ПАРАЛЛЕЛЬНЫЕ ПОДКОРПУСЫ В СТРУКТУРЕ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА: ДИАГНОСТИКА ТЕКУЩЕГО ЭТАПА И ГОРИЗОНТЫ ТРАНСФОРМАЦИИ

Нигматова Лолахон Хамидовна
доктор филологических наук (DSc),
профессор Бухарский государственный университет,
кафедра русского языка и литературы
nigmatovalolaxon@gmail.com

Аннотация: *Настоящее исследование проводит всестороннюю диагностику позиций параллельных подкорпусов в архитектуре Национального корпуса узбекского языка (НКУЯ), интерпретируя его как опорный элемент для корпусной лингвистики и инфраструктуры обработки естественного языка в условиях агглютинативных систем с ограниченными ресурсами. Опираясь на эмпирику платформы uzbekcorpus.uz и инновационные проекты 2024–2025 годов, раскрываются прорывы в алгоритмизированной аннотации, межъязыковой синергии с предобученными архитектурами и аппликации в трансляционных конвейерах. Уделяется внимание системным барьерам, вытекающим из дефицита данных, и стратегическим осям эволюции: диверсификация многоязычных конфигураций (русско-узбекская, англо-узбекская, тюрко-узбекская), механизация семантического выравнивания на сегментном и фразовом уровнях, а также имплантация в академические, индустриальные и государственные NLP-платформы. Формулируются операциональные стратегии по конструированию адаптивной инфраструктуры, способствующей глобальному цифровому продвижению узбекского языка в эру искусственного интеллекта.*

Ключевые слова: *Национальный корпус узбекского языка, параллельные подкорпусы, корпусная лингвистика, нейронный машинный перевод, обработка агглютинативных языков с низкими ресурсами, большие языковые модели.*

ВВЕДЕНИЕ

В контексте ускоренной цифровизации лингвистических ресурсов национальные корпуса эволюционируют в многомерные платформы, обеспечивающие не только консервацию и деконструкцию языковых феноменов, но и прогнозирование их динамики, особенно для языков с ограниченными данными, таких как узбекский. Национальный корпус узбекского языка (НКУЯ), эволюционирующий под эгидой ведущих институтов Узбекистана, аккумулирует гетерогенные подкорпусы, где параллельные сегменты выступают в роли интерлингвальных интерфейсов, связующих моноязыковые потоки с трансграничными взаимодействиями. В отличие от статичных монокорпусов, эти структуры не только кодифицируют имманентные паттерны, но и моделируют трансляционные векторы, семантические дрейфы и культурные

метаморфозы, способствуя интеграции в глобальные NLP-экосистемы.Данная работа, опираясь на синтез эмпирических данных из цифровых репозиториев и аналитических публикаций 2021–2025 годов, включая свежие инициативы Министерства цифровых технологий по созданию корпуса объемом 10 миллиардов слов для больших языковых моделей, проводит диагностику актуального статуса параллельных подкорпусов НКУЯ и проектирует их трансформационные горизонты. Методологический аппарат интегрирует корпусную аналитику, элементы машинного обучения и прикладную тюркологию, с акцентом на специфику агглютинативных морфосинтаксических конструкций. Расширение анализа включает сравнительный обзор с аналогичными проектами) и оценку влияния на AI-разработки, подчеркивая роль параллельных данных в преодолении барьеров низкоресурсных языков.

Актуальный статус параллельных подкорпусов НКУЯ.Архитектоника, метрики и эволюция объема.Платформа uzbekcorpus.uz, курируемая коллективом под руководством доктора наук Н. Абдурахмоновой из Национального университета Узбекистана, интегрирует субкорпусы различного профиля: авторские, диалектные, педагогические, мультимедийные и параллельные. К 2025 году совокупный объем базового массива превысил 75 миллионов токенов, с преобладанием современных нарративов и исторических артефактов, а параллельный субкорпус достиг 10 миллионов слов, включая бинарные и полилингвальные конфигурации. Приоритет отдается русско-узбекским и англо-узбекским парам, дополненным экспериментальными блоками, такими как узбекско-казахский, сформированным из литературных источников, новостных потоков и верифицированного перевода.Недавние обновления, инициированные в 2024–2025 годах, включают интеграцию мультимедийных элементов и веб-краулинговых данных, что расширяет репрезентативность. Аннотационный пайплайн сочетает полуручные и автоматизированные этапы: морфологический и синтаксический теггинг опирается на конечные автоматы (FST) и кастомизированные модели UdPipe, адаптированные к агглютинативному строю. Точность POS-категоризации варьируется от 87% до 92% на валидационных наборах, с использованием векторных эмбеддингов для выравнивания в Hugging Face экосистеме. Однако, несмотря на прогресс, метрики уступают эталонным системам, что подчеркивает необходимость масштабирования. Методологические инновации и прорывы Горизонты трансформации предполагают симбиоз с AI: параллельные corpora как трамплин для zero-shot перевода и культурно-адаптированных чатботов. Потенциальные сценарии включают интеграцию с VR-обучением и реал-тайм анализом социальных сетей, преодолевая барьеры через федеративное обучение.

ЗАКЛЮЧЕНИЕ

Параллельные подкорпусы НКУЯ символизируют переход узбекского языка к цифровому суверенитету, интегрируя лингвистическое наследие с передовыми технологиями. Диагностика фиксирует баланс прорывов (NER, sentiment analysis) и ограничений (data sparsity), а траектория, ориентированная на экспансию и AI-синергию, позиционирует НКУЯ как пивот тюркской вычислительной лингвистики.

Перспективы фокусируются на эмпирической валидации в downstream-задачах, конвертируя потенциал в трансформационные решения для культурной устойчивости и глобальной интероперабельности.

СПИСОК ЛИТЕРАТУРЫ:

1. Abdurakhmonova N. Architectural Foundations of Uzbek Digital Corpus. Tashkent, 2021.
2. Karshiev A., Tursunov M. Processing Pipeline for Agglutinative Languages in Uzbekcorpora.uz. 2022.
3. Uzbek-Kazakh Parallel Framework for Neural MT: Methodologies and Evaluations. ACL Anthology, 2024.
4. Mengliyev B. Conceptual Blueprint of National Uzbek Corpus. 2018.
5. Abdurakhmonova N. Morphosyntactic Annotation Strategies for Uzbek. 2022.
6. Formal Paradigms in Uzbek Digital Corpus. 2021.
7. Pedagogical Dimensions of Uzbek Subcorpora: Architecture and Efficacy. 2023.
8. Parallel Structures in Translation Ecosystem Development. 2024.
9. Problems and Resolutions in Uzbek National Corpus Construction. ResearchGate, 2025.
10. Multimedia Integration in Uzbek Language Corpus. Conference Proceedings, 2024.
11. Named Entity Recognition Dataset for Uzbek NLP. Mendeley Data, 2024.
12. Stemming Library for Uzbek Morphological Processing. Patent, 2024.
13. Syntactic Parser for Uzbek Sentence Decomposition. Patent, 2024.
14. Terminological Platform Based on Parallel Corpora. Patent, 2024.