

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И КОРПУСНОЙ ЛИНГВИСТИКИ К ИССЛЕДОВАНИЮ НАУЧНО-ПЕДАГОГИЧЕСКОЙ ТЕРМИНОЛОГИИ В УЗБЕКСКОМ И РУССКОМ ЯЗЫКАХ

Рахматова Гулнора Турсиновна

*Старший преподаватель кафедры языкового и литературного образования,
University of Business and Science, г. Наманган, Республика Узбекистан. E-mail:
gulnoraraxmatova429@gmail.com*

Аннотация: *Статья посвящена сопоставительному изучению научно-педагогической терминологии русского и узбекского языков корпусными и компьютерными методами. Цель работы — выявить типы терминологических соответствий в параллельных текстах и описать протокол их автоматической проверки. Материалом послужили мини-корпус из трёх пар параллельных фрагментов по дидактике, компетентностному подходу и теории воспитания (95 слов в русской и 99 слов в узбекской части) и список из 25 пар терминов с лингвистическими комментариями. В корпусе представлено 12 терминов списка. В 11 случаях узбекский эквивалент в тексте совпал с зафиксированным в списке; исключение составило русское «обучение», которое передано словом o'qitish в одном контексте из трёх, а в двух других — словом ta'lim. В узбекской части зафиксировано 10 форм на -ish/-sh, тогда как в русской части отглагольных существительных на -ние — 17 словоупотреблений; ряд русских существительных передан в узбекском тексте аналитическими сочетаниями. Предложен протокол этапа машинного обучения: извлечение терминов трансформерными моделями и сопоставление терминов по многоязычным векторным представлениям. Результаты носят пилотный характер и показывают, какие явления — многозначность ta'lim, аналитические эквиваленты, агглютинативная морфология — будут определять качество автоматической обработки.*

Ключевые слова: *научно-педагогическая терминология; корпусная лингвистика; машинное обучение; автоматическое извлечение терминов; терминологическая эквивалентность; русский язык; узбекский язык; многоязычные векторные представления.*

Abstract: *The article presents a contrastive study of scientific and pedagogical terminology in Russian and Uzbek, combining corpus-linguistic and machine-learning approaches. The aim is to identify the types of terminological correspondence in parallel texts and to outline a protocol for their automatic verification. The material consists of a mini-corpus of three pairs of parallel passages on didactics, the competence-based approach and the theory of upbringing (95 words in the Russian part and 99 words in the Uzbek part) and a list of 25 term pairs with linguistic annotations. Twelve terms from the list occur in the corpus. In eleven cases the Uzbek equivalent found in the text coincides with*

the one recorded in the list; the exception is the Russian «обучение», rendered as o'qitish in one of three contexts and as ta'lim in the other two. The Uzbek part contains 10 verbal nouns in -ish/-sh, while the Russian part contains 17 tokens of deverbal nouns in -nue, and several Russian nouns are rendered in Uzbek by analytic constructions. A protocol for the machine-learning stage is proposed: term extraction with transformer models and cross-lingual term matching with multilingual sentence embeddings. The results are preliminary and show which phenomena — the polysemy of ta'lim, analytic equivalents and agglutinative morphology — are likely to determine the quality of automatic processing.

Keywords: *scientific and pedagogical terminology; corpus linguistics; machine learning; automatic term extraction; terminological equivalence; Russian; Uzbek; multilingual embeddings.*

ВВЕДЕНИЕ

Педагогическая терминология Узбекистана развивается в условиях двуязычия. Учебники, нормативные документы и научные статьи по педагогике существуют одновременно на узбекском и русском языках, а значительная часть понятийного аппарата пришла через международные заимствования. Поэтому вопрос о том, насколько точно термин одного языка передаётся термином другого, важен не только для лингвистики: от него зависят качество подготовки учителей и единообразие образовательных стандартов.

Сопоставительные работы в этой области уже существуют. Г. Т. Рахматова анализирует культурно-специфическую педагогическую лексику русского и узбекского языков в цифровой среде и, в частности, разбирает пары «образование — ta'lim» и «обучение — o'qitish» [1]. Ф. Ч. Нормаматова сопоставляет семантику педагогических терминов узбекского и английского языков [2]. Э. Б. Яхшиева показывает, что многозначность встречается и внутри узбекской педагогической терминологии, и предлагает различать лексемы-термины и собственно термины [3]; Ф. Ниязова описывает семантические связи педагогических терминов как отдельное семантическое поле [4]. Насколько можно судить по аннотациям этих публикаций, они строятся на ручном сопоставлении примеров, а не на проверке по размеченному корпусу.

Вычислительная база узбекского языка за последние годы заметно выросла. Создан морфологически и синтаксически размеченный корпус [5]; собран крупный набор новостных, прессовых и юридических текстов по 15 категориям, на котором лучший результат показала модель BERTbek [6]; эта модель затем описана как предобученная языковая модель для узбекского языка, причём сравниваются три варианта, обученные на данных разного объёма и происхождения [7]. В 2026 г. опубликованы тематический лексикон UzThemeLex (4 945 терминов по агрономии, экономике и праву) [8] и семантическая сеть UzWordNet-GoldCore (11 178 лексических записей) [9]. Педагогической терминологии среди этих ресурсов нет, и именно этот пробел определил постановку нашей задачи.

Для извлечения терминов сегодня применяют нейросетевые модели. Обзор [10] прослеживает путь от признаковых подходов к трансформерам. Сравнительное исследование [11] показало, что одноязычные модели в большинстве изученных языков превосходят многоязычные, если не извлекать именованные сущности, а в работе [12] продемонстрировано, что дообученный BERT переносит между языками и предметными областями информацию о «терминологически окрашенных» контекстах. Итоги этих экспериментов обобщены в диссертации [13]. Для межъязыкового сопоставления терминов подходят многоязычные векторные представления предложений, такие как LaBSE, рассчитанные более чем на сто языков [14].

В русскоязычной науке корпусные методы в терминологии и в обучении языкам тоже развиваются. Ю. И. Бутенко с соавторами описывают структурные модели терминов как основу автоматической разметки корпусов научно-технических текстов [15]; О. В. Львова анализирует возможности лингвистических корпусов для формирования коммуникативных компетенций [16]. Для узбекского и родственных тюркских языков появляются параллельные ресурсы для машинного перевода [17, 18].

Таким образом, есть описания педагогической терминологии двух языков, есть корпусные и нейросетевые инструменты, но нет работ, где одно соединялось бы с другим на материале узбекско-русских параллельных текстов. Цель нашей статьи — выявить типы терминологических соответствий в таких текстах и обосновать протокол их автоматической проверки. Для этого решаются четыре задачи: описать мини-корпус и список терминов; определить, какие термины списка реализуются в текстах и какими эквивалентами; охарактеризовать морфологические модели и пласты происхождения терминов; предложить схему этапа машинного обучения. Рабочее предположение состоит в том, что часть расхождений в эквивалентности связана с многозначностью узбекского *ta'lim* и с различием словообразовательных моделей и что именно эти расхождения будут затруднять автоматическое извлечение и сопоставление терминов.

2. Материалы и методы

3. Материал исследования

Материалом послужили два набора данных. Первый — список из 25 пар «русский термин — узбекский термин». Для каждой пары указаны область (общая педагогика, дидактика, организация образования, формы образования и др.) и лингвистическое примечание о типе эквивалентности: калькирование, интернациональный компонент, различие значений и т. п. Второй набор — мини-корпус из трёх пар параллельных фрагментов: «Общая дидактика (сущность и принципы обучения)», «Современные подходы в образовании (компетентностный подход)» и «Теория и методика воспитания». Русская часть содержит 95 словоупотреблений, узбекская (латиница) — 99. При подсчёте слово определялось как последовательность букв, апострофов и дефисов, поэтому дефисные

компози́ты («духовно-нравственных», «ma'naviy-axloqiy») считались одним словом, а тире не учитывалось.

Объём корпуса невелик, и мы сознательно рассматриваем работу как пилотную: она служит проверкой процедуры и источником гипотез, но не основанием для статистических обобщений.

Корпусно-лингвистическая процедура

Анализ включал пять шагов.

1. Сплошная выборка. Каждый термин списка искали в русской части корпуса с учётом словоизменения, затем находили его контекстный эквивалент в узбекской части.

2. Контекстное выравнивание. Фрагменты сопоставляли на уровне предложений и словосочетаний; фиксировали и те пары, которых нет в списке, но которые встретились в тексте (например, «принцип — tamoyil»).

3. Классификация соответствий. Соответствие считалось полным, если в каждом контексте узбекский эквивалент совпадал с зафиксированным в списке; частичным — если эквивалент менялся от контекста к контексту; расхождением — если в тексте стоял иной термин, чем в списке.

4. Морфологическая характеристика. Подсчитывали русские отглагольные существительные на -ние и узбекские формы на -ish/-sh, а также прилагательные на -viy, -li/-lik и имена деятеля на -chi/-uvchi. Подсчёт выполнен скриптом на Python (токенизация регулярным выражением, суффиксы выделялись по окончаниям); полученные списки форм проверены вручную.

5. Характеристика по происхождению. Узбекские термины соотносили с тремя пластами: исконно тюркским, арабо-персидским и европейско-международным.

6. Протокол этапа машинного обучения

Второй этап рассчитан на расширенный материал, поскольку мини-корпус слишком мал для обучения моделей. Протокол состоит из трёх компонентов.

Предобработка: токенизация и морфологический анализ узбекского текста, лемматизация русского.

Для агглютинативного узбекского языка лемматизация критична, так как одна терминологическая единица представлена множеством словоформ.

Извлечение терминов: задача ставится как разметка последовательностей — каждому токену присваивается метка начала, продолжения термина или отсутствия термина.

Для русского текста предполагается дообучение многоязычного трансформера, для узбекского — сравнение BERTbek [7] с многоязычной моделью. Выбор опирается на результаты [11], согласно которым одноязычные модели чаще дают лучший результат, и на данные [12] о переносе между языками при нехватке размеченных данных. Качество оценивается по точности, полноте и F1-мере относительно ручной разметки.

Межъязыковое сопоставление: для русского термина и узбекских кандидатов строятся векторы LaBSE [14] и вычисляется косинусная близость; эквивалентом считается кандидат с наибольшей близостью. Метрика — доля случаев, когда верный эквивалент оказывается первым или в первой тройке. Поддержка русского и узбекского языков моделью следует проверить по актуальному списку языков перед запуском. Эталоном служит список из 25 пар.

В настоящей статье полностью реализован первый (корпусно-лингвистический) этап; второй представлен как протокол, а таблица для его результатов приведена в разделе 3.

Ограничения материала

Узбекские фрагменты по структуре воспроизводят русские, поэтому часть соответствий может отражать переводную кальку, а не независимое узуальное употребление терминов. Разметку выполнял один исследователь, согласованность между аннотаторами не измерялась. Наконец, 12 терминов, найденных в корпусе, — слишком малая выборка, чтобы говорить о частотных закономерностях.

3. Результаты

Количественные характеристики мини-корпуса

Таблица 1. Размер и лексическое разнообразие мини-корпуса

Показатель	Русская часть	Узбекская часть
Пар фрагментов	3	3
Словоупотреблений (токенов)	95	99
Разных словоформ (типов)	81	89
Отношение типов к токенам	0,85	0,90
Средняя длина слова, знаков	8,1	8,2
Знаков с пробелами	873	947

Число слов в обеих частях почти совпадает (95 и 99), однако узбекский текст длиннее по числу знаков (947 против 873) и содержит больше разных словоформ (89 против 81). На выборке в 194 слова отношение типов к токенам нестабильно, и мы не считаем его свойством языка; показатель приводится как ориентир для расширенного корпуса.

Терминологическое покрытие

Из 25 терминов списка в корпусе представлены 12 (48 %): № 2, 3, 9–16, 18 и 20. Остальные 13, среди которых «образовательный процесс», «закономерности обучения», «учебный план» и «дистанционное обучение», в текстах не встретились. Объяснение лежит в тематике фрагментов, а не в качестве списка.

Таблица 2. Термины списка в корпусе и их узбекские эквиваленты в контексте

№	Русский термин	Контекст (рус.)	Эквивалент в узбекском тексте	Соответствие
2	воспитание	«процессов обучения и воспитания»	tarbiya	полное
3	обучение	«Обучение представляет...»; «результатов обучения»; «процессов обучения»	o'qitish (T1); ta'lim (T2, T3)	частичное (1 из 3)

№	Русский термин	Контекст (рус.)	Эквивалент в узбекском тексте	Соответствие
9	учебный предмет, дисциплина	«учебных дисциплин»	o'quv fanlari	полное
10	наглядность	«принцип наглядности»	ko'rgazmalilik	полное
11	непрерывное образование	«систему непрерывного образования»	uzluksiz ta'lim	полное
12	компетентностный подход	«внедрение компетентностного подхода»	kompetensiyaviy yondashuv	полное
13	умение	«умений и навыков»; «умение применять»; «в умении... сочетать»	ko'nikma (3 из 3)	полное
14	навык	«умений и навыков»	malaka	полное
15	знание	«усвоение знаний»; «полученные знания»	bilim (2 из 2)	полное
16	оценивание	«методов оценивания»	baholash	полное
18	воспитательная работа	«Воспитательная работа направлена...»	tarbiyaviy ish	полное
20	педагогическое мастерство	«Педагогическое мастерство учителя»	pedagogik mahorat	полное

Полное соответствие зафиксировано для 11 из 12 терминов (около 92 %). Различие между «умением» и «навыком», на которое указывает комментарий к списку, в корпусе подтверждается: в паре «умений и навыков» им последовательно отвечают ko'nikma и malaka.

Расхождения и внесписочные пары

Единственное частичное соответствие относится к термину «обучение». В определении дидактики (T1) ему отвечает o'qitish, но в сочетаниях «результатов обучения» и «процессов обучения и воспитания» узбекский текст использует ta'lim. Слово ta'lim встречается в узбекской части четыре раза и покрывает три русских понятия: «образование» (uzluksiz ta'lim), «обучение» (ta'lim natijalari, ta'lim va tarbiya) и «обучающиеся» (ta'lim oluvchilar). Тем самым ta'lim выступает в нашем материале как родовое слово, а o'qitish — как название деятельности педагога.

Кроме терминов списка, контекстное выравнивание выявило пары, которых в списке нет (таблица 3).

Таблица 3. Внесписочные пары, выявленные в корпусе

Русский	Узбекский	Частота (рус./узб.)	Замечание
процесс	jarayon	2 / 2	арабо-персидская основа
принцип	tamoyil	2 / 2	вместо международного «prinsip»
содержание	mazmun	1 / 1	в «содержании учебных дисциплин»
обучающиеся	ta'lim oluvchilar	1 / 1	аналитическое сочетание
учащиеся	o'quvchilar	1 / 1	соответствует «ученик»
взаимодействие	o'zaro ta'sir	1 / 1	в списке для № 25 указаны hamkorlik, muloqot

Русский	Узбекский	Частота (рус./узб.)	Замечание
творческие способности	ijodiy qobiliyatlar	1 / 1	прилагательное на -iy

Два наблюдения относятся к самому списку. Во-первых, для педагогического взаимодействия в нём названы hamkorlik и muloqot, а в корпусе взаимодействие педагога и обучающихся передано сочетанием o'zaro ta'sir; этот вариант стоит добавить. Во-вторых, в T1 русское «усвоение знаний» представлено узбекским сочетанием bilim mashqlarini o'zlashtirish, где появляется слово mashq («упражнение»). Это может быть особенностью перевода или неточностью, и фрагмент нужно проверить.

Отдельную группу образуют русские существительные, которым в узбекском тексте соответствуют многословные сочетания: «запоминание» — yodlab olish, «пересмотр» — qayta ko'rib chiqish, «внедрение» — joriy etish. Одно русское слово становится двумя и более узбекскими токенами, а «развитие», «формирование» и «усвоение» передаются формами на -ish с каузативными и другими аффиксами (rivojlantirish, shakllantirish, o'zlashtirish).

Морфологические модели и происхождение терминов

Таблица 4. Продуктивные словообразовательные модели в корпусе

Модель	Русская часть	Узбекская часть
Отглагольное имя	-ние/-ение: 17 словоупотреблений, 15 словоформ	-ish/-sh: 10 словоупотреблений, 10 словоформ
Относительное и качественное прилагательное	не подсчитывалось	-viy: 3 (kompetensiyaviy, tarbiyaviy, ma'naviy-axloqiy); -li/-lik: 3 (ko'rgazmalilik, maqsadli, tizimli)
Имя деятеля	не подсчитывалось	-chi/-uvchi: 3 (oluvchilar, o'qituvchi, o'quvchilar)

Отглагольные существительные на -ние составляют около 18 % русских словоупотреблений (17 из 95), узбекские формы на -ish/-sh — около 10 % (10 из 99). Разницу отчасти объясняют многословные соответствия из п. 3.3: часть русских имён в узбекском тексте выражена сочетанием, а не одним словом на -ish. Узбекское ko'rgazmalilik образовано от ko'rgazma с помощью -li и -lik, русская «наглядность» — от прилагательного «наглядный» с помощью -ость; модели различны, а значение совпадает.

По происхождению узбекские термины корпуса распределяются между тремя пластами (таблица 5).

Таблица 5. Происхождение узбекских терминов корпуса

Пласт	Примеры
Тюркский	bilim, o'qitish, o'quv, ko'nikma, yondashuv, uzluksiz, ish
Арабо-персидский	ta'lim, tarbiya, mahorat, mazmun, malaka, fan, tamoyil, jarayon, natija
Международный	pedagog, pedagogik, didaktika, kompetensiya

Слово *baholash* совмещает арабский по происхождению корень *baho* с тюркскими аффиксами. Названия знания, умения и обучающего действия (*bilim*, *ko'nikma*, *o'qitish*) в корпусе тюркские, а наименования содержания, воспитания и мастерства (*ta'lim*, *tarbiya*, *mahorat*, *mazmun*, *malaka*) восходят к арабо-персидскому пласту.

Результаты этапа машинного обучения

Этап машинного обучения в рамках данной версии работы не выполнялся: вычисления по протоколу п. 2.3 не проводились, поэтому числовых значений в таблице 6 нет. Таблица задаёт форму, которую следует заполнить после запуска моделей.

Таблица 6. Форма для результатов сопоставления терминов (заполняется после запуска)

Пара терминов	Косинусная близость LaBSE	Ранг узбекского эквивалента	Метка модели извлечения (рус. / узб.)
обучение — o'qitish	[]	[]	[] / []
обучение — ta'lim	[]	[]	[] / []
образование — ta'lim	[]	[]	[] / []
умение — ko'nikma	[]	[]	[] / []
навык — malaka	[]	[]	[] / []
воспитание — tarbiya	[]	[]	[] / []

4. Обсуждение

Пилотный анализ показал высокое совпадение контекстных узбекских эквивалентов с теми, что зафиксированы в списке (11 из 12), но единственное исключение относится к центральному термину.

В работе [1] приводятся пары «образование — *ta'lim*» и «обучение — *o'qitish*»; наш материал показывает, что в связном тексте *ta'lim* обслуживает обе русские единицы, а *o'qitish* появляется в определении, где акцент сделан на деятельности педагога.

Это согласуется с выводом [3] о многозначности в узбекской педагогической терминологии и с предложением разделять лексемы-термины и собственно термины: в нашем материале *ta'lim* ведёт себя как родовой термин, а в сочетаниях *ta'lim natijalari* и *ta'lim va tarbiya* несёт частные значения.

В русле [4] можно добавить, что *ta'lim*, *tarbiya* и *o'qitish* образуют единое семантическое поле, внутри которого выбор слова определяется контекстом.

Для составителей словарей отсюда следует практический вывод: недостаточно фиксировать пару слов, нужно указывать условия выбора эквивалента. Расширение списка вариантом *o'zaro ta'sir* и пометкой о сочетании *ta'lim oluvchilar* сделало бы его полнее.

Для автоматической обработки результаты ставят три задачи. Первая связана с многословными эквивалентами (*yodlab olish*, *joriy etish*): русский однословный

термин соответствует узбекской последовательности из двух токенов, поэтому извлечение должно работать с терминами разной длины, как в [11, 12].

Вторая — агглютинативная морфология: одна терминологическая единица представлена множеством словоформ, и без морфологического анализа или субсловной токенизации частоты будут занижены; здесь BERTbek [7] выглядит естественным кандидатом.

Третья — расхождение ресурсов: UzThemeLex [8] охватывает агрономию, экономику и право, но не педагогику, поэтому размеченных педагогических данных для обучения нет, и межъязыковой перенос [12] остаётся единственным реалистичным путём, эффективность которого надо проверять.

Можно ожидать, что при сопоставлении по LaBSE [14] вектор слова *ta'lim* окажется близок и к «образованию», и к «обучению»; именно эту гипотезу должна проверить таблица 6.

Подход к описанию структурных моделей терминов [15] сопоставим с нашей морфологической характеристикой (п. 3.4), а идея использовать корпус как учебный ресурс [16] применима и к терминологии: фрагменты с контекстами терминов пригодны для занятий по профессиональному переводу педагогических текстов.

Ограничений несколько. Корпус невелик (194 слова), узбекские фрагменты воспроизводят русские и могут отражать перевод-кальку, разметку выполнял один исследователь, а этап машинного обучения не запускался. Поэтому полученные доли (48 % покрытия, 92 % полных соответствий) нельзя переносить на всю педагогическую терминологию.

Следующие шаги: расширить корпус учебниками и нормативными актами и разметить его двумя аннотаторами с оценкой согласованности; выполнить протокол п. 2.3 и заполнить таблицу 6; дополнить список недостающими вариантами; сопоставить результаты с естественными, не переводными узбекскими текстами по педагогике.

5. Заключение

В статье на материале мини-корпуса из трёх пар параллельных фрагментов и списка из 25 пар терминов описаны соответствия между русской и узбекской научно-педагогической терминологией. Основные результаты таковы. Первое: из 25 терминов списка в корпусе представлены 12, и для 11 из них узбекский эквивалент в тексте совпал с зафиксированным. Второе: исключение составило «обучение», которому в двух контекстах из трёх соответствует *ta'lim*, а не *o'qitish*; *ta'lim* покрывает понятия «образование», «обучение» и «обучающиеся». Третье: русские отглагольные существительные на -ние часто передаются в узбекском тексте формами на -ish/-sh или многословными сочетаниями (*yodlab olish*, *joriy etish*), что осложняет извлечение терминов на уровне токенов. Четвёртое: предложен протокол этапа машинного обучения, включающий извлечение

терминов трансформерными моделями и сопоставление по многоязычным векторным представлениям.

Полученные результаты пилотные. Их проверка на расширенном корпусе с двумя аннотаторами и запуск описанного протокола — ближайшая задача. Практический выход работы — основа двуязычного глоссария педагогических терминов с указанием контекстных условий выбора эквивалента.

Ключевые термины и понятия / Key terms and concepts

№	Русский	English	Пояснение
1	научно-педагогическая терминология	scientific and pedagogical terminology	термины науки об образовании и обучении
2	термин	term	слово или сочетание, называющее понятие области знания
3	терминологическая эквивалентность	terminological equivalence	соответствие терминов в двух языках
4	корпусная лингвистика	corpus linguistics	изучение языка по собранным текстам
5	параллельный корпус	parallel corpus	тексты и их переводы, сопоставленные между собой
6	машинное обучение	machine learning	построение моделей, обучаемых на данных
7	автоматическое извлечение терминов	automatic term extraction (ATE)	выделение терминов из текстов программой
8	токенизация	tokenization	разбиение текста на слова и их части
9	морфологическая разметка	morphological annotation	приписывание словоформам грамматических меток
10	лемматизация	lemmatization	приведение словоформы к словарной форме
11	векторное представление (эмбеддинг)	embedding	числовой вектор слова или предложения
12	трансформер	transformer	архитектура нейросети для обработки текста
13	предобученная языковая модель	pretrained language model	модель, заранее обученная на больших текстах
14	разметка последовательностей	sequence labeling	приписывание метки каждому токenu
15	межъязыковой перенос	cross-lingual transfer	использование знаний одного языка для другого
16	ресурсно ограниченный язык	low-resource language	язык с малым числом данных и инструментов
17	агглютинация	agglutination	присоединение к основе цепочки аффиксов
18	калькирование	calque, loan translation	пословный перенос структуры слова или сочетания
19	многозначность термина	terminological polysemy	несколько значений у одной терминологической единицы
20	компетентностный	competence-based	ориентация обучения на результат в

№	Русский	English	Пояснение
	подход	approach	виде компетенций

СПИСОК ЛИТЕРАТУРЫ:

1. Рахматова Г. Т. Сравнительный анализ культурно-специфической лексики в сфере педагогической терминологии в условиях цифровой среды (на примере узбекского и русского языков) // Oriental Renaissance: Innovative, educational, natural and social sciences. 2025. Т. 5, № 9. С. 345–355.
2. Normamatova F. Ch. Ingliz va o'zbek tillarida pedagogikaga oid terminlarning semantik tahlili // Oriental Renaissance: Innovative, educational, natural and social sciences. 2024. С. 68–72.
3. Yaxshiyeva E. B. O'zbek tilining pedagogik terminologiyasida ko'p ma'nolilik hodisasi // Oriental Renaissance: Innovative, educational, natural and social sciences. 2023. С. 562–568.
4. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language. arXiv:2210.15234. 2022.
5. Kuriyozov E., Salaev U., Matlatipov S., Matlatipov G. Text classification dataset and analysis for Uzbek language. arXiv:2302.14494. 2023.
6. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // Proceedings of SIGUL 2024. ACL Anthology: 2024.sigul-1.5.
7. Saidov B. UzThemeLex Dataset: An Uzbek Thematic Lexicon for Domain Terminology and Weakly Supervised NER. Mendeley Data, V1. 2026. DOI: 10.17632/jgfkz6493p.1.
8. Alaev R. и др. UzWordNet-GoldCore: A Gold-Core Uzbek Lexical-Semantic WordNet Dataset. Mendeley Data, V1. 2026. DOI: 10.17632/v7m7v3jdn5.1.
9. Tran H. T. H. и др. The Recent Advances in Automatic Term Extraction: A survey. arXiv:2301.06767. 2023.
10. Tran H. T. H., Martinc M., Pelicon A., Doucet A., Pollak S. Ensembling Transformers for Cross-domain Automatic Term Extraction // Proceedings of ICADL 2022. P. 90–100. arXiv:2212.05696.
11. Hazem A., Bouhandi M., Boudin F., Daille B. Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data // Proceedings of LREC 2022. Marseille, 2022. P. 648–662.
12. Tran H. T. H. Neural Approaches to Automatic Terminology Extraction: PhD dissertation. La Rochelle University; Jožef Stefan International Postgraduate School, 2024.
13. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT Sentence Embedding // Proceedings of ACL 2022. Vol. 1. DOI: 10.18653/v1/2022.acl-long.62.

14. Бутенко Ю. И., Николаева Н. С., Карцева Е. Ю. Структурные модели англоязычных терминов для автоматической обработки корпусов научно-технических текстов // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. 2022. Т. 13, № 1. С. 80–95. DOI: 10.22363/2313-2299-2022-13-1-80-95.

15. Львова О. В. Использование ситуативного вокабуляра и лингвистического квазикорпуса для изучения языка // Вестник РУДН. Серия: Информатизация образования. 2022. Т. 19, № 2. С. 90–96. DOI: 10.22363/2312-8631-2022-19-2-90-96.

16. Mamasaidov M., Shopulatov A. Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak // Proceedings of WMT 2024. P. 606–613. DOI: 10.18653/v1/2024.wmt-1.48.

17. Mamasaidov M. и др. Filling the Gap for Uzbek: Creating Translation Resources for Southern Uzbek // Proceedings of WMT 2025. ACL Anthology: 2025.wmt-1.83.