Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления

Кафедра Интеллектуальных информационных технологий

**ОТЧЁТ**

**По лабораторной работе № 3**

по дисциплине «Естественно-языковой интерфейс интеллектуальных систем»

по теме: “Синтаксический анализ текстов естественного языка”

Выполнил: Самута Д.В.

Семенов Е. Г.

гр. 221703

Проверил: Крапивин Ю. Б.

Минск 2025

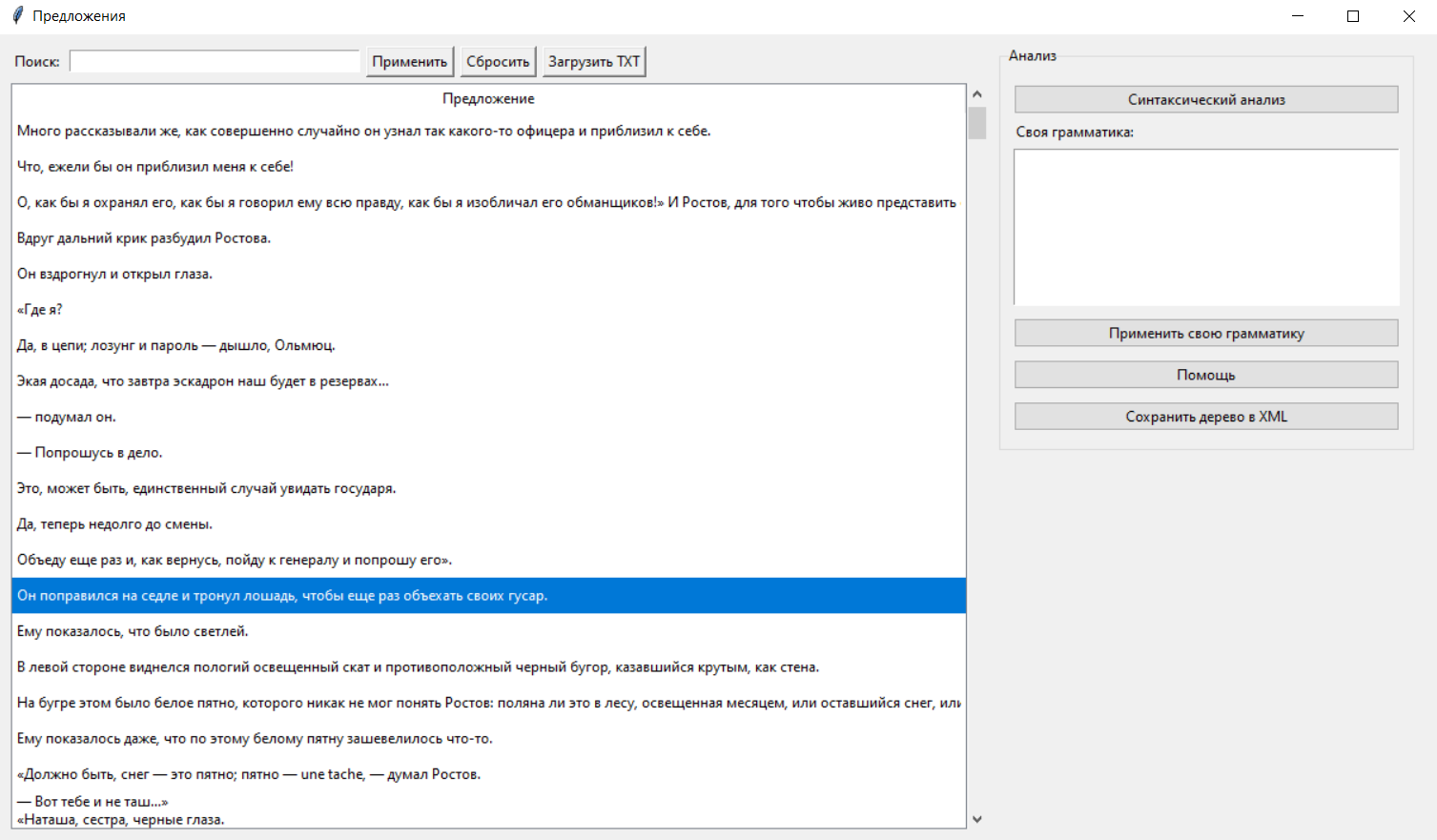
**Цель работы:** освоить принципы разработки прикладных сервисных программ для решения задачи анализа текста естественного языка.

**Задание:**

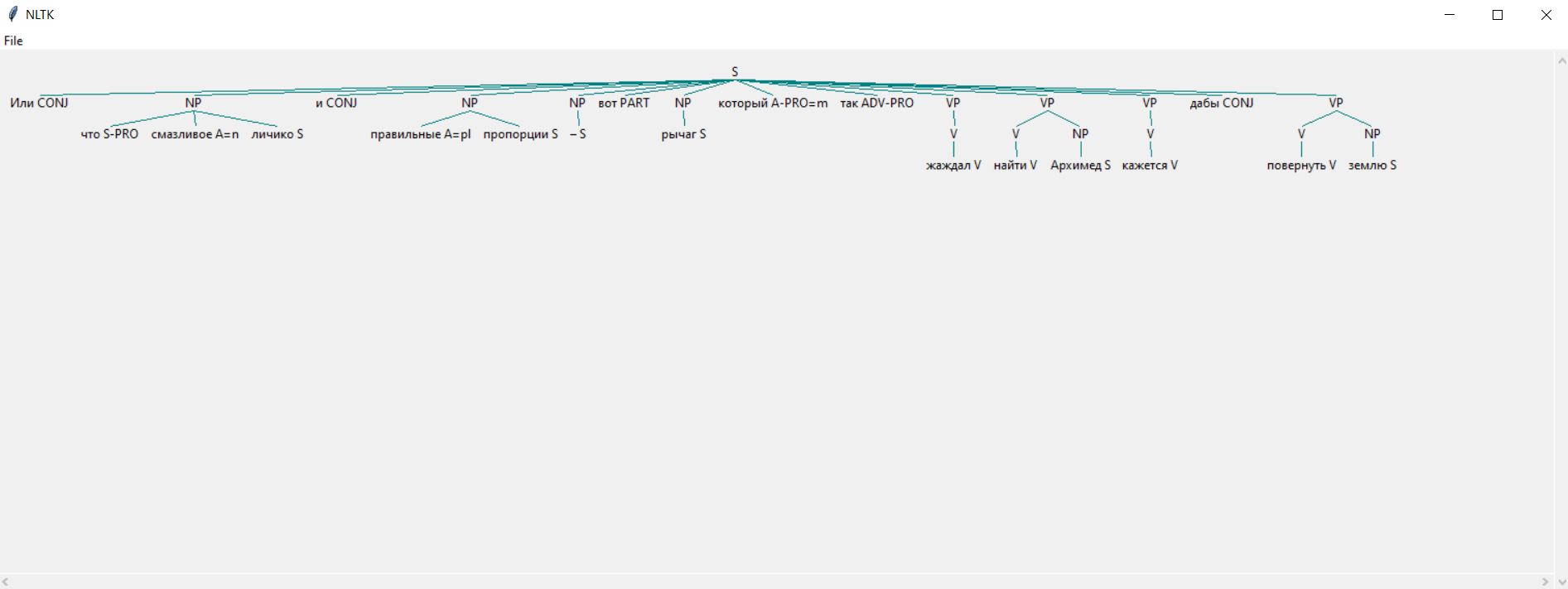
1. Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического синтаксического анализа ТЕЯ.

2. Закрепить навыки программирования при решении задач автоматической обработки ТЕЯ

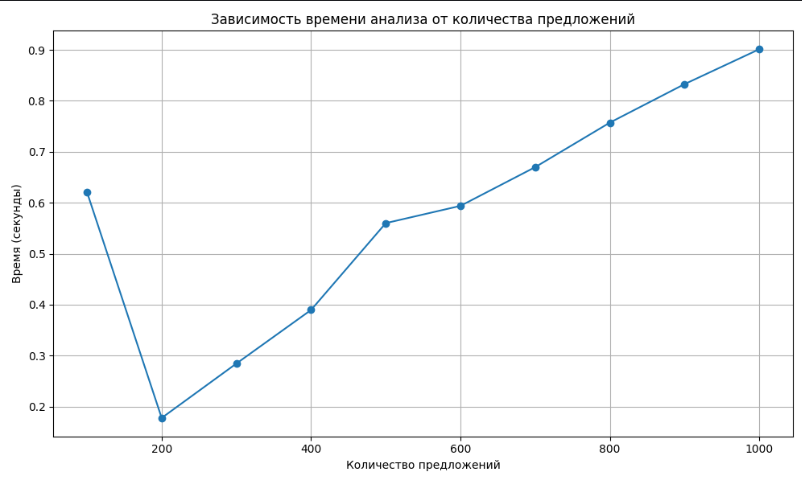
**Интерфейс**



**Интерфейс дерева разбора предложения**

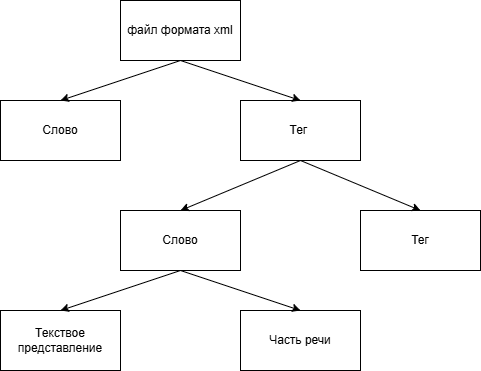


**Тест производительности:**

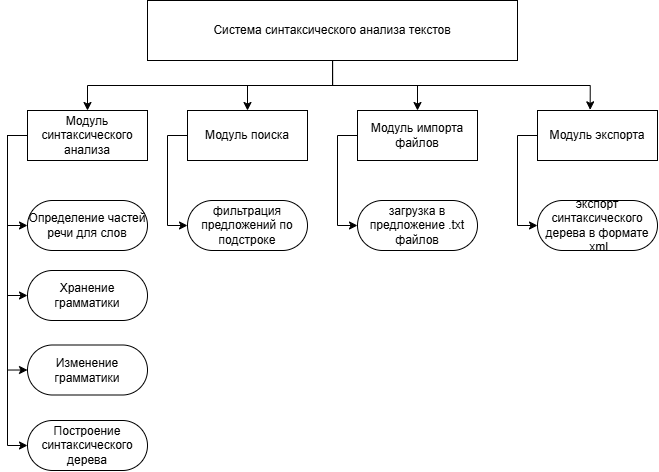


На данном графике показана зависимость времени анализа предложений от кол-ва предложений. Видно, что с увеличением кол-ва предложений время обработки также возрастает. Это указывает на **линейную зависимость** между количеством предложений и временем обработки, хотя есть небольшие отклонения, например в начале предложения, связана она с тем, что для анализа прдедложений использовался nltk, который в начале может медленнее работать, т.к. в процессе своей работы он накапливает контекст. Общая тенденция — увеличение времени обработки при увеличении кол-ва предложений.

**Структура хранения данных**

****

**Структурная схема приложения:**

****

**Использованные библиотеки:**

**json** — библиотека для работы с форматом JSON.

**re** — библиотека для обработки текста с использованием регулярных выражений.

**string** — библиотека, предоставляющая строковые константы и утилиты.

**tkinter** — библиотека для создания графических пользовательских интерфейсов (GUI).

**time** — библиотека для работы со временем и задержками.

**unittest.mock** — библиотека для создания мок-объектов в модульном тестировании.

**xml.etree.ElementTree** — библиотека для работы с XML-документами.

**nltk** — библиотека для обработки и анализа естественного языка (Natural Language Processing).

**Описание алгоритмов:**

**1. Алгоритм построения списка предложений для текста**

1. **Начало.**
2. **Выбор файла пользователем:**
   * Через диалоговое окно пользователь выбирает текстовый файл.
3. **Очистка текста (при необходимости):**
   * Удаляются лишние пробелы, управляющие символы, заменяются переносы строк.
4. **Разбиение текста на предложения:**
   * Регулярное выражение, например re.split(r'(?<=[.!?])\s)+', text).
5. **Формирование списка предложений:**
   * Каждое предложение сохраняется как отдельный элемент списка.
6. **Конец.**

**2. Алгоритм формирования дерева синтаксического анализа предложения:**

1. Начало.
2. Выбор предложения для анализа:
   * Из исходного текста извлекаются предложения.
   * Пользователь выбирает одно предложение для анализа.
3. Предобработка текста:
   * Выполняется токенизация предложения (разделение на слова).
   * Из предложения удаляются знаки препинания.
4. Частеречная разметка:
   * Каждое слово размечается по частям речи с использованием библиотеки NLTK.
   * Разметка проводится с использованием русскоязычной модели.
5. Определение синтаксических структур:
   * С помощью регулярного грамматического шаблона задаются правила выделения фраз (NP, VP, PP и др.).
   * На основе этих правил строится синтаксическое дерево (разбор по шаблонам).
6. Визуализация дерева:
   * Создаётся графическое окно с холстом.
   * На холст выводится древовидная структура предложения, отражающая иерархию синтаксических компонентов.
7. Сохранение результата:
   * Структура дерева сохраняется в XML-файл, где ключом является предложение, а значением — дерево разбора.
8. Завершение работы:
   * После завершения анализа возможно закрытие окна или выбор нового предложения.
9. Конец.

**Вывод:** в ходе выполнения лабораторной работы был реализован синтаксический анализ текста с использованием библиотеки NLTK для русского языка, включающий этапы токенизации, частеречной разметки и построения синтаксического дерева. На основе предложений текста были построены деревья, отображающие их синтаксическую структуру, что позволило наглядно продемонстрировать именные и глагольные фразы, а также другие синтаксические компоненты. Визуализация результатов анализа обеспечила понимание структуры предложений, что является основой для дальнейшей работы с текстовыми данными.