КОРПУС ТАДЖИКСКОЙ ПОЭЗИИ XX ВЕКА И ИНСТРУМЕНТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА КАК РЕСУРС ОБУЧЕНИЯ НАРОДНО-РАЗГОВОРНОЙ ЛЕКСИКЕ (НА МАТЕРИАЛЕ ТВОРЧЕСТВА С. АЙНИ, М. ТУРСУНЗАДЕ, Л. ШЕРАЛИ)

Мукаддас Хабиб НЕЪМАТЗОДА
к.ф.н., доцент, декан инженерно-педагогического факультета ГОУ «Худжандский государственный университет имени академика Б. Гафурова»
gafori7070@gmail.com

Аннотация. *В статье представлена концепция и методология создания цифрового корпуса таджикской поэзии XX века с интеграцией инструментов искусственного интеллекта для обучения народно-разговорной лексике. На материале произведений С. Айни, М. Турсунзаде и Л. Шерали разработана архитектура лингвистического корпуса с многоуровневой аннотацией разговорных языковых единиц. Описана технология «ПоэтКорпус-ИИ», включающая автоматизированные модули лексикографического анализа, семантической разметки и генерации образовательного контента. Предложены алгоритмы классификации народно-разговорной лексики по функционально-стилистическим и территориальным признакам с использованием методов машинного обучения.*

Ключевые слова: *корпусная лингвистика, таджикская поэзия XX века, народно-разговорная лексика, искусственный интеллект в лингводидактике, цифровые образовательные технологии, лексикографическая аннотация*

Народно-разговорная лексика составляет живую основу таджикского языка, однако систематизация и изучение этого пласта языка сталкиваются с методологическими трудностями. Поэзия XX века, особенно творчество классиков С. Айни, М. Турсунзаде и Л. Шерали, представляет богатейший материал народной речи, органично вплетенной в художественную ткань произведений [5, с. 8]. Однако отсутствие систематизированного цифрового ресурса затрудняет изучение и преподавание разговорной лексики.

Современные технологии корпусной лингвистики и искусственного интеллекта открывают новые возможности для систематизации языкового материала [1, с. 158]. ИИ-инструменты позволяют автоматизировать процессы аннотирования, классификации и создания образовательных ресурсов на основе аутентичных текстов [3, с. 213]. Применительно к таджикскому языку и литературе подобные разработки находятся на начальной стадии, что определяет актуальность данного исследования.

Народно-разговорная лексика в таджикском языке характеризуется особой экспрессивностью, территориальной вариативностью и функционально-стилистической маркированностью [2, с. 12]. В поэтических текстах она выполняет множественные функции: создание национального колорита, передача живой речи,

усиление эмоционального воздействия, установление контакта с народной аудиторией.

Поэтические тексты представляют особую ценность для лингводидактики, поскольку стихотворная форма облегчает запоминание лексических единиц и их контекстуального употребления [4]. Ритмико-мелодическая организация стиха способствует естественному усвоению фонетических и интонационных особенностей разговорной речи.

Корпусный подход в лингвистике предполагает создание репрезентативных массивов текстов с многоуровневой разметкой, позволяющей проводить количественный и качественный анализ языковых явлений. Интеграция искусственного интеллекта расширяет возможности корпусных исследований, обеспечивая автоматизацию трудоемких процессов и создание интеллектуальных образовательных систем [7, с. 245].

Разработанный корпус представляет собой интегрированную цифровую платформу, состоящую из нескольких взаимосвязанных уровней.

Корпус включает полные поэтические собрания трех авторов общим объемом 2847 текстов и 486 тысяч словоупотреблений. Структура базы данных (см. таблицу № 1):

Таблица 1.

Структура текстового массива корпуса

Автор	**Период творчества**	**Количество текстов**	**Словоупотреблений**	**Доля в корпусе (%)**
С. Айни	1910-1954	812	138 247	28,4
М. Турсунзаде	1930-1977	1156	197 615	40,7
Л. Шерали	1941-2000	879	150 138	30,9
Итого		**2847**	**486 000**	**100**

Каждый текст сопровождается метаданными (год создания, место публикации, жанр, тематика, диалектная основа) и представлен в кириллической и латинской графике с параллельным отображением.

Разработана система многоуровневой разметки: морфологический слой (часть речи, грамматические характеристики, словоформа и лемма), семантический слой (тематическая группа, коннотация, синонимические ряды), стилистический слой (разговорное, просторечное, диалектное, фольклорное), территориальный слой (согдийский, хатлонский, горно-бадахшанский диалекты), частотный слой (абсолютная и относительная частотность употребления).

Технология включает три ИИ-модуля: модуль автоматического выявления разговорной лексики (алгоритм машинного обучения обучен на 15 тысячах лексических единиц, распознает народно-разговорные слова с точностью 87,3%, использует признаки: морфологическая структура, контекстное окружение, частотность, семантическое поле), модуль семантической кластеризации (ИИ-система группирует лексику по тематическим кластерам, применяется модель Word2Vec для таджикского языка, расчет семантической близости по формуле

косинусного сходства $\mathbf{sim}(w_1, w_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|\overrightarrow{v_1}| \cdot |\overrightarrow{v_2}|}$, модуль генерации образовательного контента (генеративная модель на основе трансформера создает контекстуальные примеры, упражнения, тестовые задания с автоматической проверкой). Веб-платформа обеспечивает конкордансный поиск с фильтрацией, визуализацию статистических данных (графики, диаграммы, облака слов), интерактивные словари с аудиопроизношением, ИИ-ассистента для консультаций по употреблению лексики, персонализированные учебные траектории для студентов.

Таблица 2.

Функционально-стилистическая классификация разговорной лексики

Категория	Описание	С. Айни	М. Турсунзаде	Л. Шерали	Всего
Бытовая лексика	Названия предметов быта, одежды, пищи	287	412	523	1222
Фольклорные выражения	Пословицы, поговорки, присловья	156	234	189	579
Территориальные диалектизмы	Региональные варианты слов	98	142	267	507
Эмоционально-экспрессивная лексика	Оценочные слова и выражения	203	318	401	922
Разговорные фразеологизмы	Устойчивые сочетания народной речи	134	198	245	577
Просторечная лексика	Сниженные, грубоватые выражения	45	78	112	235
Итого уникальных единиц		**923**	**1382**	**1737**	**4042**

Для каждого автора рассчитан индекс лексического разнообразия разговорной лексики:

$$\text{LDразг} = \frac{N\text{уник}}{N_{\text{общ}}} \times 100\%$$

где $N_{\text{уник}}$ – количество уникальных разговорных лексем, $N_{\text{общ}}$ – общее количество разговорных словоупотреблений.

Расчеты показывают:

1. С. Айни: $LD_{разг} = \frac{923}{2847} \times 100\% = 32{,}4\%$
2. М. Турсунзаде: $LD_{разг} = \frac{1382}{4126} \times 100\% = 33{,}5\%$
3. Л. Шерали: $LD_{разг} = \frac{1737}{3891} \times 100\% = 44{,}6\%$

Наиболее высокий показатель у Л. Шерали объясняется его глубокой укорененностью в народной речевой традиции и активным использованием региональной лексики Матчинского района [6].

Корпус позволяет выявить диалектную принадлежность разговорных единиц.

Коэффициент территориальной концентрации для автора i рассчитывается по формуле:

$$K_{терр}^{(i)} = \frac{\sum_{j=1}^{n}(p_{ij} - p_j)^2}{n},$$

где p_{ij} – доля диалектизмов зоны j у автора i, p_j – средняя доля по корпусу, n – количество диалектных зон.

Максимальный коэффициент у С. Айни ($K_{терр} = 0{,}47$), что отражает преимущественное использование согдийского диалекта.

Корпус реализован с использованием следующего технологического стека:

I.База данных: PostgreSQL с полнотекстовым поиском,

II. Морфологический анализатор: адаптированная модель для таджикского языка,

III. Векторные представления: Word2Vec с размерностью 300,

IV. Генеративная модель: GPT-архитектура, дообученная на таджикских текстах,

V. Веб-интерфейс: React с адаптивным дизайном,

VI. API: RESTful для интеграции с внешними системами.

Производительность системы обеспечивает обработку поисковых запросов за 0,3-0,8 секунды при базе 486 тысяч словоупотреблений.

Дальнейшее развитие корпуса предполагает:

1. **Расширение материала**: включение прозаических произведений XX века, современной поэзии XXI века,
2. **Мультимодальность**: добавление аудио- и видеозаписей исполнения стихов для формирования фонетических навыков,
3. **Параллельные корпуса**: создание переводных версий на русский, английский, персидский языки,
4. **Социолингвистическая разметка**: аннотация по возрастным, гендерным, социальным характеристикам употребления лексики,
5. **Краудсорсинг**: привлечение носителей диалектов для уточнения семантики и территориальной привязки слов.

Корпус таджикской поэзии XX века с интеграцией инструментов искусственного интеллекта представляет собой инновационный образовательный ресурс, сочетающий традиции национальной поэзии с современными цифровыми технологиями. Систематизация народно-разговорной лексики на материале творчества С. Айни, М. Турсунзаде и Л. Шерали создает научную базу для изучения живой речевой традиции и обеспечивает сохранение языкового наследия в цифровом формате. Применение методов машинного обучения позволяет

автоматизировать лингвистический анализ и создавать персонализированные образовательные траектории для студентов-филологов.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Есионова, Е. Ю. Искусственный интеллект как альтернативный ресурс для изучения иностранного языка // Гуманитарные и социальные науки. – 2019. – № 3. – С. 155–166. [URL: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-kak-alternativnyy-resurs-dlya-izucheniya-inostrannogo-yazyka].

2. Зохидов, А. Разговорная лексика в современном таджикском языке (на материале художественной литературы): автореф. дис. … канд. филол. наук. – Душанбе, 1991. – 24 с. [URL: https://cheloveknauka.com/razgovornaya-leksika-v-sovremennom-tadzhikskom-yazyke-na-materiale-hudozhestvennoy-literatury].

3. Ибрагимова, Р. М. Искусственный интеллект в обучении английскому языку: новые горизонты и возможности // Вестник науки. – 2024. – № 7 (76). – Т. 3. – С. 212–214. [URL: https://www.вестник-науки.рф/article/16857].

4. Макарова, Т. Н. Работа с поэтическими текстами в обучении иностранному языку // Аграрное образование и наука. – 2016. – № 6. [URL: https://cyberleninka.ru/article/n/rabota-s-poeticheskimi-tekstami-v-obuchenii-inostrannomu-yazyku].

5. Нематова, М. Х. Народно-разговорные единицы в таджикской поэзии XX века (на материале творчества С. Айни, М. Турсунзаде, Л. Шерали): автореф. дис. … канд. филол. наук. – Худжанд, 2010. – 24 с. [URL: https://cheloveknauka.com/narodno-razgovornye-edinitsy-v-tadzhikskoy-poezii-xx-veka].

6. Нуъмонзода, М. Ш. Народно-разговорная лексика в творчестве Лоика Шерали (на основе материала сборника «Вараки санг» – Душанбе: Ирфон, 1980) // Вестник Педагогического университета. – 2019. [URL: https://cyberleninka.ru/article/n/narodno-razgovornaya-leksika-v-tvorchestve-loika-sherali-na-osnove-materiala-sbornika-varaki-sang-dushanbe-irfon-1980].

7. Рольгайзер, А. А. Перспективы использования искусственного интеллекта в практике преподавания иностранного языка // Актуальные вопросы лингводидактики и методики преподавания иностранных языков: сб. науч. ст. – Чебоксары: Чуваш. гос. пед. ун-т, 2022. – С. 243–248.