Дмитрий Шандронов – магистр, Военно-инженерный институт радиоэлектроники и связи, Министерства Обороны Республики Казахстан, Алматы, Казахстан; e-mail: shan_dima@mail.ru. ORCID: https://orcid.org/0009-0000-2361-2542.

Information about the authors

Zarina Kutpanova – master of the Department «Computer engineering», Astana IT University, Astana, Kazakhstan; e-mail: zarina.almabekkyzy@gmail.com. ORCID: https://orcid.org/0009-0007-4860-9244.

Dinara Kozhakhmetova* – PhD, Associate Professor of the Department «Automation and Information Technologies», Shakarim University, Semey, Kazakhstan; e-mail: d.kojahmetova@shakarim.kz. ORCID: https://orcid.org/0000-0002-4327-3899.

Kenesary Suleymenov – PhD, Candidate of physical-mathematical sciences of the department «Mathematical and computer modelling». LN Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail:kenessarymath@gmail.com.

Gani Baiseitov – candidate of technical Sciences, Professor R&D Center «Kazakhstan Engineering», Astana, Kazakhstan; e-mail:nesipova@rdke. ORCID: https://orcid.org/0009-0001-0154-0071.

Dmitry Shandronov – master of the Military Engineering Institute of Radio Electronics and Communications, Almaty, Kazakhstan; e-mail: shan_dima@mail.ru. ORCID: https://orcid.org/0009-0000-2361-2542.

Received 05.01.2026
Revised 19.02.2026
Accepted 20.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-14

МРНТИ: 50.13.13

М.Ж. Ергеш[1]*, Б.Ж. Ергеш[1], А.Р. Оразаева[2], А. Талгат[2]
[1]Евразийский национальный университет им. Л.Н. Гумилёва,
010000, Казахстан, Астана, ул. Сатпаева, 2
[2]АО «Казахский университет технологии и бизнеса им. К. Кулажанова»,
010000, Казахстан, Астана, ул. Мухамедханова, 37А
*e-mail: shymkent90@mail.ru

ПОСТРОЕНИЕ И ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ ВОПРОСНО-ОТВЕТНОГО КОРПУСА KAZAKH HISTORYQA

Аннотация: *Цифровизация среднего образования в Казахстане усиливает потребность в специализированных казахскоязычных ресурсах, поддерживающих интеллектуальные вопросно-ответные системы по школьным предметам. Однако существующие QA-датасеты в основном ориентированы на высокоресурсные языки и не охватывают содержание учебных дисциплин, в частности истории Казахстана. В данной работе представлен и подробно проанализирован корпус Kazakh HistoryQA – первый специализированный вопросно-ответный ресурс на казахском языке, созданный на основе шести официальных учебников по истории Казахстана для 5-10 классов. Корпус включает 208 тематических секций, около 180 тысяч токенов и 661 QA-пару с привязкой к точным координатам ответного спана, что обеспечивает совместимость со стандартными пайплайнами обучения моделей машинного чтения, включая библиотеку HuggingFace.*

Проведён многоуровневый анализ корпуса: статистика длины контекстов, вопросов и ответов, распределение материалов по классам и учебникам, типологизация вопросов на фактологические, хронологические, процедурные, причинно-следственные и сравнительные. Дополнительно выполнено морфологическое профилирование текста, построен частотный словарь и выделены предметные термины. Для оценки надежности корпуса как бенчмарка реализованы базовые модели выбора предложений (случайный выбор, TF-IDF, BM25), среди которых BM25 показал лучшие результаты. Представленный корпус создаёт основу для последующих исследований и разработки гибридных онтологически обогащённых QA-систем в сфере образования.

Ключевые слова: *Казахстанская история, OCR, QA, онтология, образовательная NLP, RAG, тезаурус, онтология.*

Введение. Развитие цифровых образовательных ресурсов в системе среднего образования Казахстана требует наличия полноценных казахскоязычных корпусов, ориентированных на типовые школьные предметы и формы учебной деятельности. В области автоматической обработки естественного языка это, прежде всего, корпусные и бенчмарк-ресурсы для задач машинного чтения и вопросно-ответных систем, способных работать с аутентичными текстами школьных учебников. При этом для предметной области истории Казахстана до настоящего времени отсутствовал единый, систематически описанный корпус, который одновременно отражал бы структуру школьного курса 5-10 классов и позволял бы проводить воспроизводимую оценку алгоритмов вопросно-ответного поиска [1].

В данной статье представляется корпус Kazakh HistoryQA, построенный на материале шести официальных учебников «История Казахстана» для 5-10 классов. Корпус включает 208 тематических секций, извлечённых из указанных учебников: 46 секций для 5 класса, 54 секции для 6 класса, 29 – для 7 класса, 28 – для 8 класса, 29 – для 9 класса и 22 – для 10 класса, что в сумме даёт 177 501 нормализованный токен. Для этих текстов сформировано 661 вопросно-ответная пара с сохранением фрагмента-доказательства и точных координат ответного спана в исходном контексте; на их основе построен датасет в формате DatasetDict с обучающей, валидационной и тестовой подвыборками (соответственно 528, 60 и 133 примера).

Типологизация вопросов по поверхностным лингвистическим маркерам выявила доминирование фактологических вопросов (473 из 661, около 72%), а также наличие хронологических (118 вопросов), процедурных (44 вопроса), причинно-следственных (24 вопроса) и единичных сравнительных заданий. Частотный анализ корпуса и CSV-онтологии показал высокую долю предметной лексики, а эвристическое морфологическое профилирование по основным казахским падежным и множественным суффиксам выявило значимую нагрузку генитивных и локативных форм (около 6-7% токенов каждая), что отражает специфику исторического дискурса с акцентом на принадлежностные и пространственно-временные отношения [2].

Методы исследования. Корпус Kazakh HistoryQA построен на основе шести учебников по предмету «История Казахстана» для 5-10 классов, представленных в виде отдельных каталогов *Kazakh_History_5–Kazakh_History_10*. Исходные тексты загружаются из файлов texts_corpus.csv для каждой книги, где для каждой записи указаны идентификаторы учебника (book_id), тематической секции (section_id) и соответствующий фрагмент учебного текста (text). Нормализованная версия текста (text_n) формируется автоматически с учётом удаления служебных символов, унификации апострофов, дефисов и пунктуации.

Дополнительно поверх текстового корпуса формируются два связанных слоя данных. Первый слой – вопросно-ответная разметка, извлекаемая из файлов qa_open.jsonl для каждого учебника и включающая поля question, answer, evidence_quote, book_id и section_id. На основе этих файлов и соответствующих текстовых секций строится итоговый набор из 661 QA-записи с точными координатами ответного спана в контексте, который затем разбивается на обучающую, валидационную и тестовую подвыборки. Второй слой – CSV-онтология, агрегируемая из файлов ontology_triples.csv и содержащая 2318 отфильтрованных определений, привязанных к книгам и секциям. Эти онтологические сведения используются для построения онто-префиксов в контексте и для лингвистического анализа терминологической плотности корпуса [3].

Результаты исследований. Рисунок 1 иллюстрирует структуру текстового корпуса по классам, показывая количество секций, извлечённых из каждого учебника. Наибольшее число секций приходится на 6 класс (54 секции), несколько меньше – на 5 класс (46 секций), тогда как для 7, 8 и 9 классов наблюдается сопоставимый уровень детальности (29, 28 и 29 секций соответственно).

Минимальное количество секций фиксируется для 10 класса (22 секции), что отражает более крупную тематическую сегментацию старшего учебника. Такое распределение важно учитывать при интерпретации статистики корпуса и при проектировании обучающих и тестовых выборок.

Таблица 1 суммирует основные количественные характеристики текстового корпуса по классам: число секций, их суммарный объём в символах и токенах, а также средние и медианные длины секций.

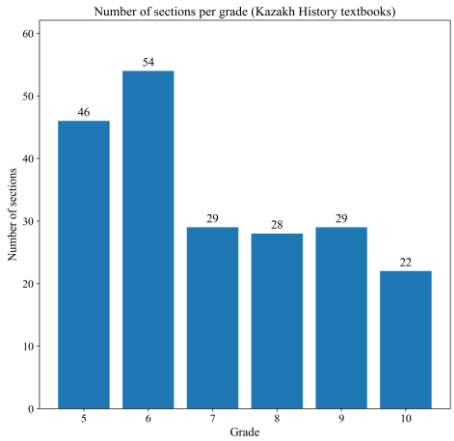

Рисунок 1 – Распределение числа тематических секций по классам в учебниках «История Казахстана» (5-10 классы)

Таблица 1 – Статистики секций корпуса Kazakh HistoryQA по классам

Класс	Число секций	Объём, символов (total_chars)	Объём, токенов (total_tokens)	Среднее число токенов в секции	Медиана токенов в секции
5	46	195 468	24 570	534.13	568.5
6	54	375 271	47 635	882.13	747.0
7	29	157 019	20 240	697.93	705.0
8	28	372 027	46 977	1 677.75	1 507.0
9	29	125 529	15 234	525.31	583.0
10	22	185 232	22 845	1 038.41	1 041.0

Видно, что при сопоставимом числе секций в 7-9 классах существенно различается плотность текста: особенно объёмными являются секции 8 и 10 классов, где среднее число токенов на секцию превышает 1000, тогда как для 5 и 9 классов типичная секция содержит около 500-550 токенов. Эти различия отражают усложнение и укрупнение тематических блоков на старших ступенях школьного курса истории.

Рисунок 2 показывает гистограмму длины секций корпуса Kazakh HistoryQA в токенах.

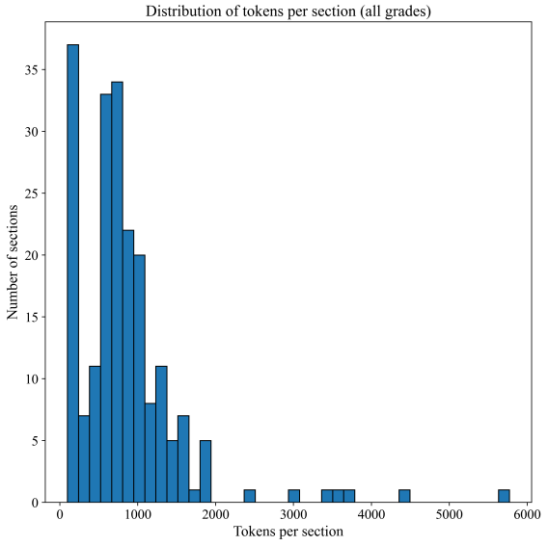

Рисунок 2 – Распределение числа токенов в секции учебника (все классы)

Распределение имеет выраженную правостороннюю асимметрию: основная масса секций сосредоточена в диапазоне до 1000 токенов, тогда как отдельные секции достигают 2000-4000 и более токенов. Наличие такого «длинного хвоста» указывает на сосуществование относительно кратких и существенно более объёмных тематических фрагментов, что важно учитывать при оценке сложности задач машинного чтения и проектировании стратегий разбиения контекста для QA-моделей [4].

Таблица 2 характеризует распределение вопросно-ответных пар по классам и их типичные длины.

Таблица 2 – Статистики вопросно-ответных пар Kazakh HistoryQA по классам

Класс	Число QA-пар	Средняя длина вопроса, токенов	Средняя длина контекста, токенов	Средняя длина ответного спана, токенов
5	106	10.13	558.41	11.79
6	198	10.05	922.63	9.65
7	131	10.53	709.85	12.66
8	13	10.46	935.31	10.62
9	132	11.39	590.31	12.74
10	81	6.86	1 045.89	7.51

Наибольшее количество QA-примеров сформировано для 6 класса (198 пар), далее следуют 7 и 9 классы, тогда как для 8 класса корпус содержит лишь 13 аннотированных вопросов. Средняя длина вопроса по большинству классов стабильно находится около 10-11 токенов, за исключением 10 класса, где вопросы заметно короче (≈7 токенов) при максимальной средней длине контекста (более 1000 токенов). Ответные спаны в среднем составляют 10-13 токенов, причём наиболее развернутые ответы наблюдаются в 7 и 9 классах, что указывает на более сложные формулировки и богаче структурированный учебный материал на этих ступенях.

Рисунок 3 демонстрирует, как аннотированные вопросно-ответные пары Kazakh HistoryQA распределены по учебникам 5-10 классов.

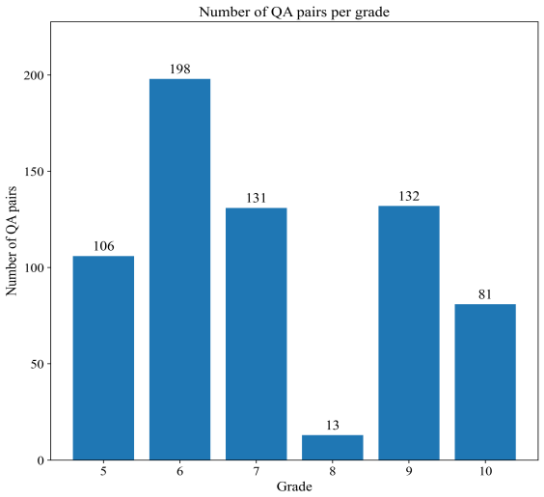

Рисунок 3 – Распределение числа вопросно-ответных пар по классам

Наибольшее количество QA-пар сформировано для 6 класса (198 примеров), далее следуют 9 (132) и 7 классы (131), тогда как для 5 и 10 классов объём разметки несколько ниже (106 и 81 пара соответственно). Минимальное число вопросов наблюдается в 8 классе (13 примеров), что отражает меньшую степень охвата данного учебника при аннотировании. Такое распределение важно учитывать при сравнении результатов моделей между классами и при построении стратифицированных обучающих и тестовых выборок.

Рисунок 4 иллюстрирует распределение длины вопросов по числу токенов для всего корпуса.

Видно, что большинство вопросов сосредоточено в интервале 8-12 токенов, формируя почти симметричное колоколоподобное распределение вокруг среднего значения около 10 токенов (табл. 3). Короткие (менее 5 токенов) и, напротив, удлинённые формулировки (более 15 токенов) встречаются существенно реже, что свидетельствует о доминировании компактных, но информативных вопросов, типичных для школьных учебных заданий [5].

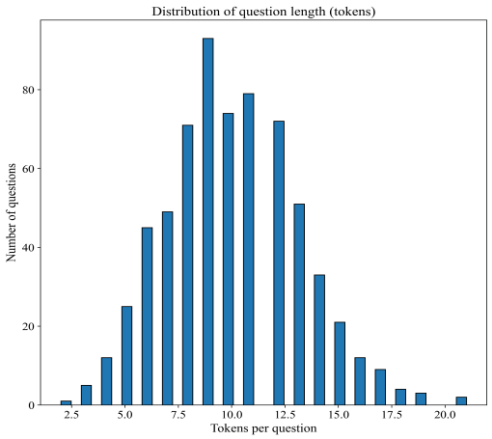

Рисунок 4 – Распределение длины вопросов в корпусе Kazakh HistoryQA (в токенах)

Таблица 3 – Распределение типов вопросов по классам в корпусе Kazakh HistoryQA

Класс	Причинно-следственные (causal)	Хронологические (chronological)	Сравнительные (comparative)	Фактологические (factoid)	Процедурные (procedural)
5	7	8	0	81	10
6	6	38	2	144	8
7	3	41	0	75	12
8	0	2	0	11	0
9	4	25	0	90	13
10	4	4	0	72	1

Хронологические вопросы наиболее представлены в 6 и 7 классах, что коррелирует с усиленным вниманием к периодизации исторических событий в соответствующих разделах учебников. Причинно-следственные и процедурные вопросы встречаются реже и распределены более неравномерно, а сравнительные задания зафиксированы только в 6 классе, что подчёркивает общую ориентацию учебного материала на воспроизведение фактов и временных связей, а не на сопоставительный анализ [7].

Таблица 4 представляет тридцать наиболее частотных токенов нормализованного корпуса объёмом 177 501 токен.

Таблица 4 – Топ-30 наиболее частотных токенов корпуса Kazakh HistoryQA (нормализованный текст, N = 177 501 токен)

№	Токен	Частота	Относительная частота	№	Токен	Частота	Относительная частота
1	мен	2 321	0.013076	16	өз	492	0.002772
2	және	1 992	0.011222	17	ол	488	0.002749
3	-	1 967	0.011082	18	ұлы	474	0.002670
4	болды	1 104	0.006220	19	олар	411	0.002315
5	қазақ	1 097	0.006180	20	б	379	0.002135
6	бұл	807	0.004546	21	әскери	377	0.002124
7	қазақстан	587	0.003307	22	•	361	0.002034
8	қандай	585	0.003296	23	де	360	0.002028
9	деп	582	0.003279	24	олардың	357	0.002011
10	да	570	0.003211	25	орталық	342	0.001927
11	туралы	565	0.003183	26	болып	320	0.001803
12	ғасырдың	564	0.003177	27	болған	310	0.001746
13	оның	548	0.003087	28	қарсы	305	0.001718
14	жылы	541	0.003048	29	пен	298	0.001679
15	хан	514	0.002896	30	үшін	293	0.001651

Как видно, наряду с функциональными элементами («мен», «және», «да», «деп») в верхней части списка присутствуют ключевые предметные лексемы, связанные с историческим дискурсом («қазақ», «қазақстан», «ғасырдың», «хан», «әскери», «орталық»). Высокая относительная частота таких терминов отражает тематическую направленность

корпуса на историю Казахстана и подтверждает его пригодность для исследования доменно-специфичных моделей обработки казахскоязычных текстов [8-12].

Обсуждение научных результатов. Полученные результаты позволяют дать обоснованную оценку как самому корпусу Kazakh HistoryQA, так и его пригодности в качестве бенчмарка для казахскоязычных вопросно-ответных моделей по истории Казахстана.

Во-первых, с точки зрения охвата учебного материала корпус достаточно репрезентативен для школьного курса 5-10 классов. Структура текста по классам (рис. 1, табл. 1) показывает, что наибольшее число секций приходится на учебники 5 и 6 классов (46 и 54 секции соответственно), тогда как в 10 классе сегментация более крупная (22 секции). Гистограмма длины секций в токенах (рис. 2) демонстрирует выраженную правостороннюю асимметрию: большая часть секций имеет объём до 1000 токенов, но присутствует «длинный хвост» из существенно более протяжённых фрагментов, особенно характерных для 8 и 10 классов. Таким образом, корпус сочетает как компактные, так и очень длинные контексты, что делает задачу машинного чтения ближе к реальному сценарию работы с учебником [13-15].

Во-вторых, распределение и статистика вопросно-ответных пар подтверждают, что корпус пригоден для обучения и тестирования моделей на разных ступенях школьного образования. По абсолютному числу QA-пар доминируют 6, 7 и 9 классы (рис. 3, табл. 2), тогда как 8 класс представлен минимальным количеством аннотаций, что следует рассматривать как ограничение текущей версии датасета. Разбиение на обучающую, валидационную и тестовую выборки сохранено стратифицированным по классам, что обеспечивает сопоставимость результатов экспериментов между уровнями. Средняя длина вопросов во всех классах составляет около 10-11 токенов (за исключением 10 класса, где вопросы короче), а средняя длина ответных спанов – порядка 10-13 токенов при средней длине контекста 500-1000 токенов (табл. 2). Распределения длины вопросов и ответов (рис. 4) подтверждают, что основная масса формулировок является компактной, но не однословной: ответы чаще всего представляют собой короткие фразы или части предложений, а не единичные токены. Для моделей это означает необходимость учитывать локальные контекстные зависимости при выделении спана, а не опираться только на точное совпадение отдельных слов.

В-третьих, анализ типов вопросов выявил важные особенности когнитивной нагрузки, закладываемой в заданиях по истории Казахстана. Распределение по классам (табл. 3) показывают доминирование фактологических вопросов (473 из 661), значительную долю хронологических (118) и более скромное представительство процедурных и причинно-следственных заданий; сравнительные вопросы встречаются единично. Это указывает на то, что в текущей версии корпуса преобладают запросы на извлечение и локализацию конкретных фактов, дат и имён, тогда как задания, требующие многошагового объяснения, сопоставления или обобщения, представлены существенно меньше. С одной стороны, такое распределение соответствует традиционной структуре школьных учебников, с другой – задаёт направление для будущего расширения корпуса за счёт более сложных типов вопросов, ориентированных на развитие исторического мышления учащихся.

В-четвёртых, частотный и морфологический анализ подтверждают высокую доменно-специфическую насыщенность корпуса. Топ-30 токенов (табл. 4) демонстрируют сочетание служебных слов и ключевых историко-политических терминов: «қазақ», «қазақстан», «ғасырдың», «хан», «әскери», «орталық» и др. Эвристический подсчёт падежных и множественных суффиксов показывает значительную долю форм с родительным и местным падежами (около 6-7% и 5-6% токенов соответственно), что отражает характерный для исторического дискурса фокус на отношениях принадлежности и локализации. Эти наблюдения важны для проектирования морфологически осведомлённых моделей обработки казахского языка и для построения онтологических представлений, которые учитывают типичные синтаксико-морфологические паттерны в учебных текстах.

Особое место в обсуждении занимает онтологический слой корпуса. CSV-онтология, включающая 2318 отфильтрованных определений, используется для построения онто-префиксов, автоматически добавляемых к контексту. Хотя в рамках данной статьи онтологическая информация не задействована непосредственно в бейзлайновых моделях, её наличие создаёт предпосылки для последующих экспериментов с гибридными архитектурами, совмещающими нейросетевой вопросно-ответный модуль с символическими знаниями предметной области .

В-пятых, результаты базовых экспериментов демонстрируют, что Kazakh HistoryQA не является тривиальным набором для поверхностных методов ранжирования. Случайный выбор предложения даёт крайне низкое качество (EM 1,5%, F1 около 8%), что фактически соответствует угадыванию. Учёт простого лексического пересечения («Max lexical overlap») уже обеспечивает заметный прирост (F1 ≈ 46 %), а переход к TF–IDF и BM25 даёт дальнейшее улучшение до 48% и 51% F1 соответственно. Модель BM25 достигает EM 23,3 %, Recall@10 около 0,77 и MRR 0,53 при латентности порядка нескольких миллисекунд, что делает её разумным нижним ориентиром для последующих нейросетевых систем. Важно отметить, что даже лучший классический бейзлайн по-прежнему заметно уступает желаемым значениям для прикладного учебного сценария, что подчёркивает наличие значительного потенциала для улучшения качества за счёт более сложных архитектур – в частности, трансформерных моделей с онтологическим обогащением контекста [16].

Вместе взятые, эти наблюдения позволяют ответить на основной исследовательский вопрос: корпус Kazakh HistoryQA обладает достаточно богатой структурой, лингвистическими характеристиками и разнообразием вопросов, чтобы служить реалистичным бенчмарком для разработки и оценки казахскоязычных вопросно-ответных моделей в предметной области истории Казахстана. В то же время выявленные ограничения – относительно небольшой общий объём QA-пар, заметный дисбаланс между классами (особенно малое представительство 8 класса) и преобладание фактологических заданий – задают контуры дальнейшего развития ресурса.

Заключение. В работе представлен и подробно проанализирован корпус Kazakh HistoryQA – специализированный казахскоязычный вопросно-ответный ресурс по истории Казахстана для 5-10 классов. На основе шести официальных учебников сформировано 208 тематических секций суммарным объёмом 177 501 токен (табл. 1), для которых вручную аннотировано 661 вопросно-ответная пара с фрагментами-доказательствами и точными координатами ответных спанов.

Проведён лингвистический и статистический анализ корпуса, включающий:

- описательные статистики по длине секций, предложений, контекстов, вопросов и ответных спанов (рис. 2, 4; табл. 1-2);
- типологизацию вопросов на фактологические, хронологические, причинно-следственные, процедурные и сравнительные с оценкой распределения по классам (рис. 6, табл. 3);
- частотный анализ лексики и онтологически отобранных токенов (табл. 4);
- эвристическое морфологическое профилирование по ключевым падежным и множественным суффиксам казахского языка.

Эти результаты показывают, что корпус сочетает разнообразные по объёму и сложности тексты, насыщен доменно-специфической лексикой и отражает типичные структурные особенности исторического учебного дискурса.

Для оценки пригодности Kazakh HistoryQA в качестве бенчмарка реализован набор простых базовых моделей выбора ответного предложения: случайный выбор, ранжирование по максимальному лексическому пересечению, TF-IDF и BM25. На тестовой выборке лучший результат показала модель BM25 (EM 23,3 %, F1 51,3 %, MRR 0,53, Recall@10 0,77; табл. 8, рис. 9), существенно превосходя более простые подходы, но оставляя значительный резерв для нейросетевых и гибридных решений. Тем самым корпус не только задаёт реалистическую нижнюю планку качества, но и позволяет количественно оценивать вклад более сложных методов.

Список литературы

1. Vartiainen T. Engaging with bad (meta) data in historical corpus linguistics / T. Vartiainen, T. Säily // Challenges in Corpus Linguistics. – John Benjamins Publishing Company, 2024. – P. 9-34.
2. Durrell M. ‘Representativeness’, ‘Bad Data’, and legitimate expectations. What can an electronic historical corpus tell us that we didn’t actually know already (and how)? / M. Durrell // Historical Corpora. Challenges and Perspectives. – Narr, 2024. – P. 13-33.
3. Historical insights at scale: A corpus-wide machine learning analysis of early modern astronomic tables / O. Eberle et al // Science Advances. – 2024. – Vol. 10, № 43. – P. eadj1719.
4. Václav C. When Data Meet Tools: Using the Monitor Corpus for the Analysis of Language Development / C. Václav, S. Martin, P. Klára // Jazykovedný Časopis. – 2025. – Vol. 76, № 1. – P. 157-166.

5. Gorozhanov A.I. Natural Language Processing and Fiction Text: Basis for Corpus Research / A.I. Gorozhanov, I.A. Guseinova, D.V. Stepanova // Vestnik Rossiiskogo universiteta druzhby narodov. Seriya: Teoriya yazyka. Semiotika. Semantika. – 2024. – Vol. 15, № 1. – P. 195-210.
6. TIGQA: An Expert Annotated Question Answering Dataset in Tigrinya / H. Teklehaymanot et al // arXiv preprint arXiv:2404. – 17194. – 2024.
7. Nyarko G.S. Leaderships' Role in Quality Assurance Moderating Curriculum Implementation: Perspectives and Preferences of Headteachers and Teachers in Basic Schools in Ghana / G.S. Nyarko // Education Quarterly Reviews. – 2025. – Vol. 8, № 3.
8. Veitsman Y. Recent advancements and challenges of Turkic Central Asian language processing / Y. Veitsman, M. Hartmann // Proceedings of the First Workshop on Language Models for Low-Resource Languages. – 2025. – P. 309-324.
9. Biyik H. Turkish Delights: A Dataset on Turkish Euphemisms / H. Biyik, P. Lee, A. Feldman // Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024). – 2024. – P. 71-80.
10. Kardeş-NLU: Transfer to Low-Resource Languages with the Help of a High-Resource Cousin – A Benchmark and Evaluation for Turkic Languages / L.K. Senel et al // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. – 2024. – Vol. 1. – P. 1672-1688.
11. Tutar K. Turkish Question-Answer Dataset Evaluated with Deep Learning / K. Tutar, O.T. Yildiz // 2024 9th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2024. – P. 101-105.
12. Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation / M.A. Bayram et al // arXiv preprint arXiv:2501.00593. – 2024.
13. Kaya Y.B. Effect of Tokenization Granularity for Turkish Large Language Models / Y.B. Kaya, A.C. Tantung // Intelligent Systems with Applications. – 2024. – Vol. 21. – Art. 200335.
14. Development and Evaluation of a Small Kazakh Language Corpus to Improve the Efficiency of Multilingual NLP Systems in Low-Resource Environments / A. Tleubayeva et al // 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST). – IEEE, 2025. – P. 1-6.
15. Aitim A. A Comparison of Kazakh Language Processing Models for Improving Semantic Search Results / A. Aitim, R. Satybaldiyeva // Eastern-European Journal of Enterprise Technologies. – 2025. – Vol. 133, № 2.
16. Machine Learning Methods for Kazakh Morphology: A Comprehensive Overview / I. Akhmetov et al // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). – IEEE, 2024. – P. 1880-1884.

Благодарность

Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP22787194).

М.Ж. Ергеш[1]*, Б.Ж. Ергеш[1], А.Р.Оразаева[2], А.Талғат[2]
[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010000, Қазақстан, Астана, ул. Сатпаева, 2
[2]Қ. Құлажанов атындағы Қазақ технология және бизнес университеті,
010000, Қазақстан, Астана, ул. Мухамедханова, 37А
*e-mail: shymkent90@mail.ru

ҚАЗАҚ ТАРИХЫ СҰРАҚ-ЖАУАП КОРПУСЫНЫҢ ҚҰРЫЛЫСЫ ЖӘНЕ ЛИНГВИСТИКАЛЫҚ ТАЛДАУЫ

Қазақстандағы орта білім беруді цифрландыру мектеп пәндері бойынша интеллектуалды сұрақ-жауап жүйелерін қолдайтын мамандандырылған қазақ тіліндегі ресурстарға деген қажеттілікті арттырып отыр. Дегенмен, қолданыстағы сапаны қамтамасыз ету деректер жиынтығы негізінен жоғары ресурсты тілдерге бағытталған және академиялық пәндердің, әсіресе Қазақстан тарихының мазмұнын қамтымайды. Бұл мақалада 5-10 сыныптарға арналған Қазақстан тарихы бойынша алты ресми оқулық негізінде жасалған қазақ тіліндегі алғашқы мамандандырылған сұрақ-жауап ресурсы – Kazakh HistoryQA корпусы ұсынылған және талданған. Корпусқа 208 тақырыптық бөлім, шамамен 180 000 токен және дәл жауап беру аралық координаталарына

сілтеме жасалған 661 QA-жұбы кіреді, бұл HuggingFace кітапханасын қоса алғанда, стандартты машиналық оқу моделінің оқыту құбырларымен үйлесімділікті қамтамасыз етеді. Корпустың көп деңгейлі талдауы жүргізілді, оның ішінде контексттердің, сұрақтар мен жауаптардың ұзақтығы, материалдарды сыныптар мен оқулық бойынша бөлу және сұрақтарды фактілік, хронологиялық, процедуралық, себеп-салдарлық және салыстырмалы деп жіктеу статистикасы жасалды. Сонымен қатар, мәтіннің морфологиялық профилі жасалды, жиілік сөздігі жасалды және пәндік терминдер анықталды. Корпустың сенімділігін эталон ретінде бағалау үшін негізгі сөйлемдерді таңдау модельдері (кездейсоқ таңдау, TF-IDF, BM25) енгізілді, BM25 ең жақсы нәтижелерді көрсетті. Ұсынылған корпус білім берудегі гибридті, онтологиялық тұрғыдан байытылған QA-жүйелерін кейінгі зерттеулер мен әзірлеу үшін негіз болып табылады.

Түйін сөздер: *Қазақстан тарихы, OCR, QA, онтология, білім беру NLP, RAG, тезаурус, онтология.*

M.Zh. Yergesh[1*], B.Zh. Yergesh[1], A. Orazayeva[2], A. Talgat[2]
[1]L.N. Gumilyov Eurasian National University,
010000, Kazakhstan, Astana, 2 Satpaev Street
[2]Kazakh University of Technology and Business named after K. Kulazhanov,
010000, Kazakhstan, Astana, 37 A Mukhamedhanov Street
*e-mail: shymkent90@mail.ru

CONSTRUCTION AND LINGUISTIC ANALYSIS OF THE QUESTION-ANSWER CORPUS OF KAZAKH HISTORYQA

The digitalization of secondary education in Kazakhstan increases the need for specialized Kazakh-language resources that support intelligent question-and-answer systems for school subjects. However, existing QA datasets are mainly focused on high-resource languages and do not cover the content of academic disciplines, particularly the history of Kazakhstan. This paper presents and analyzes in detail the Kazakh HistoryQA corpus, the first specialized question-and-answer resource in the Kazakh language, which is based on six official textbooks on the history of Kazakhstan for grades 5-10. The corpus includes 208 thematic sections, about 180 thousand tokens, and 661 QA-pairs with exact coordinates of the response span, which ensures compatibility with standard machine reading model training pipelines, including the HuggingFace library. A multi-level analysis of the corpus has been conducted: statistics on the length of contexts, questions, and answers, distribution of materials by classes and textbooks, and typology.

Key words: *History of Kazakhstan; corpus; OCR; retrieval; QA; ontology, educational NLP; RAG; question-answer; thesaurus; ontology; evaluation.*

Сведения об авторах

Манас Жантуғанұлы Ергеш* – докторант кафедры «Технологий искусственного интеллекта», Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан; e-mail: shymkent90@mail.ru. ORCID: https://orcid.org/0000-0002-2180-293X.

Бану Жантуғанқызы Ергеш – PhD, заместитель директора департамента цифрового развития Евразийского университета имени Л.Н. Гумилева; e-mail: b.yergesh@gmail.com. ORCID: https://orcid.org/0000-0002-8967-2625.

Айнур Ришатовна Оразаева – PhD, ассистент-профессор кафедры «Информационные технологии», Казахский университет технологий и бизнеса им. К. Кулажанова; e-mail: oar_is@mail.com. ORCID: https://orcid.org/0000-0002-2899-9886.

Амангуль Талгат – сеньор-лектор кафедры «Информационные технологии», Казахский университет технологий и бизнеса им. К. Кулажанова; e-mail: amangul_talgat81@mail.ru. ORCID: https://orcid.org/0000-0002-5800-9134.

Авторлар туралы мәліметтер

Манас Жантуғанұлы Ергеш* – Жасанды Интеллект Технологиялары кафедрасының докторанты, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, Астана, Қазақстан; e-mail: shymkent90@mail.ru. ORCID: https://orcid.org/0000-0002-2180-293X.

Бану Жантуғанқызы Ергеш – PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Цифрлық даму департаменті директордың орынбасары, Астана, Қазақстан; e-mail: b.yergesh@gmail.com. ORCID: https://orcid.org/0000-0002-8967-2625.

Айнур Ришатовна Оразаева – PhD, «Ақпараттық технологиялар» кафедрасы, ассистент-профессор, Қ. Құлажанов атындағы Қазақ технология және бизнес университеті; e-mail: oar_is@mail.com. ORCID: https://orcid.org/0000-0002-2899-9886.

Амангуль Талгат – сеньор-лектор, «Ақпараттық технологиялар» кафедрасы, Қ. Құлажанов атындағы Қазақ технология және бизнес университеті; e-mail: amangul_talgat81@mail.ru. ORCID: https://orcid.org/0000-0002-5800-9134.

Information about the authors

Manas Yergesh* – PhD doctoral student, Department of Artificial Intelligence Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: Shymkent90@mail.ru. ORCID: https://orcid.org/0000-0002-2180-293X.

Banu Yergesh – PhD, deputy director of the Department of Digital Development, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: b.yergesh@gmail.com. ORCID: https://orcid.org/0000-0002-8967-2625.

Ainur Orazayeva – K. Kulazhanov Kazakh University of Technology and Business, Department of Information Technology, Head of the Department, Astana, Kazakhstan; e-mail: oar_is@mail.com. ORCID: https://orcid.org/0000-0002-2899-9886.

Amangul Talgat – K. Kulazhanov Kazakh University of Technology and Business, Department of Information Technology, Head of the Department, Astana, Kazakhstan; e-mail: amangul_talgat81@mail.ru. ORCID: https://orcid.org/0000-0002-5800-9134.

Поступила в редакцию 28.01.2026
Поступила после доработки 22.02.2026
Принята к публикации 23.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-15

МРНТИ: 28.23.15

Б.С. Сарсенов*, Д. Жамангарин
[1]Евразийский национальный университет имени Л.Н. Гумилева,
010000, Республика Казахстан, г. Астана, ул. Александра Пушкина, 11
*e-mail: batyr.sarsenov@gmail.com

ОПТИМИЗАЦИЯ ПРОЦЕССОВ КОНТРОЛЯ И УПРАВЛЕНИЯ В СИСТЕМЕ КОНТРОЛЯ И УПРАВЛЕНИЯ ДОСТУПОМ КОНГРЕСС-ХОЛЛА «АУЛИЕ-АТА»

Аннотация: *В статье рассматривается задача повышения эффективности и надёжности процессов контроля и управления в системе контроля и управления доступом (СКУД) на примере объекта общественного назначения – Конгресс-Холла «Аулие-Ата». Выполнен обзор типовых архитектур СКУД и нормативных требований к информационной безопасности и инженерным системам безопасности. Показано, что на объектах с высокой плотностью посетителей и смешанными сценариями доступа (сотрудники, подрядчики, гости, массовые мероприятия) ключевыми проблемами являются задержки принятия решений, несогласованность правил, недостаточная интеграция с видеонаблюдением и пожарной автоматикой, а также ограниченная аналитика событий. Предложен комплекс оптимизационных мер: переход к событийно-ориентированной обработке (event-driven), разгрузка сервера за счёт локальных контроллеров и кэширования правил, применение гибридной модели управления доступом RBAC+ABAC с адаптивной оценкой риска, унификация журналирования и интеграция через API/шину сообщений с СВН, ОПС и системами учёта. Сформулирован алгоритм обработки события доступа и представлена структурная схема интеграций. Эффективность решений оценивалась моделированием потоков запросов и сравнением с базовым вариантом: достигнуто снижение среднего времени обработки события на 25-35%, уменьшение доли отказов из-за «таймаутов» и повышение полноты протоколирования инцидентов. Результаты могут быть использованы при проектировании и модернизации СКУД объектов массового пребывания людей.*

Ключевые слова: *СКУД; контроль доступа; оптимизация; RBAC; ABAC; оценка риска; интеграция; event-driven; аудит.*

Введение

Безопасность объектов массового пребывания людей определяется не только физической защищённостью периметра, но и качеством процессов управления доступом: насколько быстро и корректно система распознаёт субъект доступа, сопоставляет его права и контекст (время, зона, режим работы), принимает решение и фиксирует событие для