Историческая информатика
Правильная ссылка на статью:
Дебенова З.А., Цыпилова С.С., Цыренова Н.Д. Памятники на монгольской письменности: опыт создания параллельного корпуса // Историческая информатика. 2025. № 2. DOI: 10.7256/2585-7797.2025.2.73930 EDN: MMDRBC URL: https://nbpublish.com/library_read_article.php?id=73930

Памятники на монгольской письменности: опыт создания параллельного корпуса

Дебенова Зинаида Анциферовна

ORCID: 0000-0002-8824-6624

младший научный сотрудник; Институт монголоведения; буддологии и тибетологии СО РАН

670047, Россия, респ. Бурятия, г. Улан-Удэ, Октябрьский р-н, ул. Сахьяновой, д. 6

debenova@gmail.com

Цыпилова Снежана Сергеевна

ORCID: 0000-0001-9578-5493

кандидат исторических наук

младший научный сотрудник Федерального государственного бюджетного учреждения науки Институт монголоведения, буддологии и тибетологии СО РАН

670047, Россия, респ. Бурятия, г. Улан-Удэ, Октябрьский р-н, ул. Сахьяновой, д. 6

ai_qing@mail.ru

Цыренова Номинь Дондоковна

младший научный сотрудник, Институт монголоведения, буддологии и тибетологии СО РАН 670047, Россия, республика Бурятия, г. Улан-Удэ, ул. Сахьяновой, 6

670047, Россия, республика Бурятия, г. Улан-Удэ, ул. Сахьяновой, 6

nomin_n@mail.ru

Статья из рубрики "Искусственный интеллект и наука о данных"

DOI:

10.7256/2585-7797.2025.2.73930

EDN:

MMDRBC

Дата направления статьи в редакцию:

02-04-2025

Дата публикации:

21-04-2025

Аннотация: Данная статья освещает результаты работы по созданию параллельного корпуса бурятских источников на монгольской письменности. Проект осуществляется при поддержке РНФ на источниковой базе Центра восточных рукописей и ксилографов ИМБТ СО РАН. Предметом исследования является процесс создания базы данных корпуса, специфика составления, в частности выборки материалов. На данный момент в разрабатываемый корпус вошли следующие документы из архивных фондов ЦВРК ИМБТ СО РАН: тексты исторического содержания – «Краткий очерк истории хори-монгольских бурят», «Об истории местности Зугалай»; официальный документ «Протокол всебурятского собрания в Чите 1917 года»; этнографическое сочинение «Повествование о Самдан нойоне», медицинское сочинение «Заметки тибетского врача Дондуба Мункуева»; произведение буддийской дидактической литературы «Субхашита» в переводе Галсан-Жимбы Тугулдурова. Для анализа рукописных, печатных и ксилографических текстов на монгольской письменности применялись общенаучные и источниковедческие методы. Рассмотрены процессы отбора материалов, их транслитерации и перевода, а также содержательные (тематика, лексика) и технические аспекты (опечатки, пагинация, числительные). Параллельный русскоязычный вариант создаётся научной группой. Авторы подчеркивают значимость создания параллельного корпуса как ресурса для дальнейших исследований в области бурятского языкознания, переводоведения и культурологии, а также его роль в популяризации старомонгольской письменности среди широкой общественности, а также сохранении нематериального наследия Байкальского региона. Корпус представляет собой уникальную базу данных для дальнейших исследований в различных областях науки и т. д. Рассмотренные тексты послужат базой для развития алгоритмов машинного перевода, а проводимая на данном этапе работа поможет будущим разработчикам создавать более эффективные алгоритмы. Перспективным представляется создание специализированной базы данных, открытой не только для исследователей, но и для представителей образовательной сферы, профессиональных переводчиков, а также всех лиц, проявляющих научный или культурный интерес к письменному наследию.

Ключевые слова:

старомонгольская письменность, параллельный корпус, письменные источники, Бурятия, ЦВРК, Байкальский регион, нематериальное наследие, машинный перевод, оцифровка, текстовый корпус

Исследование выполнено в рамках гранта РНФ «Параллельный корпус бурятских письменных источников на монгольской письменности: на пути к машинному переводу», проект № 24-28-00942.

Введение. Байкальский регион представляет собой уникальное этнокультурное пространство России, где начиная с XVII века происходит интенсивное взаимодействие различных мировых культур и религиозных традиций. Исследование письменных памятников, созданных на территории данного региона, является важным направлением изучения его истории и культуры, способствующего более глубокому пониманию происходящих процессов. Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН (далее ЦВРК ИМБТ СО РАН) храни

значительное количество письменных памятников, включающих разнообразные тексты на тибетском и старописьменном монгольском языках. Коллекция на монгольском языке насчитывает более 6,5 тысяч единиц хранения, и большая часть имеющихся документов до сих пор не исследована.

Причины сложности изучения данной коллекции обусловлены рядом причин. Вертикальная монгольская письменность *монгол бэшэг* получила широкое распространение среди бурят в начале XVIII в. и активно использовалась для записи бурятского языка в делопроизводстве Степных дум и дацанов, при написании исторических, литературных, религиозных сочинений, а также в личном общении. В 1920-х годах, в ходе советской кампании по латинизации письменностей народов СССР, была разработана бурят-монгольская письменность на основе латиницы. Она была официально введена в 1931 году, но уже в 1939 году алфавит был сменен на кириллический. Результатом всех этих изменений стала утрата традиции обучения *монгол бэшэг*, что в свою очередь усложняет изучение сохранившихся памятников.

С 2024 года с финансовой поддержкой Российского научного фонда исследовательским коллективом ИМБТ СО РАН реализуется проект под названием «Параллельный корпус бурятских письменных источников на монгольской письменности: на пути к машинному переводу». Целью проекта является создание базы для дальнейшей разработки параллельного корпуса письменных памятников на старомонгольской письменности и их переводов на русский язык. В задачи проекта входит оцифровка и аннотирование текстов, выработка рекомендаций и правил подготовки датасетов для машинного обучения в области OCR и машинного перевода. В данной статье предоставлен обзор проделанной работы: поэтапное описание принципов выборки материалов для включения в корпус, специфики транслитерации рукописных источников, ксилографических и печатных текстов, особенности переводов описанных источников.

Материалы и методы. Одним из приоритетных направлений работы ЦВРК ИМБТ СО РАН сегодня является прикладное применение современных технологий, таких как оцифровка, машинное обучение и искусственный интеллект. Сочетание традиционных методов источниковедческого и лингвистического анализа с современными технологиями обработки данных способствует развитию междисциплинарных дисциплин, одной из которых является корпусная лингвистика.

Создание корпусов – структурированных наборов текстов в электронном формате – открывает новые перспективы и возможности для исследования языковых процессов и явлений, значительно упрощает анализ большого объема материалов и обеспечивает доступ к ним большему числу людей. Развитие корпусной лингвистики в свою очередь стимулирует проведение прочих междисциплинарных исследований, а также способствует созданию дополнительных инструментов для сохранения языков. Существуют разные виды текстовых корпусов, каждый из которых создается для определенных целей. На сегодняшний день отечественными исследователями уже реализуются такие проекты как корпус бурятского языка [1], диахронический корпус бурятского языка [2; 3] и диалектный корпус бурятского языка [2; 4].

В корпус литературного бурятского языка включены произведения художественной, учебно-научной и общественно-публицистической литературы, написанные на кириллице в период с начала XX века до настоящего времени. Всего в нем насчитывается около 2,5 миллионов словоупотреблений. Данный корпус активно используется в научных исследованиях по бурятскому языку, включая лексикографические, морфологические, синтаксические и семантические анализы. Его данные служат основой для подготовки

научных статей и диссертаций, а также базой для создания учебных материалов и методических пособий по бурятскому языку.

Диахронический корпус включает в себя тексты летописей и других исторических документов на старомонгольском письме в латинизированной транслитерации и предназначен для исследования исторического развития бурятского языка и его литературных традиций. Всего в нем содержится около 82 тысяч словоупотреблений. Эти данные позволяют исследовать язык в его историческом контексте, выявлять изменения, происходившие на протяжении долгого периода времени. Данные корпуса служат основой для сравнительно-исторических исследований, а сам проект способствует сохранению культурного наследия бурятского народа.

Проект диалектного корпуса включает в себя звуковые записи, которые отражают фонетические и грамматические особенности различных бурятских говоров. Его основная цель заключается в обеспечении открытого доступа к структурированным и аннотированным звуковым данным бурятских диалектов. Корпус организован в виде геоинформационной системы, что позволяет учитывать территориальное распределение диалектов и их связь с родоплеменными объединениями. Также к подобным проектам в рамках монгольской языковой группы относятся корпус калмыцкого языка [5] и монгольский корпус [6; 7].

Авторами данного исследования проводится работа по созданию первого параллельного корпуса источников на классическом монгольском письме. Результатом проекта станет электронная база данных, содержащая тексты на двух языках (бурятском и русском), которые являются переводами друг друга. Подобные корпусы используются для изучения и анализа переводов, межъязыковых соответствий и различий, а также для создания и улучшения систем машинного перевода.

Создание параллельного корпуса включает в себя следующие этапы: выбор исходных текстов из фондов ЦВРК ИМБТ СО РАН, оцифровка, транслитерация и перевод на русский язык, выравнивание и лингвистическая разметка.

Результаты.

Тематическая выборка текстов и их содержание. Работа с письменными памятниками на бурятском языке предполагает повышенное внимание к ряду языковых, культурных и исторических факторов. Во-первых, в бурятских памятниках присутствует большое количество заимствований из других языков, и в особенности из русского. Во-вторых, отсутствует стандартизированная орфография, что обуславливает широкую ее вариативность, особенно в плане терминов, выходящих за пределы общеупотребительной лексики, а также для географических названий, личных имен, заимствованных слов и т. д. В-третьих, в грамматике языка бурятских памятников присутствует значительные отличия от классических форм старописьменного монгольского языка. Л. Б. Бадмаева предполагает следующее: «Полагаем, что многочисленные отступления от правил старописьменного монгольского языка, допущенные в бурятских летописях, могут быть объяснены, с одной стороны, давлением стихии разговорного бурятского языка, с другой – подобного рода отклонения не были кодифицированы, хотя и признается, что в Бурятии формировался письменный диалект (извод) старомонгольского языка» [8, с. 45].

Все эти факторы определили ряд требований к содержанию источниковой базы параллельного корпуса, так как используемые материалы будут также применяться при

создании обучающих датасетов для машинного перевода. Одним из главных факторов выборки была репрезентативность создаваемого корпуса: в его составе должны быть представлены тексты, принадлежащие к разным жанрам и тематикам, составленные разными авторами. Таким образом, рабочей группой был выбран ряд текстов, которые в свою очередь можно разделить на несколько жанрово-тематических групп, далее мы опишем некоторые из них. К первой группе относятся тексты исторического содержания, в которую вошли следующие документы:

Краткий очерк истории хори-монгольских бурят (*qori mongγol buriyad ulus-un quriyangγui teüke orusibai*) [9]. Автором данного текста является бурятский врач и краевед Даши Бубеев. Письменный памятник посвящен истории этнической группы хори-бурят, и представляет собой рукописную летопись, составленную на основе авторских воспоминаний. В источнике описываются события из истории хоринских бурят, последовательно представленные в пяти разделах. Точная дата составления неизвестна, предположительно летопись была написана после 1936 г. [10, с. 213] Исторические рамки данного сочинения охватывают период с легендарного Хоридай-мэргэна по 1930 год – указанный период включает в себя события, связанные с присоединением Бурятии к Российской империи. Данный текст является примером летописной традиции хоринцев – одной из самых многочисленных этнических групп бурят, а также отличается богатым художественным языком и большим количеством исторических сведений. В то же время сочинение Д. Бубеева демонстрирует более позднюю стилистику, характерную для сочинений XX в., выраженную в краткости предложений. Все эти факторы делают данный источник важным текстом в контексте создания параллельного корпуса.

Об истории местности Зугалаи˘ (ǰuγalai nutuγ tuqai) [11]. Еще одна рукопись авторства Д. Бубеева. Памятник представляет собой историко-географическое сочинение на старописьменном монгольском языке и является примером дореволюционной бурятской краеведческой традиции историописания. Сочинение было написано в 1964 г. и содержит краткие сведения об истории и культуре местности Зугалай, полученные автором от местных жителей, за исторический период с 1600 г. по 1964 г. В тексте источника встречаются топонимы, связанные с данной местностью, имена выдающихся личностей, названия дацанов, а также заимствования из русского языка. В «Истории Зугалая» прослеживается особый авторский стиль, а также при его написании Даши Бубеев привлекал собственные полевые материалы, что делает данное произведение ценным источником не только письменной, но отчасти и устной традиции.

К группе официальных документов относится следующий источник:

Протокол всебурятского собрания, собранного в городе Чите 23 апреля 1917 года (1917-duγar on-u apreli sarayin 23 edür=e čita qota-dur čiγulaγsan bügüde buriyad-un čiγulγan-u protoqol bičig) [12]. Данный памятник является важным историческим документом, в котором отражены первые шаги бурятского национального движения в начале XX века. Сам съезд, также известный как Первый общебурятский съезд бурят-монголов Восточной Сибири, проходил с 23 по 25 апреля 1917 года в Чите. На нём были приняты решения, направленные на национальное возрождение бурятского народа, такие как национализация школы, введение делопроизводства на бурятском языке, организация издательской деятельности и т. д. Текст данного протокола набран на печатной машинке, что делает корпус более репрезентативным в контексте приведенной графики. Помимо самого протокола к документу прилагается телеграмма, отправленная Первому министр-председателю Временного правительства России Львову, и часть проекта Бурят-монгольской автономии, составленного М. Н. Богдановым. Данный протокол полезен для

изучения установленных речевых стандартов и канцеляризмов периода первой половины XX в., привносит в корпус значительное количество профессиональной терминологии и архаизмов. В тексте также имеется большое количество имен собственных и названий административных единиц.

Сочинения этнографического характера:

Повествование о Самдан нойоне, привычка извлекать выгоду за одалживание денег и различных предметов (samdan noyan-u teüke, mal-un toγ=a luγ=a, busud-tu mönggün ba eldeb ǰögeri önggüged qončin ködülmürilegülegsen jangsil anu) [13]. Рукописное сочинение посвящено человеку по имени Самдан нойон, владевшему большим хозяйством и отличавшимся жадностью: «История о том, как Самдан нойон умножал свое состояние», делал деньги или «каким образом эксплуатировал народ» [14, с. 149]. Автором данного сочинения является некий Ц. Самданов – предположительно, родственника самого Самдан нойона [14, с. 146]. В тексте источника содержатся уникальные этнографические сведения, информация о культурных, социальных и бытовых аспектах жизни забайкальских бурят: описываются реалии жизни скотоводов, их взаимоотношения с властью и религией в лице буддийской церкви, приводится много информации об экономических взаимоотношениях. Привлечение этнографических сочинений при создании параллельного корпуса позволит понять контекст употребления определенных слов и выражений, что значительно повысит качество моделей машинного перевода.

Медицинские сочинения:

Заметки тибетского врача Дондуба Мункуева о лечении различных заболеваний лекарственными средствами тибетской медицины (qoγusun t□bed-ün emči neremǰi-tü dondub m□ngke-yin / t□bed emnel-ün yosun-iyar / emčilelge-yin tuqai-du ebedčin eldeb ǰüil·üd-tü em □gtügsen temdeg-tü bičig dangča bolai)[15]. Данная рукопись также была написана Дондубом Ендоновым (Дондуб Мункуев – еще один вариант имени автора). В отличие от предыдущих сочинений данная рукопись полностью посвящена медицинской тематике и содержит в себе сведения из традиционной тибетской медицины, ее основным положениям и различным практикам. Сочинение также содержит в себе многочисленные вставки на тибетском языке, особенно относительно названий отдельных болезней и лекарственных прописей. Установлено, что рукопись была закончена Д. Ендоновым 10 февраля 1935 г. и представляет собой изложение его личного опыта врачебной практики. Традиционная медицина в Бурятии имеет глубокие исторические корни, связанные с развитием тибетской медицины и буддизма, поэтому включение подобных медицинских текстов позволяет обогатить создаваемый корпус грамматическими конструкциями и медицинскими терминами, а также использовать его для изучения межъязыковых связей бурятского и тибетского языков.

Буддийская дидактическая литература:

Субхашита - трактат, называемый драгоценная сокровищница полезных изречений (sayin ügetü erdeni-yin sang subhasida kemegdekü sastir orusiba) [16]. Данное произведение является важным образцом индо-тибетской литературы, который стал неотъемлемой частью монгольской и бурятской литературной традиции. Оригинальное сочинение было составлено Сакья-Пандитой Гунга-Джалцаном (1182–1251). Данный трактат содержит 457 четверостиший, в которых обсуждаются моральные и этические аспекты жизни, а также принципы правильного поведения. Текст охватывает темы достойного и недостойного поведения, разумных и неразумных действий, честных и нечестных поступков, а также их

последствия. Каждое четверостишие предлагает практические советы и мудрость, которые направлены на применение в повседневной жизни, в его содержании отражаются интертекстуальные связи с индийской литературной традицией. Субхашита была переведена на множество языков, включая монгольский. Существуют шесть переводов на классическом монгольском письме, среди которых выделяется перевод Галсан-Жимбы Тугулдурова, печатавшийся в Агинском дацане для широкой аудитории ксилографическим способом, экземпляр которого хранится в ЦВРК ИМБТ СО РАН и был включен в данный корпус. Этот источник содержит в себе большое количество терминов, относящихся к философии, религиозным практикам, а также много топонимов и имен собственных, заимствованных из тибетского и санскритского языков.

Стоит отметить, что описанные выше тексты не являются полным списком источников, которые будут использованы при создании корпуса. Научным коллективом планируется привлечение дополнительных источников по истории распространения буддизма, а также прочие исторические сочинения.

Особенности транслитерации и перевода текстов. Следующим после выборки текстов этапом работы является их транслитерация – переложение с вертикального монгольского письма на общепринятую систему на основе латиницы. То, что значительная часть привлеченных источников представляет собой рукописи и обозначенная выше проблема в виде отсутствия стандартизированной орфографии несколько усложняют данный процесс.

Научным коллективом были составлены правила для разметки и разработана особая система символов, которая применяется для обозначения следующих вещей: конец предложения, конец строки, вставки, зачеркнутые фрагменты, тибетские глоссы, межстрочные леммы, грамматические словоформы, все знаки препинания также обозначаются специальными символами. В привлеченных источниках встречаются разные виды числительных: написанные арабскими цифрами и традиционными монгольскими, последние при транслитерации отмечаются особым знаком. Также специальными знаками отмечаются уровни текстов: главы, части, параграфы и т.д.

Следующим этапом создания параллельного корпуса является выравнивание текстов, который подразумевает сопоставление фрагментов оригинала и перевода. Выравнивание может проводиться на уровне предложений или слов, однако это часто связано с трудностями из-за различий в структуре языков. Так как в рамках данного проекта привлекаются тексты, ранее не переведенные на русский язык, данный этап предваряется работой по их переводу.

Необходимо отметить, что материалы параллельного корпуса в рамках следующего этапа данного проекта будут использованы при обучении нейросетевых моделей OCR и машинного перевода. В связи с этим перевод источников выполняется с учетом синтаксиса оригинальных сочинений в ущерб их художественной ценности, а местами и логики русского языка. Так, русский язык допускает свободный порядок слов благодаря падежной системе, но чаще всего используется SVO (субъект-глагол-объект) порядок. В монгольских языках, к которым относится и бурятский, порядок слов чаще всего фиксированный – SOV (субъект-объект-глагол), что влияет на структуру предложений.

Заключение. Создание параллельного корпуса источников на старомонгольской письменности представляет собой важный шаг в области корпусной лингвистики и существенно способствует сохранению культурного и языкового наследия Байкальского региона. Данный проект позволяет не только систематизировать и оцифровать

разнообразные тексты, но и создать уникальную базу данных для дальнейших исследований в области лингвистики, литературоведения, исторической науки и т. д.

Особенно значима роль параллельного корпуса в создании алгоритмов машинного перевода. Для успешного функционирования систем машинного перевода требуется большое количество качественных и верно аннотированных текстов на исходном и целевом языках. Параллельный корпус на старомонгольской письменности обеспечивает такие данные, что в будущем поможет разработчикам машинного перевода создавать более точные и эффективные алгоритмы.

Проекты подобного характера разрабатываются сотрудниками ИМБТ СО РАН. Так, в период с октября по декабрь 2021 года был реализован пилотный проект по созданию датасета для обучения моделей оптического распознавания символов тибетского языка. В рамках проекта была впервые использована технология глубокого обучения для разработки модели распознавания тибетской письменности. Для этого было отсканировано несколько редких тибетских изданий, а основой для работы стало ксилографическое Чонэское издание *Кангьюра* XVIII века. На их основе были подготовлены датасеты. Алгоритм, полученный в результате машинного обучения с помощью этих датасетов, достиг точности 94% при распознавании графем, что в сумме обеспечило около 80% точности декодирования текста [17].

Кроме того, создание и использование параллельного корпуса способствует популяризации старомонгольского письма среди широкой аудитории. В планах создать базу данных, доступную не только для узкого круга специалистов, но и для студентов, преподавателей, переводчиков и всех интересующихся.

Библиография

1. Бурятский корпус [Электронный ресурс]. – Режим доступа: https://buryat.web-corpora.net/ (дата обращения: 16.09.2024).
2. Диахронический корпус бурятского языка [Электронный ресурс]. – Режим доступа: http://annals.imbtarchive.ru/ (дата обращения: 16.09.2024).
3. Ринчинов О.С. Диахронический корпус бурятского языка как цифровой инструмент исторических исследований: подходы, решения, экспериментальные исследования // Историческая информатика. 2020. № 2. С. 26-34. DOI: 10.7256/2585-7797.2020.2.33446 URL: https://nbpublish.com/library_read_article.php?id=33446
4. Ринчинов О. С., Абаева Ю. Д. Геоинформационный веб-ресурс "Диалектный корпус бурятского языка" // Филологические науки. Вопросы теории и практики. 2023. Т. 16, № 1. С. 328-334. DOI: 10.30853/phil20230006. EDN: KTPPAM.
5. Национальный корпус калмыцкого языка [Электронный ресурс]. – Режим доступа: http://kalmcorpora.ru/ (дата обращения: 16.09.2024).
6. Монгольский корпус [Электронный ресурс]. – Режим доступа: http://web-corpora.net/MongolianCorpus/search/index.php?interface_language=ru (дата обращения: 16.09.2024).
7. Бадмаева Л. Д. Монголоязычные корпуса: современное состояние // Вестник Бурятского государственного университета. Филология. 2015. № 10. С. 148-152.
8. Бадмаева Л. Б. Языковое пространство бурятского летописного текста / Л. Б. Бадмаева ; отв. ред. Л. Д. Шагдаров ; Федеральное гос. бюджетное учреждение науки Ин-т монголоведения, буддологии и тибетологии Сибирского отд-ния РАН. Улан-Удэ : Изд-во Бурятского науч. центра СО РАН, 2012. ISBN 978-5-7925-0340-3.
9. Цыренова Н. Д. История хоринских бурят: рукопись Даши Бубеева // Гуманитарный вектор. 2020. Т. 15, № 3. С. 153-160. DOI: 10.21209/1996-7853-2020-15-3-153-160. EDN:

NWQLEX.
10. Цыренова Н. Д., Ван И. Д. Об одном историческом сочинении бурятского летописца Д. Бубеева // Духовное наследие народов Центральной Азии. Улан-Удэ : Изд-во БНЦ СО РАН, 2020. С. 212-214.
11. Цыренова Н. Д. Краевед Даши Бубеев и его рукопись "Об истории местности Зугалай" как источник по истории Агинского округа // Вестник Бурятского научного центра Сибирского отделения Российской академии наук. 2023. № 1(49). С. 97-103. DOI: 10.31554/2222-9175-2023-49-97-103. EDN: UFBUBV.
12. Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук. MII-680.
13. Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук. MI-32.
14. Галданова Г. Р., Дашибалов Б.-Ц. Рукописное наследие Ц. Самданова // Культура Центральной Азии: письменные источники. Вып. 2. Сб. ст. Улан-Удэ: Изд-во БНЦ СО РАН, 1998. С. 145-166.
15. Ванчикова Ц. П., Жабон Ю. Ж., Цыренова Н. Д., Дашиева С. Б. Рукопись бурятского эмчи-ламы Д. Ендонова из монгольской коллекции ЦВРК ИМБТ Сибирского отделения РАН // Вестник архивиста. 2020. № 4. С. 1255-1266. DOI: 10.28995/2073-0101-2020-4-1255-1266. EDN: MNRCML.
16. Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук. MII-269.
17. Базаров Б. В., Ринчинов О. С., Базаров А. А. Цифровая трансформация письменного наследия тибетского буддизма: состояние и перспективы // Oriental Studies. 2022. Vol. 15, No. 4. P. 740-750. DOI: 10.22162/2619-0990-2022-61-4-740-750. EDN: VPYBAW.

Результаты процедуры рецензирования статьи

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

Рецензируемая статья посвящена изучению использования технологий искусственного интеллекта для исследования текстов на старописьменном монгольском языке.
Методология исследования базируется сочетании традиционных методов источниковедческого и лингвистического анализа с современными технологиями обработки данных (таких как оцифровка, машинное обучение и искусственный интеллект), а также применении подходов и методов корпусной лингвистики.
Актуальность работы авторы связывают с наличием множества неисследованных документов на старописьменном монгольском языке, необходимостью разработки моделей машинного обучения для автоматизации их перевода, и возможностью получения новых исторических знаний.
Научная новизна рецензируемого исследования, по мнению рецензента, состоит в поэтапном описании принципов выборки материалов для включения в корпус, специфики транслитерации рукописных источников, ксилографических и печатных текстов, особенностей переводов описанных источников с использованием современных методов машинного обучения и искусственного интеллекта.
В публикации структурно выделены следующие разделы и подразделы: Введение, Материалы и методы, Результаты, Тематическая выборка текстов и их содержание, Особенности транслитерации и перевода текстов, Заключение и Библиография.
В статье освещена деятельность Центра восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН по сохранению письменных

памятников, включающих разнообразные тексты на тибетском и старописьменном монгольском языках; отмечено, большая часть имеющихся документов до сих пор не исследована; показана важность оцифровки и аннотирования текстов, выработки рекомендаций и правил подготовки датасетов для машинного обучения в области оптического распознавания и машинного перевода. Под корпусами в рецензируемой работе понимаются структурированные наборы текстов в электронном формате. В публикации приведены примеры использована технологии глубокого обучения для разработки модели распознавания письменности на различных языках. Отмечена роль параллельного корпуса в создании алгоритмов машинного перевода, поскольку для успешного функционирования систем машинного перевода требуется большое количество качественных и верно аннотированных текстов на исходном и целевом языках. Авторы справедливо полагают, что параллельный корпус на старомонгольской письменности обеспечивает такие данные, и в будущем это поможет разработчикам машинного перевода создавать более точные и эффективные алгоритмы.

Библиографический список включает 17 источников – публикации отечественных ученых по теме статьи на русском языке, а также интернет-ресурсы. На источники в тексте имеются адресные ссылки, подтверждающие наличие апелляции к оппонентам.

Из недостатков публикации следует отметить, что в тексте имеются несогласованные причастные обороты, например, во втором предложении статьи, а также неудачные словосочетания, например, «междисциплинарных дисциплин», «существенно способствует».

В целом же статья отражает результаты проведенного авторами исследования, соответствует направлению журнала «Историческая информатика», содержит элементы научной новизны и практической значимости, может вызвать интерес у читателей, может быть рекомендована к опубликованию после корректировок в соответствии с высказанными замечаниями.