

I SHO'BA

KORPUS LINGVISTIKASINING DOLZARB MASALALARI

ИННОВАЦИИ В ИЗУЧЕНИИ ПОЭЗИИ АБАЯ: РАЗРАБОТКА КАЗАХСКОГО ПОЭТИЧЕСКОГО ПОДКОРПУСА

Мадиева Гульмира

доктор филологических наук, профессор
Казахский национальный университет имени аль-Фараби.

Пирманова Күнсүлу

старший научный сотрудник, постдокторант
Казахский национальный университет имени аль-Фараби.

Серикбаева Айжан

старший научный сотрудник, постдокторант
Казахский национальный университет имени аль-Фараби.

Аннотация: Предлагаемое исследование, посвященное разработке казахского поэтического подкорпуса, актуально в контексте сохранения культурного наследия и применения современных технологий для его анализа. Работа фокусируется на творчестве известного казахского поэта Абая Кунанбаева с целью создания цифровой базы для морфологического и поэтического изучения его стихов. Особенностью проекта является отсутствие аналогов такого рода исследований, что подчеркивает его инновационный характер. Методологические основы проекта включают в себя разработку критериев разметки и адаптацию инструментов обработки естественного языка. Первый этап исследования охватывает разработку теоретической базы и первичный анализ текстов, что обеспечит качественный фундамент для последующего глубокого изучения и классификации данных. Проект имеет значительные перспективы для развития цифровой гуманитарной науки и обогащения культурного пространства Казахстана, открывая новые горизонты для изучения национальной литературы и культуры в цифровую эпоху.

Ключевые слова: цифровые гуманитарные науки, поэтический корпус, казахская поэзия, поэзия Абая, морфологический анализ, обработка естественного языка, культурное наследие.

Annotation: The study dedicated to the development of the Kazakh poetic corpus is pertinent in the context of cultural heritage preservation and the application of modern technologies for its analysis. The focus of the work is on the writings of the renowned Kazakh poet Abai Kunanbaev, creating a digital database for the morphological and poetic study of his verses. A distinctive feature of the project is the absence of direct analogs in such research, highlighting its innovative nature. The methodological foundations of the project include the development of markup criteria and the adaptation of natural language processing tools. The initial stage of the research encompasses the development of a theoretical base and preliminary text analysis, which

will provide a qualitative foundation for further in-depth study and data classification. The project holds significant prospects for the development of digital humanities and enriching the cultural space of Kazakhstan, opening new horizons for the study of national literature and culture in the digital age.

Keywords: digital humanities, poetic corpus, Kazakh poetry, Abai's poetry, morphological analysis, natural language processing, cultural heritage.

В современном мире, в котором цифровые технологии используются практически во всех сферах жизни, становится все более очевидной важность развития цифровых гуманитарных наук. Одним из направлений в этой области является создание и анализ поэтических корпусов, что позволяет не только сохранить культурное наследие, но и предоставить новые возможности для его изучения и интерпретации [Жубанов А., 2013: 25]. Разработка казахского поэтического подкорпуса, сосредоточенная на морфологической и поэтической разметке стихов Абая Кунанбаева, отражает современные тенденции в этой области.

Актуальность и новизна данной темы обусловлены несколькими факторами. Во-первых, творчество Абая Кунанбаева является средоточием, сердцевинной казахской литературы и культуры, однако комплексный цифровой анализ его творчества только начинается. Во-вторых, разработка такого корпуса позволит применять современные методы обработки естественного языка и лингвистического анализа, что откроет новые горизонты для исследования поэтики казахского языка и культуры.

Основанием для разработки данной темы служат исходные данные, полученные из предыдущих лингвистических исследований [Жубанов А., 2016: 159] казахского языка, а также существующие цифровые ресурсы казахской литературы. Важно отметить, что наблюдается недостаток комплексных работ, посвященных анализу стихотворений Абая с использованием современных технологий.

Необходимость проведения научно-исследовательской работы (НИР) обусловлена стремлением расширить понимание поэтической истории казахской культуры и литературы, а также обеспечить сохранение и популяризацию национального наследия в цифровую эпоху. Научно-технический уровень разработки предполагает создание цифровой базы данных, обладающей высоким уровнем доступности и функциональности для проведения морфологического и поэтического анализа.

В рамках подготовки проекта были проведены исследования, которые показали отсутствие прямых аналогов разрабатываемого подкорпуса, что подчеркивает его уникальность и инновационный потенциал. Кроме того, важным аспектом является метрологическое обеспечение НИР по проекту "Разработка казахского поэтического корпуса: морфологическая и поэтическая разметка стихов Абая", которое включает в себя разработку стандартов для морфологической и поэтической разметки текстов.

Связь данной работы с другими научно-исследовательскими работами выражается в ее вкладе в развитие цифровой лингвистики, а также в обогащении культурных и образовательных цифровых ресурсов Казахстана. Разработка казахского поэтического корпуса является не только актуальной задачей, но и важным шагом на пути сохранения и развития национального культурного наследия.

Цель проведенного исследования: разработка и проверка методологических основ для морфологической и поэтической разметки текстов, что включает в себя выбор соответствующих инструментов и технологий, а также определение методических подходов для анализа и классификации стихов Абая.

Задачи проведенного исследования:

- разработка критериев для морфологической и поэтической разметки стихов;
- выбор и адаптация инструментов обработки естественного языка для поэтического наследия Абая;
- проведение первичного анализа выбранных произведений Абая для оценки эффективности методов;
- оценка текущего состояния исследований в данной области для интеграции лучших практик.

Этот этап является фундаментальным для всей программы, так как заложенные здесь методологические основы определяют качество и точность всего последующего анализа. Он направлен на устранение начальных технических и методологических препятствий, что позволит эффективно перейти к следующим этапам исследования, включая расширенный анализ и классификацию данных, а также разработку пользовательского интерфейса для доступа к разрабатываемому подкорпусу.

Работа, выполненная на данном этапе, обеспечивает необходимую основу для дальнейшего углубленного исследования стихов Абая, что способствует достижению конечной цели программы – созданию полноценного, структурированного и легко доступного цифрового малого поэтического корпуса, который будет служить важным ресурсом для исследователей, студентов и всех, кто интересуется казахской литературой и культурой.

С помощью поэтического подкорпуса Абая станет возможным облегчить проведение различных лингвистических исследований поэтического творчества Абая, анализ структуры стихотворений Абая, изучение опыта использования языковых средств для создания образов и эмоционального воздействия, а также исследование развития казахской поэзии как исторического процесса.

1. Выбор направления исследований.

В рамках программы "Разработка казахского поэтического корпуса: морфологическая и поэтическая разметка стихов Абая" было выбрано направление исследований, основанное на современных методах обработки естественного языка и цифровой лингвистике. Основная цель – создание структурированного и аннотированного поэтического корпуса стихотворений Абая Кунанбаева, который будет способствовать глубокому пониманию его творчества и казахской поэтической традиции.

Методы решения задач включали использование алгоритмов машинного обучения для морфологической, поэтической разметки и анализа текстов. Сравнительная оценка с аналогами показала, что применение этих методов для казахского языка является новаторским подходом, способным обеспечить более глубокий и точный анализ.

2. Процесс теоретических и экспериментальных исследований.

Теоретические исследования включали анализ лингвистических особенностей казахского языка и поэтических структур в поэзии Абая: различные методы компьютерной

лингвистики, включая статистический анализ и методы машинного обучения, для обработки текстов и извлечения необходимой информации.

Экспериментальные работы охватывали разработку и тестирование алгоритмов для морфологической разметки, а также создание интерфейса для доступа к корпусу. Было проведено несколько циклов тестирования для оценки точности и эффективности разработанных методов.

3. Обобщение и оценка результатов исследований.

Полученные предварительные результаты показали, что технико-экономическая эффективность проекта проявляется в возможности использования корпуса в образовательных и научных целях, который является уникальным в своем роде благодаря специфики казахского языка и поэзии, что способствует повышению интереса к казахской культуре и языку. В то же время, была выявлена необходимость дополнительных исследований для дальнейшего улучшения алгоритмов морфологической разметки и расширения функциональности корпуса, включая возможности для интерактивного анализа и более глубокой статистической обработки данных.

В целом, поэтический подкорпус Абая ориентирован на то, чтобы сделать поэзию более доступной и понятной как для академического сообщества, так и для широкой публики, а также способствовать более глубокому пониманию и оценке поэтического наследия [<http://web-corpora.net>].

Морфологическая разметка стихов Абая включает в себя процесс присвоения каждому слову в тексте меток, обозначающих их морфологические характеристики. Для этого можно использовать следующие методы и условные обозначения:

Методы морфологической разметки:

Автоматическая разметка: Использование компьютерных программ, которые автоматически анализируют текст и присваивают метки на основе предварительно заданных правил и словарей. Такие программы могут быть обучены на специализированных корпусах казахского языка.

Ручная разметка: Анализ и разметка текста специалистами-лингвистами. Этот метод обеспечивает более высокую точность, особенно в случае сложных или неоднозначных языковых конструкций.

Гибридный подход: Сочетание автоматической и ручной разметки для увеличения точности и эффективности процесса.

Условные обозначения для морфологической разметки [Жубанов А., 2015: 37]:

Для казахского языка могут быть использованы следующие условные обозначения, общепринятые в знаковой системе национальных корпусов:

Существительное (С): N

Прилагательное (П): ADJ

Глагол (Г): V

Наречие (Нар): ADV

Местоимение (Мест): PRON

Числительное (Числ): NUM

Предлог (Предл): PREP

Союз (Союз): CONJ

Междометие (Межд): INTJ

Падеж (например, именительный, винительный): NOM, ACC и т.д.

Время (например, настоящее, прошедшее): PRES, PAST и т.д.

Число (единственное, множественное): SG, PL

Эти обозначения применяются к каждому слову или словоформе в стихотворении для указания их морфологических характеристик. Это позволяет проводить более детальный анализ текста и изучать его языковые особенности.

Создание моделей морфологической разметки для стихов Абая Кунанбайулы представляет собой сложный процесс, который включает несколько этапов:

1. Сбор и подготовка данных:

Репозиторий стихов Абая: централизованное хранилище всех доступных стихотворений Абая, которые собраны из различных источников. Репозиторий предоставляет доступ к этим данным для лингвистических исследований в разных целях с использованием поэтического подкорпуса Абая, разработки языковых моделей, обучения машинного обучения и других целей, связанных с изучением или анализом языка творчества Абая.

Предварительная обработка: чистка текста от технических ошибок (например, исправление опечаток, стандартизация орфографии).

2. Разработка морфологической разметки:

Определение тегов: создание списка морфологических тегов, применимых конкретно к поэтическому языку Абая (например, часть речи, падеж, число, время и т.д.).

Ручная разметка: начальная ручная разметка выборки текстов для создания обучающего набора данных.

3. Обучение модели:

Выбор алгоритма: Определение подходящего алгоритма машинного обучения (например, CRF, LSTM, Transformer).

Обучение: использование размеченных вручную данных для обучения модели.

Валидация и тестирование: проверка эффективности модели на отдельном наборе тестовых данных.

4. Оптимизация и масштабирование:

Итеративная оптимизация: настройка параметров модели и алгоритмов для улучшения точности.

Расширение данных: постепенное добавление новых данных для улучшения обобщающей способности модели.

Технические аспекты:

Использование НЛП библиотек: применение библиотек обработки естественного языка, таких как NLTK, spaCy или машинного обучения (TensorFlow, PyTorch) для реализации модели.

Работа с подкорпусом: управление и обработка больших текстовых данных с использованием соответствующих инструментов и технологий.

Создание такой модели, безусловно, требует значительных усилий и сотрудничества лингвистов, программистов и специалистов по обработке естественного языка, в нашем случае – поэтическое наследие Абая. Это многоступенчатый процесс, требующий постоянного обновления и оптимизации для обеспечения высокой точности и полезности в исследовательских и образовательных целях.

Для примера модели морфологической разметки стихотворной строки Абая "Желсіз түнде жарық ай" (переводится как "Яркая луна в безветренную ночь"), можно использовать следующие условные обозначения:

Желсіз (безветренный):

Часть речи: Прилагательное (ADJ)

Падеж: Именительный (NOM)

Число: Единственное (SG)

түнде (ночь):

Часть речи: Наречие (ADV)

Падеж: В предложном (LOC)

Число: Единственное (SG)

жарық (яркий):

Часть речи: Прилагательное (ADJ)

Падеж: Именительный (NOM)

Число: Единственное (SG)

ай (луна):

Часть речи: Существительное (N)

Падеж: Именительный (NOM)

Число: Единственное (SG)

Таким образом, морфологическая разметка этой строки выглядит следующим образом:

Желсіз (ADJ, NOM, SG)

түнде (ADV, LOC, SG)

жарық (ADJ, NOM, SG)

ай (N, NOM, SG)

Эта разметка позволяет провести детальный анализ морфологических особенностей поэтического языка, использованного в стихотворении Абая, и способствует глубокому пониманию текста как с лингвистической, так и с литературной точек зрения.

Благодарность: Исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № BR21882334).

Список использованных источников:

1. Жұбанов А.Қ. Мәтіндер корпусы – лингвистер үшін ас күшті ақпараттық ресурс // «Фонетика ғылымдарының өзекті мәселелері: даму көздері мен болашағы» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2012. – 186-193 бб.
2. Жубанов А.К., Кабетенов А.Т. Казахская компьютерная лингвистика в период независимости Республики Казахстан // Сборник: Проблемы современной прикладной лингвистики (ответственный - А.В. Зубов). – Минск: МГЛУ, 2013. декабря. – С. 23-30.
3. Жубанов А.К. Принципы автоматизации морфологической разметки текстов Национального корпуса казахского языка (НККЯ) // Материалы международной научно-практической конференции «Контрастивные исследования и прикладная лингвистика». Минск: МГЛУ, 2015. – 227 с.

4. Жұбанов А. Қазақ тілінің ұлттық корпусы – қуатты ақпараттық ресурс // Актуальные проблемы диалектологии языков народов России. XVI Всероссийская конференция. – Уфа, 1–4 июня 2016 г. – 157-161-бб.
5. <http://web-corpora.net>