### 📄 Журнал разработки тезауруса и лексикографического ядра (Uzbek-German Thesaurus)

#### 📅 Начало: 2025-07-04

#### 👤 Руководитель: ABDUKARIMOV ABDURASHID ABDULXAMITOVICH

## 🧱 Этап I: Подготовка и импорт источников

### Используемые файлы (основные словари):

* O'zbek tilining izohli lug'ati. 1-3 jild
* Etimologik lug'at (Rahmatullayev)
* Tarix terminlari lug'ati (Bekmuhamedov)
* O'zbekcha-nemischa o'quv lug'ati
* Turkcha-ozbekcha va ozbekcha-turkcha lug'at
* Phrasebook (Awde, Hikmatullaeva)

### Используемые корпуса:

* [uzbekcorpus.uz](https://uzbekcorpus.uz)
* [uzWac](https://www.sketchengine.eu/uzwac-uzbek-corpus/)
* [Leipzig Corpora](https://corpora.uni-leipzig.de/en?corpusId=uzb_community_2017)

## ⚙️ Этап II: Построение ядра лемм и дефиниций

**Инструменты:**

* pdfplumber для извлечения текста из PDF
* re (regex) для поиска шаблонов: lemma — definition
* json для сериализации данных

**Формат вывода:**

{  
 "lemma": "ilm",  
 "definition": "bilim, ilm-fan",  
 "source": "izohli lug'at, 1-jild"  
}

### Проблемы и замечания:

* Многостраничные PDF-файлы могут вызывать time-out.
* Часть страниц плохо извлекается (шум OCR, дефекты шрифта).
* Некоторые словари не используют “—”, приходится варьировать шаблон.

### Рекомендации:

* Автоматически сохранять JSON-логи постранично.
* Добавить поддержку UTF8-гаммы с нестандартными апострофами: ‘ʼʻ

## 🧠 Этап III: Международный опыт и стандарты

### Используемые практики:

* **Langenscheidt/Duden/Oxford**: форматы карточек, коллокации, частота
* **Multitran/Systran/DeepL**: вариативность перевода, API-интеграции
* **TshwaneLex/Shoebox**: формат TEI/XML + JSON для хранения
* **SketchEngine**: извлечение частот, коллокаций, контекста
* **Text+ (IDS Mannheim)**: инфраструктура FAIR и автоматизация

## 🔧 Этап IV: Подключение к редакторам и CAT-инструментам

### Модули:

* Word Add-In / OfficeJS: выделение → перевод → замена / комментарий
* PDF Annotator: обработка примечаний, синхронизация с тезаурусом
* CAT Tool Export: TBX / TMX / XLIFF поддержка (на базе JSON)

## 🗂 Планы на следующие шаги:

📌 Лог будет дополняться по мере каждого действия и анализа. Все скрипты и их выводы будут прикрепляться по этапам.