

**КОРПУСНЫЙ АНАЛИЗ И ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ТЕКСТА:
ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ НАУЧНО-ПЕДАГОГИЧЕСКОЙ ТЕРМИНОЛОГИИ В
УЗБЕКСКОМ И РУССКОМ ЯЗЫКАХ**

Рахматова Гулнора Турсиновна

gulnoraraxmatova429@gmail.com

Старший преподаватель кафедры

языкового и литературного образования,

UNIVERSITY OF BUSINESS AND SCIENCE

(Наманган, Узбекистан);

Махмудова Манзура Мухарамовна

учитель русского языка Политехникума №1

Чустского района, Наманганская область,

Республика Узбекистан.

Аннотация: *Актуальность. Данная статья посвящена сопоставительному исследованию научно-педагогической терминологии в узбекском и русском языках с использованием методов корпусной лингвистики и интеллектуальной обработки текста (NLP). В условиях цифровизации образования и интеграции научных пространств возникает необходимость автоматизированного анализа и унификации терминосистем.*

Цель. Выявление структурно-семантических и деривационных особенностей педагогических терминов, а также определение степени межъязыковой эквивалентности в двух разноструктурных языках.

Методы. Используются методы корпусного анализа (выделение ключевых слов, анализ конкордансов), статистические метрики TF-IDF, а также дистрибутивный семантический анализ. Эмпирическую базу составили специально сформированные параллельные и сопоставимые корпуса научно-педагогических текстов объемом 1,2 млн словоупотреблений.

Результаты. Определены наиболее продуктивные модели терминообразования. В узбекском языке преобладают аналитические конструкции (субстантивные словосочетания) и аффиксация (суффиксы -лик, -чилик), в русском — морфологический способ (префиксально-суффиксальный) и многокомпонентные атрибутивные группы. Выявлены зоны терминологической лакуарности и асимметрии, обусловленные лингвокультурными факторами.

Практическая значимость. Результаты исследования могут быть применены при создании двуязычных автоматизированных словарей, систем машинного перевода педагогического дискурса и разработке интеллектуальных образовательных платформ.

Ключевые слова: корпусная лингвистика, интеллектуальная обработка текста, NLP, педагогическая терминология, узбекский язык, русский язык, сопоставительный анализ, TF-IDF, конкорданс.

Abstract: *Background. The article is devoted to a comparative study of scientific and pedagogical terminology in the Uzbek and Russian languages using the methods of corpus linguistics and natural language processing (NLP). In the context of the digitalization of education and the integration of scientific spaces, there is a need for automated analysis and unification of terminology systems.*

Purpose. *Identification of structural, semantic, and derivational features of pedagogical terms, as well as determining the degree of interlinguistic equivalence in two structurally different languages.*

Methods. *Corpus analysis methods (keyword extraction, concordance analysis), TF-IDF statistical metrics, and distributive semantic analysis were used. The empirical database consisted of specially formed parallel and comparable corpora of scientific and pedagogical texts with a volume of 1.2 million tokens.*

Results. *The most productive models of term formation were identified. In the Uzbek language, analytical constructions (substantive phrases) and affixation (suffixes -lik, -chilik) predominate, while in Russian, the morphological method (prefixal-suffixal) and multicomponent attributive groups prevail. Areas of terminological lacunarity and asymmetry caused by linguacultural factors are identified.*

Practical value. *The results of the study can be applied to the creation of bilingual automated dictionaries, machine translation systems for pedagogical discourse, and the development of intelligent educational platforms.*

Keywords: *corpus linguistics, natural language processing, NLP, pedagogical terminology, Uzbek language, Russian language, comparative analysis, TF-IDF, concordance.*

1. ВВЕДЕНИЕ (INTRODUCTION)

Современный этап развития лингвистической науки характеризуется активным внедрением технологий искусственного интеллекта и методов интеллектуальной обработки текста (Natural Language Processing — NLP). Корпусные методы перешли из разряда вспомогательных в категорию базовых инструментов количественного и качественного анализа языковых систем. Одной из наиболее динамично развивающихся и одновременно сложных для унификации подсистем языка является терминология, в частности — научно-педагогическая терминосистема.

Педагогическая наука в Узбекистане и странах СНГ переживает период коренной трансформации, вызванной переходом на компетентностный подход, цифровизацией образовательной среды и интеграцией в международное образовательное пространство. Этот процесс сопровождается лавинообразным появлением новых

терминов, переосмыслением старых понятий и активным межъязыковым заимствованием.

Сопоставительное изучение научно-педагогической терминологии (НПТ) узбекского и русского языков представляет особый научный интерес, так как данные языки относятся к совершенно разным типологическим группам: узбекский язык является агглютинативным (тюркская семья), а русский — флективным (индоевропейская семья). Различие грамматического строя накладывает существенный отпечаток на способы номинации педагогических понятий. Несмотря на наличие фундаментальных трудов по сопоставительному терминоведению (Ш. Шахабиддинова, Р. Данияров, В.М. Лейчик, С.В. Гринев-Гриневи́ч), специфика именно педагогической терминологии в паре «узбекский-русский» с применением автоматизированных методов обработки больших текстовых данных (Big Data) остается малоизученной. Большинство существующих словарей и справочников составлены традиционными методами, что не позволяет фиксировать реальное функционирование терминов в живом научном дискурсе.

Целью данного исследования является выявление структурно-семантических, деривационных и статистических особенностей функционирования научно-педагогической терминологии в узбекском и русском языках на основе корпусного анализа и алгоритмов интеллектуальной обработки текста.

2. МЕТОДЫ И МАТЕРИАЛЫ (METHODS)

Для достижения поставленной цели автором был сформирован специализированный сопоставимый и параллельный научно-педагогический корпус (UzRu-PedCorp).

2.1. Формирование корпуса

Материалом для формирования корпуса послужили:

– Научные статьи из рецензируемых журналов (ВАК Республики Узбекистан и ВАК Российской Федерации) по специальности «Педагогика» за период 2020–2026 гг.

– Авторефераты диссертаций.

– Современные учебники и учебные пособия для высших учебных заведений.

Общий объем корпуса составил 1 214 500 словоупотреблений (токенов):

– Узбекский подкорпус (Uz-SubCorp) — 594 200 токенов.

– Русский подкорпус (Ru-SubCorp) — 620 300 токенов.

2.2. Этапы интеллектуальной обработки текста (NLP-конвейер)

Процесс обработки текстовых массивов включал следующие шаги:

1. Токенизация и очистка: Удаление стоп-слов, знаков препинания, цифровых обозначений и метаразметки.

2. Лемматизация и POS-теггинг (определение частей речи):

– Для русского языка использовалась библиотека PyMorphy

–Для узбекского языка применялся разработанный авторами морфологический анализатор на основе правил (Rule-based) и словаря основ, учитывающий специфику агглютинативного наращивания аффиксов.

3. Экстракция терминов (Term Extraction):

Выделение однокомпонентных и многокомпонентных терминологических единиц осуществлялось с помощью статистической метрики TF-IDF (Term Frequency-Inverse Document Frequency) в сочетании с лингвистическими фильтрами (шаблонами частей речи, например:

Noun + Noun, Adj + Noun).

Формула расчета веса терминологичности кандидата (TF-IDF):

$$TF - IDF(t, d, D) = TF(t, d) \times$$

где t — термин-кандидат, d — конкретный документ, D — вся коллекция документов корпуса.

4. Анализ конкордансов и коллокаций:

Использовались метрики взаимной информации (MI — Mutual Information) и T-score для определения устойчивости многокомпонентных терминов.

3. РЕЗУЛЬТАТЫ (RESULTS)

В ходе интеллектуального анализа текста из корпусов было извлечено и верифицировано экспертным путем 2450 терминологических единиц (1180 в узбекском и 1270 в русском подкорпусах).

3.1. Структурно-количественный анализ терминосистем

Автоматизированный анализ распределения терминов по количеству компонентов показал существенные различия между языками (Таблица 1).

Таблица 1. Распределение педагогических терминов по количеству компонентов (в %)

Количество компонентов	Узбекский подкорпус (Uz-SubCorp)	Русский подкорпус (Ru-SubCorp)
Однокомпонентные (моноксемные)	28,4%	34,2%
Двухкомпонентные	48,6%	41,5%
Трехкомпонентные	18,2%	19,1%
Многокомпонентные (4 и более)	4,8%	5,2%

Как видно из таблицы, в обоих языках преобладают составные термины (полилексемные), однако в узбекском языке доля двухкомпонентных единиц выше (48,6%). Это связано с активным использованием изафетных конструкций (примыкания

существительных) типа «существительное + существительное» (например, ta'lim sifati, dars samaradorligi).

3.2. Деривационные особенности терминообразования

Узбекский язык (Агглютинативная деривация)

Корпусный анализ позволил выявить наиболее продуктивные словообразовательные аффиксы в узбекской педагогической терминологии:

-lik: Абстрактные существительные, обозначающие свойство или качество (kompetentlilik — компетентность, ijodkorlik — творчество/креативность, * mustaqillik* — самостоятельность). Частотность данного аффикса среди производных терминов составляет 42\%.

-chi: Обозначение деятеля, субъекта образовательного процесса (tarbiyachi — воспитатель, izlanuvchi — исследователь).

-ish / -uv: Обозначение процесса опредмеченного действия (o'qitish — обучение, boshqaruv — управление).

Спецификой узбекского языка является синтаксический способ — образование терминов-сочетаний по модели N + N (изафет II типа, где второе слово принимает аффикс принадлежности -i/-si): ta'lim jarayoni (процесс образования), tarbiya usuli (метод воспитания).

Русский язык (Флективно-фузионная деривация)

В русском подкорпусе доминируют морфологические способы:

Суффиксальный: суффиксы -ость (личность, успешность), -ние (обучение, воспитание, оценивание), -изация (цифровизация, гуманизация).

Префиксально-суффиксальный: со-действие, со-творчество. Синтаксическая модель в русском языке чаще всего строится по схеме Adj + N (прилагательное + существительное): дистанционное обучение, инклюзивное образование.

3.3. Семантическая плотность и ключевые концепты (TF-IDF анализ)

Топ-10 высокочастотных терминов с наибольшим весом TF-IDF в обоих подкорпусах отражают современную парадигму образования (Таблица 2).

Таблица 2. Топ-10 педагогических терминов по метрике TF-IDF

Р анг	Узбекский термин	Частот а (IPM*)	Русский эквивалент	Частота (IPM*)
1	Ta'lim tizimi	2450	Система образования	2610
2	Kompetensiya	1820	Компетенция	1950

3	Pedagogik texnologiya	1640	Педагогическая технология	1580
4	Mustaqil ta'lim	1210	Самостоятельна я работа	1340
5	Innovatsiya	1150	Инновация	1100
6	Baholash mezonlari	980	Критерии оценивания	1020
7	Masofaviy o'qitish	890	Дистанционное обучение	940
8	Kreativlik	750	Креативность	710
9	Shaxsga yo'naltirilgan	680	Личностно- ориентированный	730
10	Raqamlashtirish	620	Цифровизация	690

*Примечание: IPM (Instances Per Million) — количество вхождений на миллион словоупотреблений.

4. ОБСУЖДЕНИЕ (DISCUSSION)

Сопоставительный анализ, подкрепленный данными интеллектуальной обработки текста, выявил ряд зон лингвистической асимметрии между узбекской и русской педагогическими терминосистемами.

4.1. Межъязыковая асимметрия и лакунарность

Интеллектуальный анализ контекстов (конкордансов) показал, что полной межъязыковой эквивалентности удастся достичь в основном в сфере интернационализмов (компетенция — kompetensiya, инновация — innovatsiya). Однако в пласте исконной лексики наблюдается семантическое расхождение. Например, русский термин «обучение» в зависимости от контекста переводится на узбекский язык двояко:

O'qitish — если акцент делается на деятельность преподавателя (процесс передачи знаний).

O'rganish — если речь идет о деятельности учащегося (процесс усвоения). В обобщенном научно-методическом контексте чаще используется синтетический термин ta'lim, который, однако, шире по значению и часто эквивалентен русскому «образование».

Интересна динамика термина «самостоятельная работа». В русском дискурсе это устоявшееся выражение. В узбекских корпусах зафиксирован сдвиг в сторону *mustaqil ta'lim* (буквально: самостоятельное образование), что подчеркивает большую автономию субъекта в современной узбекской педагогической модели.

4.2. Влияние цифровизации на деривационные процессы. Методы NLP позволили зафиксировать «неологический бум» в обоих языках за последние 5 лет. Происходит калькирование англоязычных терминов. В узбекском языке наблюдается гибридное терминообразование: к заимствованной основе добавляется национальный аффикс: *blended learning* \rightarrow *aralash ta'lim*, *digitalization* \rightarrow *raqamlashtirish*.

В русском языке процесс идет по пути суффиксации: геймификация, вебинаризация.

Корпусный анализ конкордансов показывает, что новые термины в узбекском языке имеют более высокую вариативность (колебания между использованием заимствования и кальки), в то время как в русском языке терминография быстрее кодифицирует новые единицы.

5. ЗАКЛЮЧЕНИЕ (CONCLUSION)

Применение корпусного анализа и методов интеллектуальной обработки текста (NLP) позволило провести объективное, количественно обоснованное исследование научно-педагогической терминологии узбекского и русского языков.

Типологическая обусловленность. Структура терминов жестко детерминирована грамматическим строем. Флективный характер русского языка способствует развитию морфологической деривации и атрибутивных цепочек (Adj + N). Агглютинативная природа узбекского языка определяет доминирование аффиксации (особенно суффикса *-lik*) и беспредложных субстантивных сочетаний (изафет).

Конвергенция терминосистем. Под влиянием глобализации и цифровизации образования (развитие концептов *e-learning*, *STEM*, компетенции) ядра обеих терминосистем демонстрируют высокую степень семантической конвергенции, выраженную в росте индекса TF-IDF для интернациональных терминов.

Перспективы. Разработанные в ходе исследования алгоритмы экстракции терминов и морфологического анализа узбекского текста могут быть интегрированы в интеллектуальные лингвистические системы: автоматизированные переводчики, терминологические базы данных нового поколения и системы электронного обучения (LMS).

ЛИТЕРАТУРА (REFERENCES):

1. Абдурахманова Н. З. Лингвистические основы создания национального корпуса узбекского языка // СФЕРА НАУКИ: Международный научный журнал. — 2021. — № 11(2). — С. 34–39.

2. Баранов А. Н. Введение в прикладную лингвистику: учебное пособие. 4-е изд., испр. и доп. — М.: Флинта, 2022. — 368 с.
3. Гринев-Гриневиц С. В., Сорокина Э. А. Современные тенденции развития терминоведения // Вестник МГОУ. Серия: Лингвистика. — 2020. — № 2. — С. 6–15.
4. Гринев-Гриневиц С. В. Терминоведение: учеб. пособие для студ. высш. учеб. заведений. — М.: Академия, 2008. — 304 с.
Данияров Р. Развитие узбекской технической терминологии. — Ташкент: Фан, 1989. — 132 с.
5. Захаров В. П., Хохлова М. В. Корпусная лингвистика: институционализация и перспективы развития // Научно-техническая информация. Серия 2: Информационные процессы и системы. — 2023. — № 5. — С. 18–27.
6. Искандарова Ш. М. Теоретические и практические вопросы узбекской компьютерной лингвистики // Иностранные языки в Узбекистане. — 2022. — № 4 (45). — С. 112–124.
7. Лейчик В. М. Терминоведение: Предмет, методы, структура. — Изд. 6-е. — М.: ЛЕНАНД, 2024. — 256 с.
8. Мамедов А. Н., Каримов С. С. Интеллектуальный анализ научно-педагогического текста: квантитативные методы и алгоритмы NLP // Вестник Национального университета Узбекистана. — 2025. — № 1/2. — С. 143–151.
9. Шахабиддинова Ш. Х., Ахмедова Н. А. Агглютинативная деривация в корпусных исследованиях тюркских языков // Филологические науки. Вопросы теории и практики. — 2022. — Т. 15, № 8. — С. 2541–2547.
10. Шахабиддинова Ш. Х. Тюркское морфемное и словообразовательное моделирование в сопоставительном аспекте // Востоковедение. — 2021. — № 2. — С. 45–54.
11. Abdurakhmanova N., Mengliev B. Development of Morphological Analyzer for the Uzbek Language based on Finite State Transducers // International Journal of Advanced Computer Science and Applications (IJACSA). — 2021. — Vol. 12, No. 5. — P. 412–419.
12. Kurova M. I., Tarasova E. V. Computational and Corpus Tools for Educational Terminology Extraction // Journal of Psycholinguistic Research. — 2024. — Vol. 53, No. 2. — P. 201–216.
13. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. — MIT Press, 1999. — 680 p.
14. Rakhmatova G. T. The Role of Pedagogical Terms in Modern Pedagogy: Problems and Solutions (A Comparison of Russian and Uzbek Terms)— 2025. Volume 5 | Issue 4 | Apr - 2025 88–95.
15. Salimov B. R. Linguistic modeling of educational discourse in bilingual environments // Central Asian Journal of Theoretical and Applied Sciences. — 2023. — Vol. 4, No. 6. — P. 77–84.