексировано Российской империей в 1876 году.

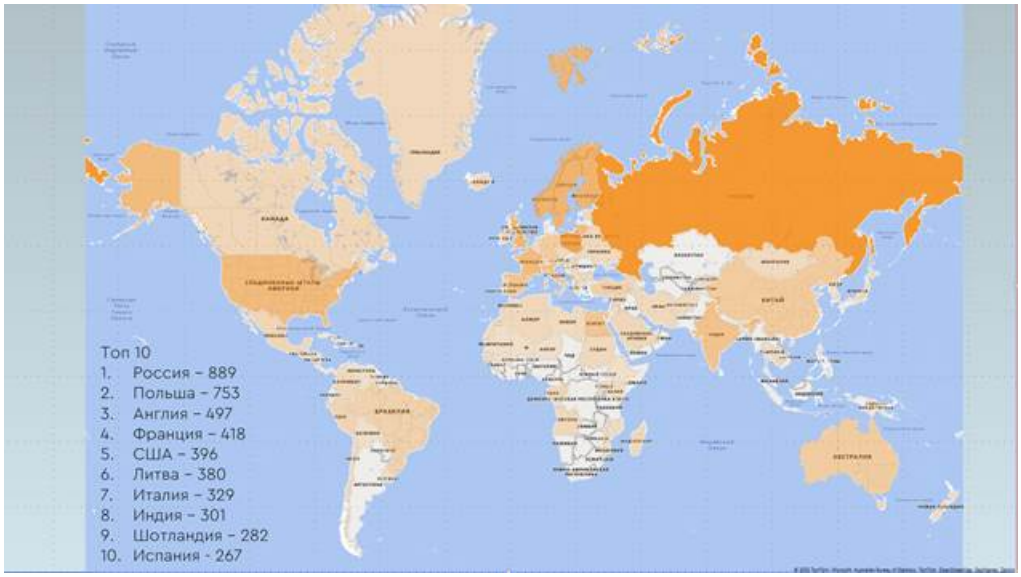

Рисунок 4 – Географическое распределение стран, представленных в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков

На рисунке 5 представлено частотное распределение городов, упомянутых в исследуемом корпусе. Среди наиболее часто упоминаемых оказались Киев, Москва и Санкт-Петербург. Примечательно, что на четвертом месте находится индийский город Мумбаи.

Анализ также показывает, что города, расположенные на территории Европы, упоминаются в корпусе значительно чаще по сравнению с городами Африки, Южной Америки и Австралии. На территории Северной Америки преобладают города Соединенных Штатов, в то время как города Канады полностью отсутствуют, а из городов Мексики представлен лишь Мехико.

В целом, распределение упоминаемых городов соответствует ожидаемым результатам: чаще всего это столицы стран или крупные города, обладающие значительным региональным значением.

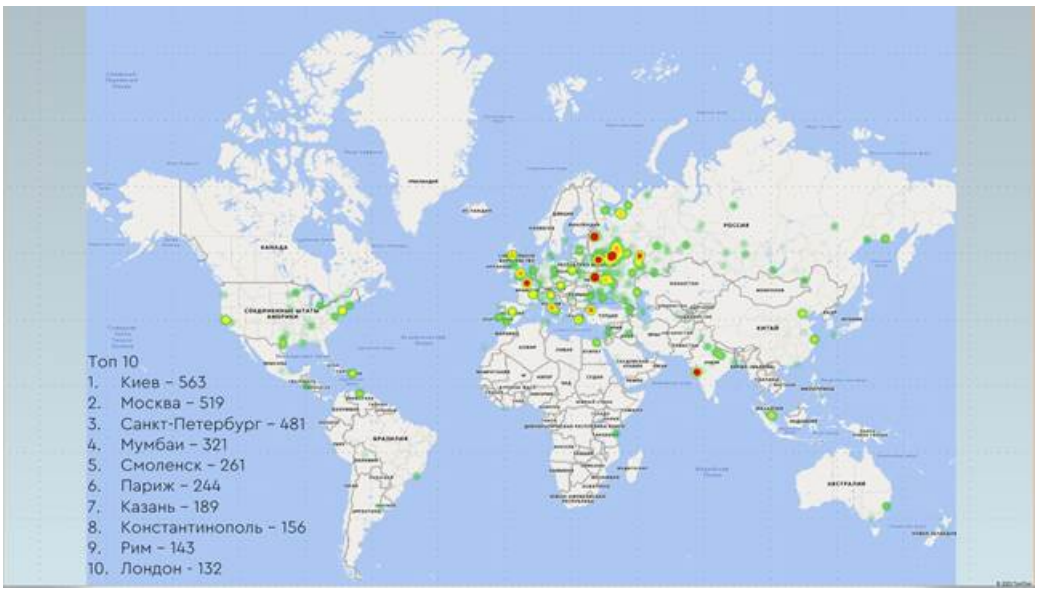

Рисунок 5 – Географическое распределение городов, представленных в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков

Ограничения исследования

В процессе исследования географического распределения стран и городов, представленных в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX — начала XX веков, были выявлены определенные ограничения, связанные с неравномерностью распределения именованных сущностей.

Для проверки этой гипотезы была составлена таблица (Таблица 1), в которой приведены наиболее часто упоминаемые страны и города в отдельных изданиях. Эта таблица помогает выявить возможные аномалии в распределении упоминаний и оценить, насколько равномерно представлены географические объекты в различных произведениях корпуса. В частности, было обнаружено, что некоторые именованные сущности могут упоминаться значительно чаще в одном произведении, чем в остальных текстах корпуса.

Таблица 1 – Неравномерное распределение именованных сущностей в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков

Книга	Наименование	Кол-во упоминаний	Всего упоминаний	Процент от общего
Знаменитые исследователи и путешественники	Испания	147	267	55,06
Сборник статей, выбранных из произведений русской литературы	Санкт-Петербург	186	481	38,67
Сборник статей, выбранных из произведений русской литературы	Россия	82	889	9,22
Сборник статей, выбранных из произведений русской литературы	Москва	173	519	33,33
Как я отыскал Ливингстона	Мумбаи	89	321	27,73
Иллюстрированная история России в разговорах для детей	Киев	100	563	17,76
Иллюстрированная история России в разговорах для детей	Москва	217	519	41,81
Катакомбы	Рим	86	143	60,14
Квентин Дорвард	Франция	181	418	43,30
Книга для первоначального чтения	Санкт-Петербург	71	481	14,76
Куль Хлеба	Россия	56	889	6,30

Треволнения одного китайца в Китае	Китай	48	193	24,87
Рассказы из путешествий по Африке	Мумбаи	49	321	15,26
История России в рассказах	Киев	61	563	10,83
Фрегат Паллада	Англия	80	497	16,10
Паровой дом	Индия	100	301	33,22

Исходя из данных Таблицы 1 можно сделать выводы о неравномерном употреблении некоторых географических наименований в корпусе. Более половины от общего числа употреблений Испании приходится на книгу «Знаменитые исследователи и путешественники» Жюля Верна 1873 года издания. Аналогичная ситуация с Римом, который употребляется в произведении «Катакомбы» Евгении Тур (издание 1866 г.) 86 раз из 143, что составляет 60% от общего употребления. Однако наиболее неравномерно распределено употребление упоминания города Москва, около 75% приходится на два книжных издания («Сборник статей, выбранных из произведений русской литературы», В.А. Яковлев, 1874 г; «Иллюстрированная история России в разговорах для детей», автор неизвестен, 1863г). Примечательный в исследовании, Мумбаи распределен по двум изданиям («Как я отыскал Ливингстона», Г.М. Стэнли, 1873 г; «Рассказы из путешествий по Африке» М.Б. Чистяков, 1897 г.) на 45% от общего употребления. Подобные выводы можно сделать и по другим городам, представленным в Таблице 1.

Еще одно важное ограничение связано с качеством сканированных страниц. Не все книги сохранились в удовлетворительном состоянии, что потребовало значительных усилий для приведения текстов в машиночитаемый формат. В некоторых случаях приходилось вручную корректировать отдельные слова, а в более сложных случаях – расшифровывать целые страницы. Лишь в редких случаях, когда восстановить страницу не представлялось возможным, даже с помощью ручной обработки, такие страницы исключались из анализа. Это могло оказать незначительное влияние на конечные результаты исследования.

Техническим ограничением исследования стало использование библиотеки Spacy для распознавания именованных сущностей. Spacy не поддерживает работу с текстами, объем которых превышает один миллион символов, что требовало разделения больших текстов на части, а после обработки – их повторного объединения. Таким образом, для более эффективного использования Spacy целесообразно применять её для анализа меньших по объему корпусов.

Выявленные ограничения не помешали достижению целей исследования, а скорее предоставили возможность для дальнейших размышлений и анализа. Безусловно, невозможно утверждать, что полученные результаты полностью отражают объективную реальность, так как художественные произведения не могут в точности воспроизвести реальные события и представления. Например, нельзя с уверенностью сказать, что провинциальный житель конца XIX – начала XX века воспринимал мир именно так, как это отражено на созданных в ходе исследования картах. Велика вероятность, что информация о городах и странах передавалась устно, в том числе через рассказы путешественников, торговцев, переселенцев и ссыльных, что могло оказать влияние на восприятие географической картины мира того времени.

Заключение

Настоящее исследование является одним из первых, посвященных выявлению именованных сущностей в корпусе раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков и их картированию. В ходе работы были проанализированы тексты Минусинской общественной библиотеки с использованием современных методов геопарсинга. Исследование выявило ряд особенностей, таких как неравномерное распределение географических наименований в корпусе, а также ограничения, связанные с качеством оцифрованных текстов и техническими аспектами автоматического выявления сущностей.

Сравнивая полученные результаты с другими исследованиями в области геопарсинга, можно отметить, что несмотря на общий успех в применении методов распознавания именованных сущностей, ограничения, выявленные в ходе работы, схожи с проблемами, упомянутыми в других научных трудах. Например, как и в исследованиях [4], NER показал смещение в сторону более известных и часто упоминаемых локаций, что подтверждается неравномерностью распределения топонимов в нашем корпусе. Также отмеченная проблема топонимической двусмысленности [8] нашла свое отражение в нашем исследовании, где некоторые топонимы могли иметь несколько значений или быть нераспознанными системой из-за отсутствия дополнительного контекста.

Выявленные ограничения по объему данных библиотеки Spacy показали, что для более крупных корпусов требуется использование более мощных инструментов, таких как Flair, TagMe или нейросетевые решения, предлагаемые [9]. Важно отметить, что наш корпус отличался языковой спецификой, что требует дополнительных усилий для адаптации существующих инструментов NER к дореформенным текстам, в отличие от англоязычных или других современных корпусов, анализируемых в предыдущих исследованиях [11].

Таким образом, наше исследование вносит значительный вклад в область геопарсинга и NER для русскоязычных текстов XIX – XX веков. Несмотря на выявленные трудности, такие как качество исходных материалов и технические ограничения, исследование открывает новые перспективы для развития методов NLP для корпусов на русском языке. Дальнейшая работа в этом направлении может включать анализ всех книжных изданий из Минусинской библиотеки, сравнение нескольких корпусов: собрание книг из библиотек Сибири и собрание библиотек центральных регионов, анализ личного общения с помощью сохранившихся писем и отображение границ стран с использованием исторических карт.

Библиография

1. Ли Дж., Сан А., Хан Дж., Ли К. Обзор глубокого обучения для распознавания именованных сущностей // IEEE Transactions on Knowledge and Data Engineering. 2020. С. 122-127.
2. Надео Д., Секин С. Обзор распознавания и классификации именованных сущностей // Международный журнал по компьютерной лингвистике и приложениям. 2007. С. 3-26.
3. Ламп Г., Баллестерос М., Субраманиан С., Каваками К., Дайер К. Нейронные архитектуры для распознавания именованных сущностей // Материалы конференции Североамериканского отделения Ассоциации компьютерной лингвистики: технологии обработки естественного языка. 2016. С. 260-270.
4. Лю З., Янович К., Цай Л., Чжу Р., Май Г., Ши М. Геопарсинг: решение или предвзятость? Оценка географических предвзятостей в геопарсинге // AGILE: серия "ГИС-наука". 2022. С. 13.

5. Бургмайстер М. Измерение городских изменений в текстах о путешествиях на примере города Грац в XIX веке // magazen. 2022. Т. 3, № 1. С. 61-90.
6. Эванс Э., Уилкенс М. Нация, этническая принадлежность и география британской художественной литературы, 1880–1940 гг. // Журнал культурного анализа. 2018. С. 48.
7. Смайл Р., Грегори И., Тейлор Дж. Качественная география в цифровых текстах: представление исторических пространственных идентичностей в Озерном крае // Международный журнал гуманитарных и художественных вычислений. 2019. С. 28-38.
8. Файз Дж., Монкла Л., Мартинс Б. Глубокое обучение для распознавания топонимов: геокодирование на основе пар топонимов // Международный журнал ISPRS по геоинформации. 2021. С. 16.
9. Тамбускио М., Эндрюс Т.Л. Геолокация и распознавание именованных сущностей в древних текстах: тематическое исследование армянской истории Гевунда // Конференция по исследованиям в области гуманитарных наук. 17-19 ноября 2021 года. Амстердам, 2021. С. 136-148.
10. Санджакомо А., Хогенбирк Х., Танасеску Р., Караисл А., Уайт Н. Чтение в тумане: высококачественное оптическое распознавание символов на основе свободно доступных оцифрованных книг раннего Нового времени // Digital Scholarship in the Humanities. 2022. Т. 37, № 4. С. 1197-1209. DOI: 10.1093/llc/fqac014 EDN: IWWDWY
11. Стракова Й., Страка М., Хайич Й. Нейронные архитектуры для вложенного NER с помощью линеаризации // Материалы 57-й ежегодной конференции Ассоциации компьютерной лингвистики. 2019. С. 6.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Представленная статья на тему «Мир глазами образованного человека г. Минусинска конца XIX - начала XX веков: распределение частотности географических названий в книгах Минусинской общественной библиотеки» соответствует тематике журнала «Историческая информатика» и посвящена актуальному исследованию.
В водной части статьи авторы уделяют внимание исследованию, направленному на выявление именованных сущностей и их картирование в корпусе детской литературы из собрания книг Минусинской общественной библиотеки конца XIX – начала XX века. В работе представлен алгоритм геопарсинга и картирования географических именованных сущностей в текстах, написанных дореформенной орфографией. Особое внимание уделено ограничениям примененных методов, таким как неравномерное распределение именованных сущностей в корпусе. По мнению авторов это приводит к тому, что значительная часть упоминаний одной и той же сущности может быть сосредоточена в одном источнике, что отрицательно сказывается на репрезентативности результатов картирования.
Авторами самостоятельно проведен географический анализ литературного корпуса раздела детской литературы собрания книг Минусинской общественной библиотеки конца XIX – начала XX веков. В исследование было включено 121 произведение, написанное между 1719 и 1905 годами. В работе показано, как формировалось пространственное представление о мире у жителя провинциального сибирского города.
В качестве материалов и методов авторами выбраны: приведение к машиночитаемому виду, геопарсинг, картирование.
Практическая значимость четко обоснована и заключается в значительном вкладе в области геопарсинга и NER для русскоязычных текстов XIX – XX веков. Исследование

открывает новые перспективы для развития методов NLP для корпусов на русском языке. Авторами указана целесообразность перспектив дальнейшего исследования, которая заключается в анализе всех книжных изданий из Минусинской библиотеки, сравнение нескольких корпусов: собрание книг из библиотек Сибири и собрание библиотек центральных регионов, анализ личного общения с помощью сохранившихся писем и отображение границ стран с использованием исторических карт.
Статья по объему соответствует рекомендуемому объему от 12 000 знаков. Стиль и язык изложения является достаточно доступным для широкого круга читателей. Авторами статьи проведен широкий аналитический обзор отечественной и зарубежной литературы. Статья достаточно структурирована - в наличии введение, заключение, внутреннее членение основной части (обзор связанных работ, результаты и интерпретация).

К недостаткам можно отнести следующие моменты: не сформулирован объект и предмет исследования; отсутствует научная новизна.
Рекомендуется сформулировать объект и предмет исследования; обозначить научную новизну.
Статья «Мир глазами образованного человека г. Минусинска конца XIX - начала XX веков: распределение частотности географических названий в книгах Минусинской общественной библиотеки» требует доработки по указанным выше замечаниям. После внесения поправок рекомендуется к повторному рассмотрению редакцией рецензируемого научного журнала.

Результаты процедуры повторного рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Предметом исследования является распределение частотности географических названий в книгах Минусинской общественной библиотеки конца XIX – начала XX веков. Авторы сосредоточились на анализе детской литературы, чтобы выявить, как географические объекты представлены в текстах и как это могло формировать мировоззрение читателей того времени. Исследование также затрагивает технические аспекты обработки текстов, включая геопарсинг и картирование. Несмотря на ограничения исследовательской базы – 121 произведение – , статья имеет значительную ценность как историческое исследование, вносит вклад в развитие Digital Humanities.
Использованы методы NLP (Natural Language Processing), такие как геопарсинг и картирование. Это позволяет автоматизировать обработку больших объемов данных, что особенно важно для работы с архивами и библиотеками; визуализировать результаты, исследовать тексты, которые ранее не анализировались с точки зрения их географического содержания. Таким образом, статья демонстрирует, как современные технологии могут быть применены для изучения исторических источников, поднимает важные методологические вопросы, связанные с работой с историческими текстами – проблемы оцифровки и распознавания текстов плохой сохранности, ограничения современных инструментов NLP (например, Spacy) для работы с историческими корпусами, неравномерность распределения данных. Эти проблемы актуальны для всех, кто работает с историческими текстами в цифровую эпоху. Предложенная в статье технология исследования включает несколько этапов: оцифровка и преобразование текстов, геопарсинг, картирование. Результаты геопарсинга визуализировались с помощью Microsoft Excel, где использовались градиентные заливки для стран и тепловые карты для городов. Методология в целом соответствует современным подходам к NLP,

однако использование Spacy для анализа дореформенных текстов вызывает вопросы, так как эта библиотека не оптимизирована для работы с историческими текстами.

Исследование актуально в контексте развития цифровых гуманитарных наук и NLP. Оно вносит вклад в изучение русскоязычных текстов, которые до сих пор остаются недостаточно исследованными по сравнению с англоязычными корпусами. Кроме того, работа имеет историческую ценность, так как позволяет понять, как формировалось географическое восприятие мира у жителей провинциального города в конце XIX – начале XX веков.

Научная новизна выражена тем, что анализ географических названий в корпусе дореволюционной детской литературы был проведен впервые. Выявлены ограничения методов геопарсинга применительно к дореформенным текстам. Следует отметить, что новизна несколько ограничена использованием уже существующих инструментов (Spacy) без их значительной модификации для работы с историческими текстами.

Стиль текста научный, с использованием терминологии NLP и исторического анализа. Структура работы логична, состоит из необходимых элементов: введение, обзор литературы, описание методов, результаты и их интерпретация, заключение. Вместе с тем, в вводной части выбор корпуса произведений, как и библиотеки Минусинска не обоснован и выглядит случайным. Исследовательская база, использованная в работе, недостаточно репрезентативна для широких выводов о «мире образованного человека» конца XIX – начала XX веков. Текст перегружен техническими деталями, в целом статья сосредоточена на внедрении современных методов NLP, основное внимание уделено техническим аспектам исследования – геопарсинг, обработка текстов, визуализация данных – в ущерб исторической интерпретации. Такие вопросы, как провинциальная культура, история детской литературы или проблемы исторической памяти остались нераскрытыми.

Библиография включает современные работы по NLP, геопарсингу и историческому анализу текстов, публикации за последние 5 лет составляют 50% списка. Большинство ссылок относятся к англоязычным исследованиям, что подчеркивает недостаток работ по русскоязычным корпусам. Не хватает ссылок на исследования, посвященные именно дореформенным текстам, что могло бы усилить аргументацию авторов. В исследование не вовлечены работы по истории провинциальной культуры, детской литературы, истории Минусинска.

Авторы признают ограничения своего исследования, такие как неравномерное распределение географических названий в корпусе, проблемы с качеством оцифрованных текстов и ограничения библиотеки Spacy. Они также отмечают, что результаты могут быть не полностью репрезентативными из-за специфики художественных текстов. Однако авторы не уделяют достаточного внимания возможным альтернативным подходам, таким как использование более мощных инструментов (например, Flair или нейросетевых моделей), что могло бы улучшить качество анализа.

Выводы исследования логичны и соответствуют поставленным задачам. Авторы подчеркивают, что их работа вносит вклад в развитие методов NLP для русскоязычных текстов и открывает новые перспективы для дальнейших исследований. Однако выводы могли бы быть более конкретными, например, с указанием на то, какие именно аспекты географического восприятия мира были выявлены. В выводах исследования «Мир глазами образованного человека г. Минусинска конца XIX – начала XX веков» отражен через анализ частотности упоминаний географических названий в книгах Минусинской общественной библиотеки. Вот что фактически написано в выводах: исследование является одним из первых, посвященных анализу географических названий в корпусе детской литературы Минусинской библиотеки; выявлены технические ограничения, такие как неравномерность распределения топонимов и проблемы с качеством оцифрованных

текстов; подчеркивается, что результаты не могут полностью отражать объективную реальность, так как художественные тексты не всегда точно воспроизводят географические данные; указывается на необходимость дальнейших исследований, включая анализ других корпусов и сравнение с библиотеками центральных регионов. Отмечается, что художественные тексты не всегда точно отражают реальную географическую картину мира, что «мир глазами образованного человека Минусинска» был европоцентричным с акцентом на Европу и Россию, ограниченным с минимальным представлением о других частях света, фрагментарным с неравномерным распределением упоминаний географических объектов. Эти выводы логически вытекают из анализа частотности упоминаний стран и городов.

Интерес читательской аудитории будет зависеть от её специализации. Для исследователей в области NLP и цифровых гуманитарных наук работа представляет значительный интерес, особенно в части анализа дореформенных текстов. Для более широкой аудитории, включая историков и культурологов, исследование также может быть полезным, хотя может и не оправдать ожидания.

Несмотря на то, что многие вопросы остались нераскрытыми, статья научно значима – демонстрирует возможность применения методов NLP для анализа исторических текстов. Важно и то, что статья привлекает внимание к провинциальной культуре – упоминание Минусинска и его общественной библиотеки может вдохновить других исследователей на изучение региональной истории.