



Uzbek-German Digital Thesaurus Platform

Author: Abdurashid Abdukarimov\ **Project Start:** July 2025\ **Affiliation:** Codex Terra, Green UZ, OpenAI collaborative project\ **Legal Status:** Protected by the Berne Convention, Uzbek Copyright Law, and OpenAI session archive (SHA-256 ID reference pending)



Оглавление

1. Введение
 2. Структура тезауруса и база данных
 3. Этапы разработки
 4. Тематические направления
 5. AR/AI/EdTech Модули
 6. Научная и правовая база
 7. Вывод
 8. Список литературы
-

I. Введение

Проект 2: Мультиязычная онлайн-энциклопедия на базе тезауруса

Цель: создать доступную, расширяемую, инклюзивную энциклопедию нового типа, которая соединяет культуры, языки, поколения и знания, формируя цифровой мост между народами.

♦ **Источник знаний:** на базе мультиязычного тезауруса, интеграция с Википедией, Wikidata, национальными корпусами и цифровыми архивами вузов, библиотек и культурных институций (DE, UZ, EN)

♦ **Функции:**

- Статьи в формате JSON/HTML с мета-разметкой и AR/AI поддержкой
- Визуализация знаний: карты, схемы, словесные сети, инфографика
- Автоматическая генерация глоссариев по темам, базам и доменам
- Голосовой ввод/прослушивание (TTS), поддержка жестового языка

♦ **Расширяемость и редакция:** Каждый пользователь может адаптировать материал под свои цели (как Lego): добавить, перевести, визуализировать, экспортировать в PDF, AR или преподнести в форме лекции

♦ **Социальная интеграция:** Публикация статей и фрагментов в Telegram, Instagram, Facebook, LinkedIn

♦ **Миссия:** Создание образовательной среды без языковых и цифровых барьеров для всех возрастов, включая людей с ограничениями по зрению, слуху, обучению. Модель ориентирована на страны с растущей демографией и дефицитом преподавателей, учебников и доступа к библиотекам.

Эта энциклопедия становится гуманитарной инфраструктурой: она не только учит, но и объединяет, сохраняя культурное наследие и способствуя цифровой демократии.

Проект 1: Немецко-узбекский — узбекско-немецкий интерактивный онлайн-словарь-тезаурус

Цель: создать цифровую платформу, объединяющую функционал двуязычного словаря и тезауруса с мультимодальной поддержкой (AR, TTS, JSON, NLP), направленную на перевод, обучение и лингвистическую интеграцию.

♦ Двуязычная поддержка: полная симметрия перевода DE ↔ UZ\ ♦ JSON-структуры, позволяющие использовать данные в CAT-инструментах и образовательных системах\ ♦ Интерфейс: веб и мобильный, доступный без регистрации, с возможностью локализации под миграционные или образовательные потребности\ ♦ Связь с корпусами: автоматическая загрузка примеров из DWDS, uzWas и Leipzig Corpora\ ♦ Расширяемость: пользователь может вносить примеры, термины, фразеологизмы, использовать систему голосования

В основе проекта лежит идея демократизации доступа к знаниям, устранения языковых барьеров и продвижения инклюзивного цифрового обучения в странах с растущей демографией и ограниченными образовательными ресурсами.

Проект направлен на разработку мультязычного тезауруса нового поколения, ориентированного на узбекский и немецкий языки, с интеграцией ИИ, корпусной лингвистики и цифровых технологий (AR, TTS, JSON, NLP). Актуальность определяется необходимостью поддержки образования, миграции, терминологической точности и переводов. Повышенный интерес подтверждается и международным научным сообществом, где тема развивается в рамках проектов eLex, AMuSE и национальных программ. Вся разработка зафиксирована в авторских декларациях, защищена в системе Codex Terra и сопровождается открытым философским и гуманитарным манифестом.

II. Структура тезауруса и база данных

- **Lemma:** основная форма слова (например: ilm)
- **POS:** часть речи (например: s.m. — существительное мужского рода)
- **Definition:** краткое определение
- **Translation:** эквивалент на немецком языке
- **Domain:** предметная область (туризм, логистика, медицина)
- **Frequency:** частота в корпусе
- **Register:** стиль (официальный, разговорный, научный)

```
{
  "lemma": "kitob",
  "pos": "noun",
  "definition": "bilim manbai",
  "translation": {"de": "das Buch"},
  "frequency": 124,
  "register": "neutral",
  "examples": [
    {"uz": "Men kitob o'qiyapman.", "de": "Ich lese ein Buch."}
```

```
]
}
```

Также создаются расширенные лексические гнёзда (пример: музей), включающие определения, перевод, домены, подтипы, примеры, связанные термины и мультязычные конструкции (UZ, DE, EN).

III. Этапы разработки

Корпусная и парсинговая база Основу языкового наполнения составляет корпусная лингвистика и автоматический парсинг текстов. Для узбекского языка используются:

- Национальный корпус узбекского языка (NKMUZ — www.corpus.uz)
- Узбекский веб-корпус uzWac (SketchEngine)
- Корпус университетов (TATU, NUUZ)
- Частные коллекции речевых актов (оцифровка интервью, радио, TikTok)

Для немецкого языка:

- DWDS (Digitales Wörterbuch der deutschen Sprache — www.dwds.de)
- DeReKo (Institut für Deutsche Sprache)
- OpenThesaurus.de и EuroParl
- Wikipedia+Wiktionary dumps (DE-XML)

Парсинг выполняется с помощью:

- spaCy + Uzbek/German модели POS/NER
- stanza + `uzbek_morphology/stanza_de_models`
- собственный модуль `UzTagger` на базе CRF и BiLSTM

Результаты парсинга используются для извлечения лемм, частот, коллокаций, построения гнёзд и семантических графов.

Этап I. Подготовка источников Словари: *izohli lug'at*, *etimologik lug'at*, *tarixiy lug'at*, DWDS.de, Multitran, Phrasebook

Этап II. Извлечение и аннотирование Инструменты: pdfplumber, spaCy, stanza, regex, OCR. Форматы: JSON, TEI/XML, CSV

Этап III. Корпусная обработка и семантика SketchEngine, uzWac, Leipzig, FAIR Text+, WordNet

Этап IV. Интеграция в цифровые системы

- CAT-инструменты: TBX, TMX, XLIFF
- Word Add-In
- ERP-системы: SAP, 1C
- Мобильные AR-приложения
- Портал узбекской лексикографической платформы

IV. Тематические направления

A. Туризм: buchen, check-in, Trinkgeld → buyurtma qilmoq, ro'yxatdan o'tish, choypuli

B. Логистика: Fracht, Drohne, Telematik → yuk, dron, telematika

C. Образование и миграция: Studium, Sprachkurs, Zertifikat → ta'lim, til kursi, sertifikat

D. Транспорт: метро, трамвай, электротранспорт → metro, tramvay, elektr transporti → тематическая энциклопедия «Метро Ташкента»

V. AR/AI/EdTech Модули

Галереи, музеи, клубы по интересам

♦ Цифровые галереи и выставки:


- Пользователь может создавать собственные тематические галереи (фото, арт, AR-объекты)
- Виртуальные выставки с мультязычными подписями, навигацией и возможностью комментирования
- Интеграция с национальными художественными фондами, студенческими галереями, школьными творческими секциями

♦ Интерактивные музеи знаний:

- Автоматически собранные энциклопедии по темам (языки, культура, наука, спорт)
- Мультимедийные туры с голосовым сопровождением, TTS, знаковым языком
- Виртуальные реконструкции: Codex Terra, музей метро, музей узбекской письменности

♦ Клубы по интересам:

- Тематические группы (языки, IT, литература, театр, кулинария, робототехника и т.д.)
- Онлайн-занятия, голосовые чаты, видеостримы, совместные проекты
- Цифровые архивы и дневники клубной активности
- Интеграция с Discord, Telegram, Google Meet

 Эти разделы формируют **социальную образовательную инфраструктуру**, где каждый может учиться, творить, вести исследование и делиться результатами как часть глобального сообщества.

Модули сертификации, экзаменов и конкурсного отбора

♦ Международные языковые экзамены:

- Интеграция платформ TestDaF, TELC, Goethe-Institut, TOEFL, IELTS и CEFR
- Подготовка и пробное тестирование с автоматической проверкой и аналитикой
- Получение цифрового сертификата (верифицированного или обучающего уровня)

♦ Дистанционные образовательные платформы:

- Встроенный интерфейс для онлайн-школ, курсов, университетских MOOC

- Взаимодействие с национальными системами (OpenU, UzEdu, EU eLearning)
- Модули для формирования портфолио учащегося

♦ **Национальные конкурсы и олимпиады:**

- Автоматическое проведение тематических конкурсов (лучший ученик/студент месяца/года)
- Балльная система, рейтинги, онлайн-награды и публикация результатов
- Поддержка школьных, университетских и международных конкурсов по языкам, литературе, STEM и др.

7 Эти функции превращают платформу в систему не только обучения, но и **сертификации, карьерного роста и цифрового признания достижений.**

Опциональные ИИ-модули образования и лингвистической поддержки

♦ **ИИ-помощник правописания и лексики** — исправление орфографии, пунктуации, стилистики на UZ, DE, EN ♦ **Генератор школьных заданий** — автоматическое создание упражнений, сочинений, заданий по теме ♦ **Онлайн-викторины и диктанты** — генерация интерактивных викторин, диктантов по базе тезауруса и школьных стандартов ♦ **Подготовка к экзаменам** — модуль готовит пользователя к:

- Немецкому языковому экзамену (TestDaF, TELC, DSH)
- Национальным экзаменам Узбекистана (ДТМ, CEFR UZ)
- Международным тестам (IELTS, TOEFL, Goethe Zertifikat)

♦ **ИИ-репетитор по всем школьным предметам**


- Математика, история, химия, физика, биология, география, литература, право, программирование, экология
- Работа на базе открытых библиотек университетов, курсов, Khan Academy, Coursera, университетских MOOC
- Локализация под каждую страну: национальные стандарты образования (UZ, DE, EN, INT)

♦ **Интеграция в обучающие платформы**

- Встроенные модули в школьные LMS, Moodle, Google Classroom
- Генерация планов уроков, интерактивных тестов, электронных дневников

Все модули адаптированы под:

- людей с ОВЗ (поддержка жестового языка, TTS, крупный шрифт)
- мультиязычную навигацию
- мобильные и офлайн-платформы

 **Результат:** система становится образовательной инфраструктурой нового типа — открытой, адаптивной, доступной, защищённой, локализуемой и расширяемой.

Сборка системы: от протокола к модулю (инструкция IKEA)

Модуль: Лексический процессор (Uzbek/German)

Цель: Автоматически анализировать слова в тексте, определять их часть речи, лемму и морфологические признаки для последующего включения в тезаурус.

1. Компоненты (ПО и ресурсы):

- spaCy + uzbek POS модель (uzbek_pos_custom_model) + de_core_news_md
- stanza + uzbek_morphology , de_morphology
- UzTagger (на базе CRF или BiLSTM)
- корпуса: uzWac , NKMUZ , DWDS , EuroParl
- инструменты: pdfplumber , pymupdf , csvkit

2. Шаги сборки:

Шаг 1: Установка окружения

```
pip install spacy stanza pandas sklearn
python -m spacy download de_core_news_md
```

Шаг 2: Инициализация моделей

```
import stanza
stanza.download('uz')
stanza.download('de')
```

Шаг 3: Парсинг PDF/Text → JSON

```
import pdfplumber
with pdfplumber.open("sample.pdf") as pdf:
    text = ""
    ".join([page.extract_text() for page in pdf.pages])
```

Шаг 4: Обработка текста

```
import spacy
nlp_de = spacy.load("de_core_news_md")
doc = nlp_de(text)
for token in doc:
    print(token.text, token.lemma_, token.pos_)
```


Шаг 5: Форматирование в JSON




```
{
  "lemma": "entwickeln",
  "pos": "VERB",
  "morph": "Infinitive",
  "translation": "rivojlantirmoq",
  "domain": "ICT"
}
```

3. Тестирование:

- Запусти на файле `sample_uzbek_german.txt`
- Сравни частоты и леммы с DWDS и uzWac
- Проверь сохранение в `data/thesaurus/entries.json`

4. Связь с тезаурусом: Результат парсера поступает в модуль построения гнёзд и автоматически группируется по лемме, частоте и семантике.

 Результат: модуль может быть встроен в любой тезаурус, CMS, переводчик или онлайн-курс.

-  **AR-лексика:** визуализация предметов с переводом
- **Обратная связь:** пользователь может добавлять примеры, голосовать
-  **EdTech:** упражнения, квизы, викторины, мультимедийные энциклопедии
-  **Codex Terra:** лингвистическая капсула памяти, защита культурного наследия

VI. Научная и правовая база

- Диссертации (UZ/DE): структура, методология, цифровая реализация
- Журналы: фиксация этапов, JSON/корпус/гнёзда
- Авторские декларации (июль 2025): охрана авторских прав и приоритета
- Codex Terra Defense Kit: философия, гуманитарный манифест, FAQ

VII. Вывод

Создание мультязычного тезауруса — ключевой вклад в цифровую лексикографию, межкультурную коммуникацию и языковое наследие. Платформа охватывает AI, AR, EdTech, лингвистику и прикладные дисциплины. Это не просто словарь — это семантическая инфраструктура знаний Центральной Азии.

VIII. Список литературы

1. Abdurakhmonova, N., & Shamieva, N. (2024). Creating an English-Uzbek Bilingual Thesaurus of Frequently Used Adjectives in Uzbek Corpus. *IEEE Proceedings*. <https://doi.org/10.1109/piere62470.2024.10804960>
2. Abdurakhmonova, N., Shamieva, N., & Adali, E. (2024). Exploring the Semantic Complexity of Adjective-Noun Collocations between Uzbek and English. *UBMK Conference*. <https://doi.org/10.1109/ubmk63289.2024.10773511>

3. Tilovova, G. (2024). Roadmap of Creating a Bilingual Uzbek-German Electronic Dictionary of Forestry Terms. *Journal of Interdisciplinary Studies in Education*. <https://doi.org/10.32674/3c15dj73>
4. Matlatipov, S., Tukeyev, U., & Aripov, M. (2020). Towards the Uzbek Language Endings as a Language Resource. *Springer: Computational Collective Intelligence*. https://doi.org/10.1007/978-3-030-63119-2_59
5. Matlatipov, S., Aripov, M., & Abdurakhmonova, N. (2018). Modeling WordNet Type Thesaurus for Uzbek Language Semantic Dictionary. *International Journal of Systems Engineering*. <https://doi.org/10.11648/J.IJSE.20180201.16>
6. Dahal, C. (2024). Revolutionizing Education through AI-Powered Inclusive Learning Systems. *AAAI*. <https://doi.org/10.1609/aaai.v38i21.30546>
7. Vijayalakshmi, B., Deshpande, K.A., & Krishan, N. (2024). Use of AI to Augment Multilingual Content in Cyberspace for Development – An Indian Case Study. *GHTC IEEE*. <https://doi.org/10.1109/ghtc62424.2024.10771552>
8. Msweli, Z.P., & Ajani, O.A. (2024). Enhancing Digital Inclusion in Multilingual Education: Role of Technology in South African HE. *IJITSS*. [44 .2024.3134">https://doi.org/10.31435/ijitss.4\(44\).2024.3134](https://doi.org/10.31435/ijitss.4(44).2024.3134)
9. Byakodi, R. (2024). Use of AI in India and International Education in Developmental Countries. *IJRASET*. <https://doi.org/10.22214/ijraset.2024.64252>
10. Barausse, E. (2023). Promoting Digital Inclusion with AI: Reaching Marginalized Communities. *OSF Preprints*. <https://doi.org/10.31219/osf.io/c9qdx>