ЭТАП 3: Создание SQLite индекса словоформ - ЗАВЕРШЕН

Обзор выполненных задач

✅ **Создание структуры SQLite базы данных**
- Создан файл `index.sqlite` (3.3 MB)
- Таблица `wordforms` с полной структурой согласно спецификации
- 4 индекса для быстрого поиска: по словоформе, лемме, части речи и частотности

✅ **Обработка морфологических данных**
- Обработано 29,627 токенов из `morphology_data.jsonl`
- Проиндексировано 8,433 уникальных словоформ (превышает ожидаемые 8,062)
- Определены основные леммы и POS-теги для каждой словоформы
- Собраны морфологические признаки с подсчетом частотности

✅ **Создание индекса с примерами предложений**
- Каждая словоформа содержит от 1 до 5 примеров предложений
- Среднее количество примеров на словоформу: 1.50
- Реализован алгоритм выбора разнообразных примеров
- Фильтрация технических и слишком коротких предложений

✅ **Оптимизация базы данных**
- Созданы 4 индекса для быстрого поиска
- Выполнены операции VACUUM и ANALYZE для оптимизации
- Использован режим WAL для лучшей производительности
- Корректная обработка UTF-8 и словацких диакритических знаков

✅ **Создание дополнительного JSONL экспорта**
- Файл `index_export.jsonl` (3.3 MB) с тем же содержимым
- 8,433 записи в формате JSON
- Сохранены все диакритические знаки

✅ **Статистика и отчетность**
- Файл `index_stats.json` (31 KB) с подробной статистикой
- 464 уникальных морфологических признака
- Распределение по частям речи

Ключевые результаты

Статистика индекса:

- **Всего словоформ**: 8,433
- **Время создания**: 1.09 секунд
- **Средние примеры на словоформу**: 1.50
- **Уникальных морфологических признаков**: 464

Распределение по частям речи:

1. **NOUN** (существительные): 3,513 (41.7%)
2. **VERB** (глаголы): 2,019 (23.9%)
3. **ADJ** (прилагательные): 1,280 (15.2%)
4. **ADV** (наречия): 513 (6.1%)
5. **PROPN** (имена собственные): 212 (2.5%)
6. **DET** (определители): 180 (2.1%)
7. **PRON** (местоимения): 151 (1.8%)
8. **PART** (частицы): 132 (1.6%)
9. **ADP** (предлоги): 106 (1.3%)
10. **NUM** (числительные): 93 (1.1%)

Топ-10 самых частых словоформ:

1. `,` (запятая) - 2,762 раза
2. `=` (символ равенства) - 2,750 раз
3. `.` (точка) - 1,267 раз
4. `:` (двоеточие) - 592 раза
5. `v` (предлог “в”) - 491 раз
6. `sa` (местоимение “себя”) - 377 раз
7. `!` (восклицательный знак) - 330 раз
8. `;` (точка с запятой) - 328 раз
9. `–` (тире) - 281 раз
10. `na` (предлог “на”) - 254 раза

Качество данных

✅ Сохранение диакритических знаков

Корректно сохранены все словацкие диакритические знаки:
- `áäčďéíĺľňóôŕšťúýž`
- Примеры: `Sándor`, `Agneška`, `Akým`

✅ Примеры предложений

Каждая словоформа содержит реальные примеры использования:


```
{
  "wordform": "Je",
  "lemma": "byť",
  "upos": "VERB",
  "feats": "Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|Verb-
Form=Fin",
  "frequency": 4,
  "sentences": [
    "Je, vidím, cítim, celok života!",
    "Je šťastný, môž'-li hrdou lýrou svit jej slávy velebiť..."
  ]
}
```

✅ Производительность

- Быстрый поиск благодаря индексам
- Компактный размер базы данных (3.3 MB)
- Оптимизированная структура

Созданные файлы

1. `index.sqlite` (3.3 MB) - основная SQLite база данных с индексом
2. `index_export.jsonl` (3.3 MB) - JSONL экспорт для удобства использования
3. `index_stats.json` (31 KB) - подробная статистика по индексу
4. `index_builder.py` - скрипт для создания индекса (воспроизводимость)

Готовность к использованию

Индекс полностью готов для использования в корректоре:
- ✅ Быстрый поиск по любой словоформе
- ✅ Примеры реального использования в предложениях
- ✅ Полная морфологическая информация
- ✅ Корректная обработка словацкого языка
- ✅ Оптимизированная производительность

Следующие шаги

Этап 3 успешно завершен. Индекс готов для:
1. Интеграции в корректор орфографии
2. Упаковки в финальный архив
3. Использования для проверки и исправления текстов

Статус: ✅ ЗАВЕРШЕН УСПЕШНО
Дата: 30 августа 2025
Время выполнения: 1.09 секунд