КОРПУСНЫЕ МЕТОДЫ И ИИ В ИССЛЕДОВАНИЯХ УЗБЕКСКОЙ ПРЕССЫ РУБЕЖА XX ВЕКА

Эркинжон Камилович САТИБАЛДИЕВ
Старший преподаватель
Узбекский государственный университет мировых языков

Аннотация. *Статья посвящена анализу возможностей искусственного интеллекта (ИИ) и корпусных методов в изучении узбекской прессы рубежа XX века. Показано, что сочетание нейросетевых технологий, статистических моделей и гибридных лингвистических инструментов существенно расширяет аналитический потенциал гуманитарных исследований, позволяя выявлять динамику политического, культурного и образовательного дискурса в газетах начала XX века. Особое внимание уделяется формированию специализированных корпусов узбекской прессы, проблемам нормализации многослойной орфографии, методам автоматического извлечения тем, анализа авторского стиля, семантического дрейфа и количественной реконструкции медиапространства.*

Ключевые слова: *искусственный интеллект, корпусная лингвистика, узбекская пресса, джадидизм, исторические медиатексты, семантическое моделирование, цифровые гуманитарные науки.*

Узбекистан начала XX века был одним из важнейших центров общественного обновления в Средней Азии. Газеты и журналы, издававшиеся на узбекском, тюркском и татарском языках, отражали ключевые тенденции эпохи: образовательные реформы, модернизацию общественной мысли, культурно-национальное пробуждение, рост политической активности, столкновение традиций и нового социального порядка. Однако исследование этих источников сталкивается с трудностями, связанными с их фрагментарностью, графической неоднородностью, нестабильностью орфографии и огромным объёмом материалов.

Развитие ИИ и корпусной лингвистики открывает возможность системного изучения медийного пространства рубежа XX века, позволяя анализировать большие массивы исторических текстов и выявлять закономерности, ранее недоступные традиционным филологическим методам.

Создание корпусов узбекской прессы: задачи и трудности

Корпус узбекской прессы рубежа XX века должен учитывать специфику исторических текстов:

– использование арабографической письменности,
– орфографические колебания,
– параллельное существование чагатайской нормы и ранней литературной узбекской нормы,
– разнообразие жанров (редакционные статьи, научно-просветительские колонки, фельетоны, объявления, письма читателей).

Нормализация таких текстов требует специальных алгоритмов. Нейросетевые OCR-модели обучаются на изображениях газетных полос, учитывают особенности дореволюционных шрифтов и способны распознавать как печатные, так и рукописные элементы [1].

После распознавания тексты проходят этап морфологической нормализации, где ИИ-модели приводят вариативные формы слов к единому лемматизированному виду. Этот шаг важен для корректного статистического анализа, поскольку орфография начала XX века крайне нестабильна [2].

Статистические методы: частотность, коллокации и реконструкция узбекского медийного языка

Корпусные методы позволяют вычислять частоты ключевых слов и выражений, выявлять устойчивые словосочетания, фиксировать развитие медийного стиля.

Анализ словников газет показал, что тексты тех лет насыщены терминами, связанными с модернизацией и образованием: *ma'rifat*, *islahat*, *taraqqiyot*, *millat*, *xalq*, *ilm*, *maktab* и др. Частотный профиль помогает определить, когда те или иные темы становятся центральными в общественном дискурсе [3].

Методы коллокационного анализа позволяют изучать контексты, в которых употребляются ключевые термины. Например, понятие *millat* («нация/народ») часто сопутствует лексемам, связанным с просвещением и единством, что свидетельствует о формировании национальной концептосферы.

Семантическое моделирование и выявление дискурсивных трендов

Семантические модели на основе векторных представлений слов (word embeddings) позволяют проследить семантический дрейф понятий. Переход от культурно-просветительской трактовки реформ к политизированной риторике 1920-х годов отражается в изменении ближайших семантических соседей таких слов, как *taraqqiyot* («прогресс») или *hurriyat* («свобода») [4].

Тематическое моделирование (LDA) позволяет выделять ключевые темы:

- школьная реформа,
- положение женщин,
- культура и литература,
- экономика и рынок,
- религиозные дискуссии,
- урбанизация и инфраструктурные проекты.

Каждая тема имеет свою динамику, что позволяет историкам количественно реконструировать интеллектуальную повестку эпохи.

Стилометрия и вопрос авторства

Стилометрические методы, основанные на машинном обучении, дают возможность исследовать индивидуальные особенности публицистов. Для газет рубежа XX века это особенно важно, так как многие тексты публиковались без подписи или под псевдонимами.

ИИ-модели анализируют частотность служебных слов, характерные синтаксические конструкции, длину предложений, индекс разнообразия лексики. Такие алгоритмы позволяют определить вероятного автора статьи и выявить изменения его стиля в разные годы [5].

Стилометрия также помогает дифференцировать редакционную правку от авторского текста, что имеет важное значение для изучения газетных школ и формирования литературных норм.

Семантические сети и моделирование медиапространства

ИИ-технологии позволяют строить графовые модели текстовых взаимосвязей. Такие модели отображают:

– какие темы доминировали в разных периодах,
– как понятия связывались между собой,
– какие издания являлись ключевыми узлами информационной коммуникации,
– как происходила циркуляция идей между журналистами Бухары, Самарканда, Ферганы и Ташкента [6].

Этот подход даёт возможность реконструировать социальные и интеллектуальные сети, в которых существовала узбекская пресса переходного периода.

ИИ и визуальная составляющая газет

В газетах начала XX века важную роль играли иллюстрации, карикатуры и графические элементы. ИИ-модели компьютерного зрения могут классифицировать такие изображения, выявлять повторяющиеся визуальные мотивы, определять жанровые особенности сатирической графики [7].

Это особенно ценно для изучения газет *Mulla Nasriddin*, *Oyina*, *Shuhrat* и др., где визуальная коммуникация была не менее выразительной, чем текстовая.

Ограничения цифровых методов

Несмотря на мощь ИИ-алгоритмов, исследователь должен учитывать ряд ограничений:

– низкое качество старых газет и повреждённых страниц,
– нестандартные орфографические варианты,
– необходимость культурно-исторической интерпретации данных,
– риск «механического» анализа без учёта контекста эпохи [8].

ИИ дополняет, но не заменяет историческую и филологическую экспертизу.

Перспективы направления

В ближайшие годы ожидается:

– создание единого открытого корпуса узбекской прессы 1890–1930 гг.,
– интеграция моделей ИИ с историческими базами данных,
– развитие алгоритмов нормализации арабского письма, – глубокая визуально-текстовая аналитика,
– автоматическая реконструкция утраченных частей газет [9].

Эти шаги позволят существенно углубить анализ медиадискурса эпохи и предложат новый взгляд на культурную историю Узбекистана.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Springmann, M., Lüdeling, A. OCR of historical printings with neural networks. *DH Conference Proceedings*, 2017.

2. Piotrowski, M. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 2012.

3. McEnery, T., Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.

4. Hamilton, W., Leskovec, J., Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. *ACL*, 2016.

5. Holmes, D. I. The evolution of stylometry in humanities scholarship. *LLC*, 1998.

6. Moretti, F. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, 2005.

7. Yanai, K. Image recognition in historical documents. *CVPR Workshops*, 2018.

8. Jockers, M. L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

9. Underwood, T. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.