Министерство образования Республики Беларусь

Учреждение образования “Белорусский государственный университет информатики и радиоэлектроники”

Факультет информационных технологий и управления

Кафедра интеллектуальных информационных технологий

Отчёт по дисциплине

“Естественно-языковой интерфейс интеллектуальных систем”

Лабораторная работа №1

“Разработка автоматизированной системы формирования словаря естественного языка”

Выполнили студенты группы 221702: Целуйко Д.А.

Юргилевич Е.В.

Проверил: Крапивин Ю.Б.

Минск 2024

Лабораторная работа №1

**Цель работы:**

Освоить принципы разработки прикладных сервисных программ для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка.

**Задание (№1, вариант 13):**

Список слов, упорядоченный по алфавиту и включающий как

лексемы, так и словоформы, с указанием частоты встречаемости каждой из

форм. Для словоформ пользователю должна быть предоставлена возможность вводить дополнительную морфологическую информацию, а именно, отнесение слова к соответствующей части речи, указание рода, числа, падежа и т.п. При этом морфологическая информация может быть оформлена как отдельная неформатированная запись, т.е. это просто текст, который пользователь может оформлять произвольным образом.

**Язык текста:**

Английский

**Формат входного документа:**

RTF, TXT.

**Используемые средства разработки:**

Для решения поставленной задачи было разработано desktop-приложение на языке программирования **Python** с использованием библиотека **tkinter**, а также некоторые другие библиотеки

**Tkinter —** это пакет для Python, предназначенный для работы с библиотекой Tk. Библиотека Tk содержит компоненты графического интерфейса пользователя (graphical user interface – GUI). Эта библиотека написана на языке программирования Tcl.

**Spacy (spaCy)** — это программная библиотека с открытым исходным кодом для расширенной обработки естественного языка, написанная на языках программирования Python и Cython. Были догружены следующие ресурсы:

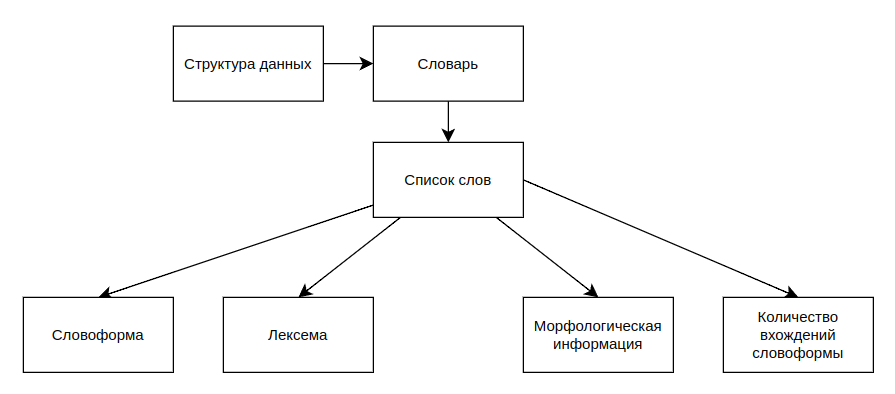
* **spacy.load('en\_core\_web\_sm') -** легкая английская языковая модель, оптимизированная для использования центральным процессором. Эта модель позволяет выполнять присвоение частей речи, синтаксический анализ зависимостей и лемматизацию токенов во входном тексте. Разработано компанией Explosion AI.

**Json** — используется в качестве стандартного текстового формата для хранения и извлечения структурированных данных, таких как информация об анализируемых словах и их морфологические данные.

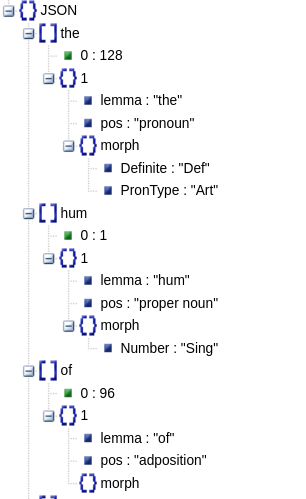
**Используемые структуры для хранения данных**Проанализированный словарь естественного языка хранился в словаре Python, встроенной структуре данных Python, представляющей собой не упорядоченную коллекцию данных. Этот словарь хранил информацию о словах, включая количество вхождений, лексему и морфологическую информацию. Данный словарь сохраняется в файл формата JSON.

Сам словарь естественного языка представляет собой встроенную структуру данных языка Python – **словарь** – неупорядоченную коллекцию данных. В данном словаре содержится информация о словоформе, его лексеме, его морфологической информации и его частоте вхождений.  
Система работает со следующими частями речи:

* ADJ(прилагательное)
* ADP(предлог)
* ADV(наречие)
* AUX(вспомогательный глагол)
* CCONJ(координационная связка)
* DET(детерминатив)
* INTJ(междометие)
* NOUN(существительное)
* NUM(числительное)
* PART(частица)
* PRON(местоимение)
* PROPN(имя собственное)
* PUNCT(пунктуация)
* SCONJ(подчинительная связь)
* SYM(символ)
* VERB(глагол)



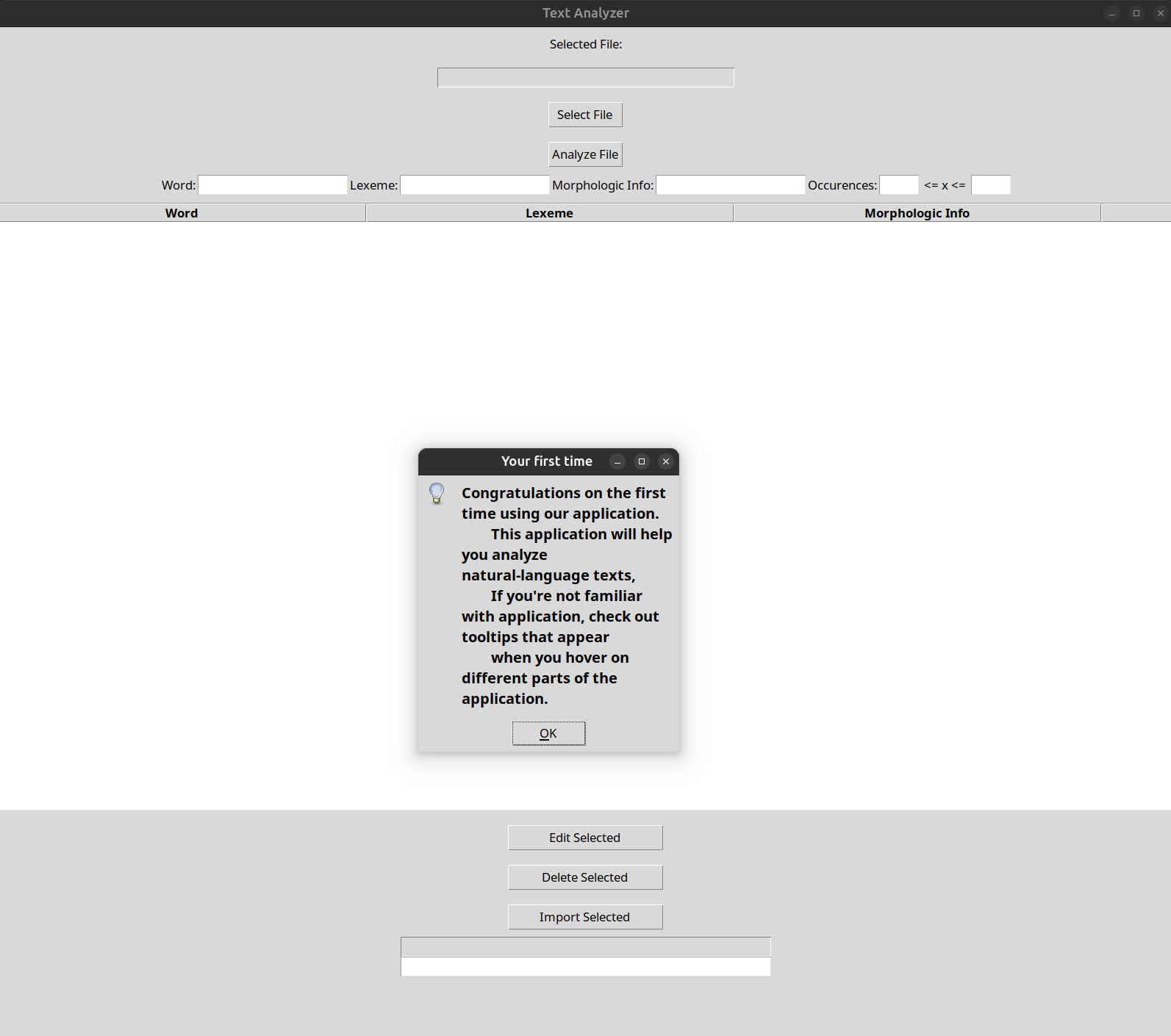
*Рис. 1 – структура хранения данных*



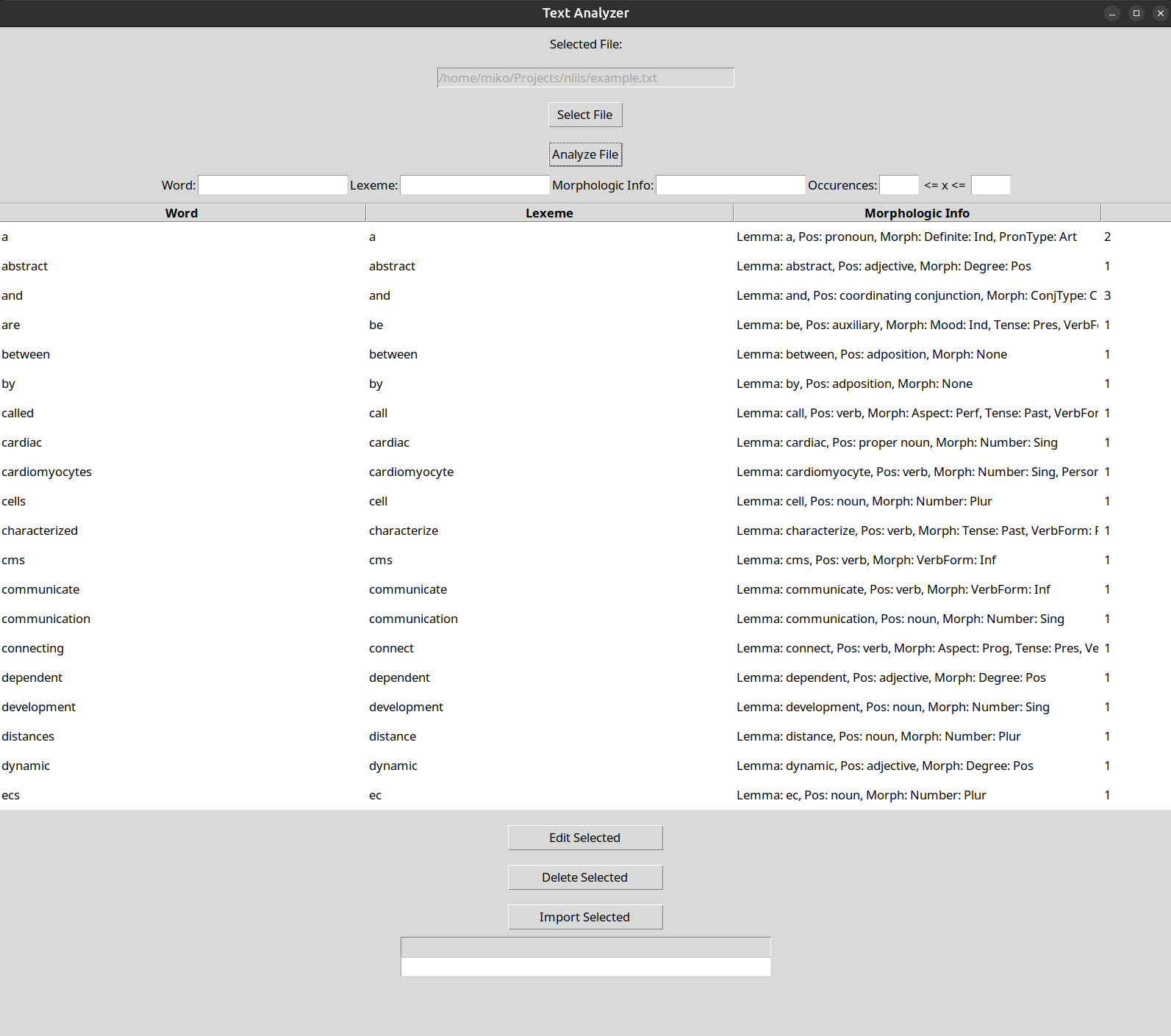
*Рис. 2 – пример хранения данных в структуре хранения*

**Скриншоты разработанной системы:**

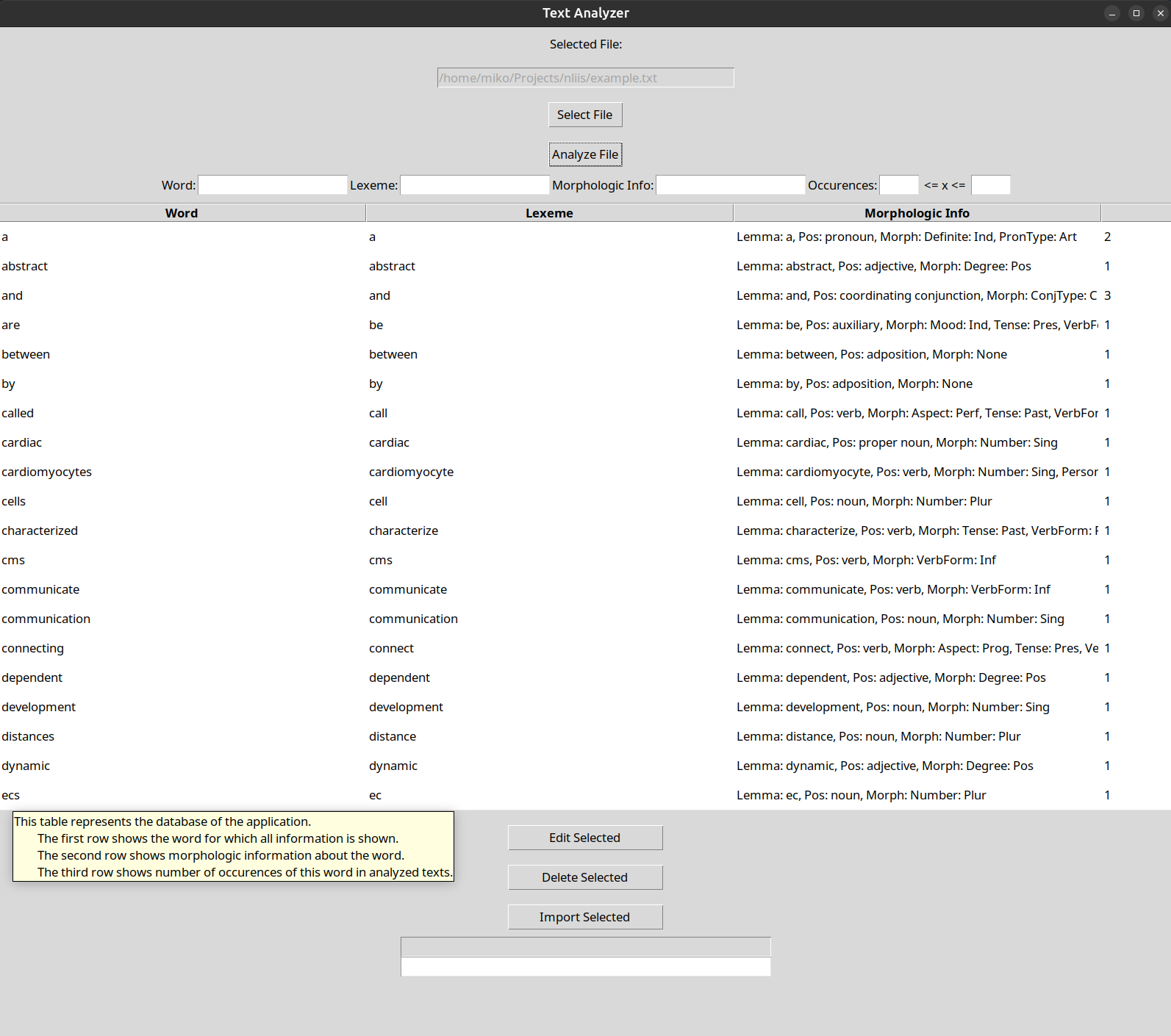
**Начальная страница:**

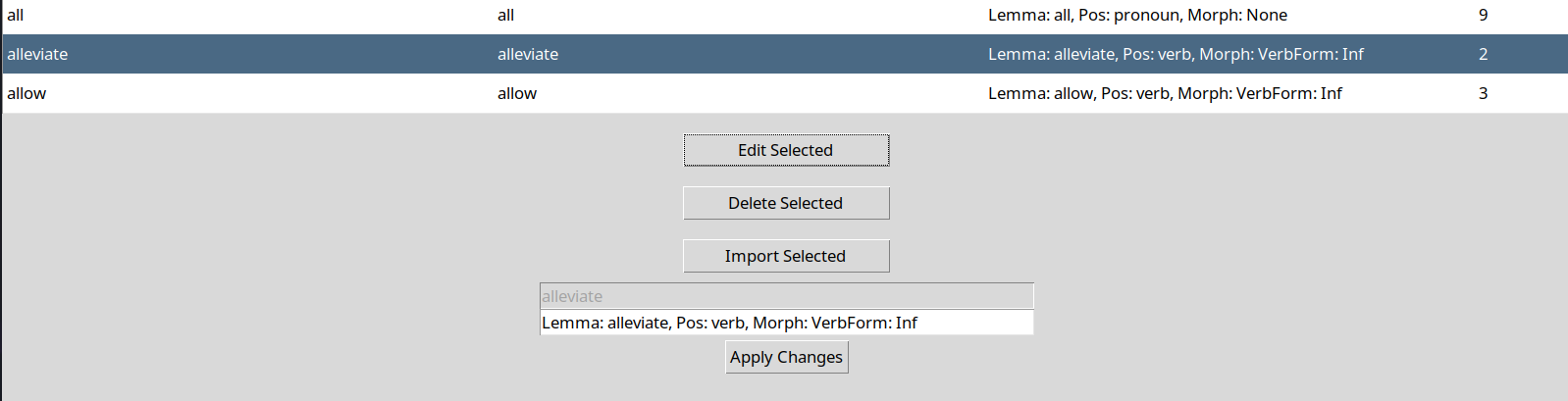
****

*Рис. 3 – начальная страница*



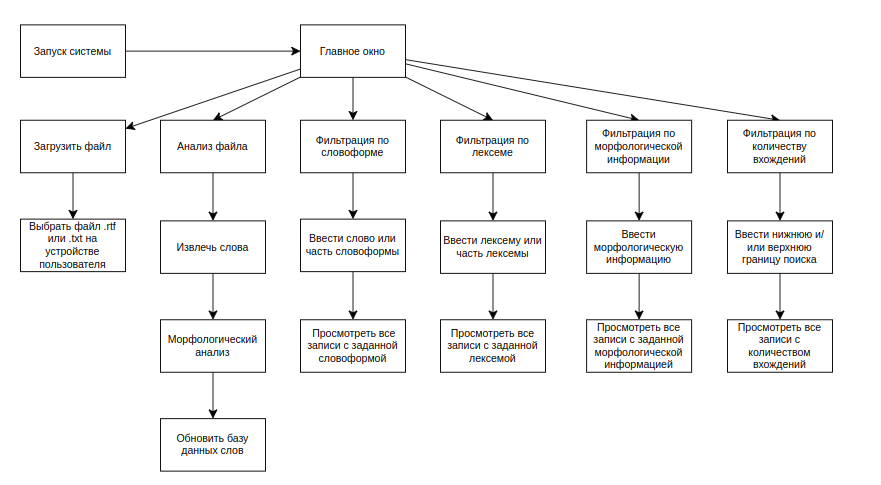
*Рис. 4 – страница со всеми словоформами и их информацией, найденными в файле example.txt*

****

*Рис. 5 – окно с помощью к таблице*

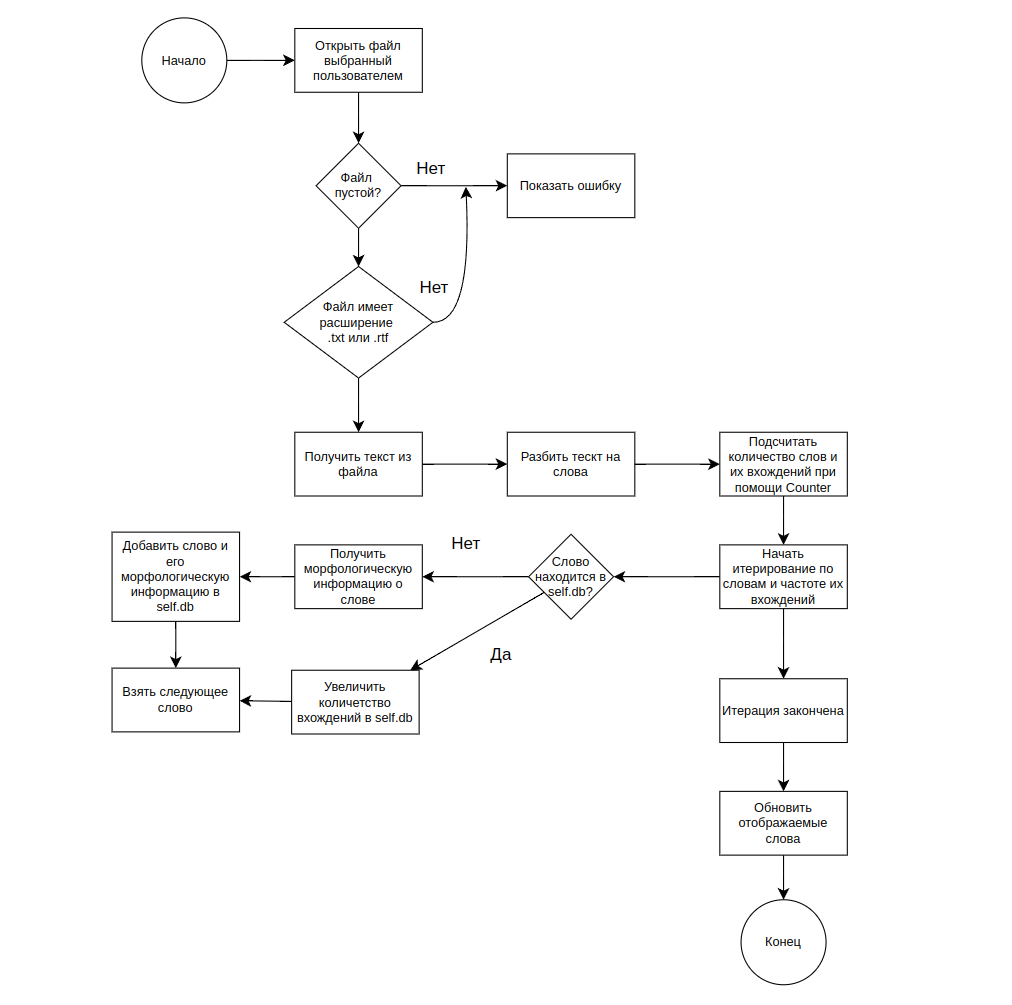
*Рис. 6 – просмотр информации о словоформе с возможностью редактирования*

**Структурно-функциональная схема приложения:**



*Рис. 7 - структурно-функциональная схема предложения*

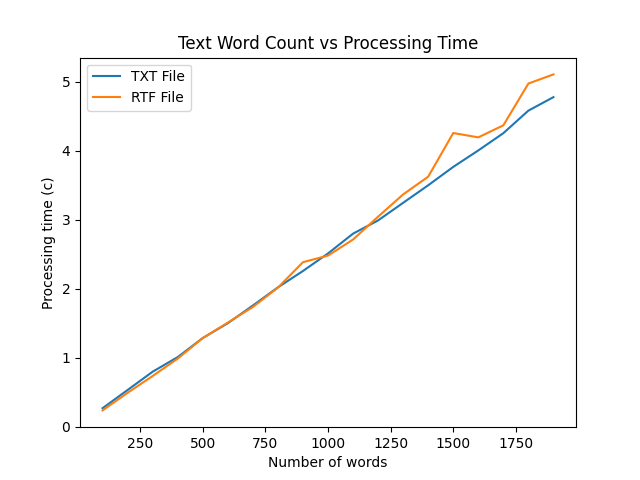
**Структурные схемы правил, используемых в системе:**



*Рис. 8 - структурная схема правил: функция анализа содержимого файла*

**Тестирование скорости программы:**

Для тестирования быстродействия приложения были использованы файлы формата .txt и .rtf, содержащие текст, размер которого составляет 100 до 2000 слов с шагом в 100.



*Рис. 9 - график зависимости времени обработки от количества слов в анализируемом тексте для различных форматов текста*

**Выводы по работе и по перспективам развития приложения:**

Программа представляет собой инструмент для анализа текстовых файлов с целью извлечения уникальных слов, определения их морфологических характеристик и подсчета частоты встречаемости. Она предоставляет пользовательский интерфейс для фильтрации и редактирования полученных данных, а также возможность экспорта информации о конкретных словах.