

# 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 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](http://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](http://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

-  **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/гнезда
- 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](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>