Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

Кафедра интеллектуальных информационных технологий

**Отчет по лабораторной работе №4**

**по курсу «ЕЯзИИС»**

**на тему: «Разработка системы машинного перевода документов»**

|  |  |
| --- | --- |
| Выполнили  студенты гр.921703: | Губаревич И.А.  Кравцов М.С.  Семякин А.В.  Новицкий С.В. |
| Проверил: | Крапивин Ю.Б. |

**МИНСК**

**2022**

**Цель работы**:

освоить на практике основные принципы машинного перевода документов.

**Условие:** 

**Информация о тестовой коллекции документов**

Была составлена тестовая коллекция документа для английского языка.

### Пример тестового документа на английском языке:

### Имя файла: text.txt

### Тип файла: текстовый документ

### Размер: 1КБ

### Содержание:

### 

**Описание структур данных, использованных для хранения входной и выходной информации**

### Входная информация:

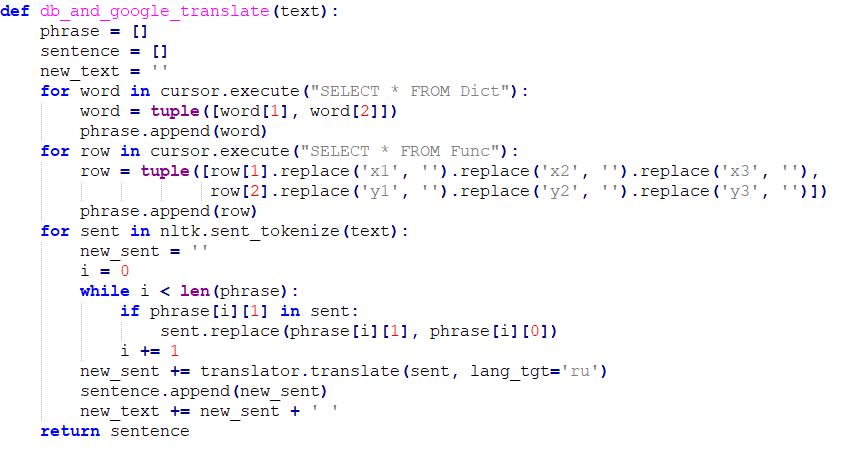
Текстовый документ, на английском языке.

### Выходная информация:

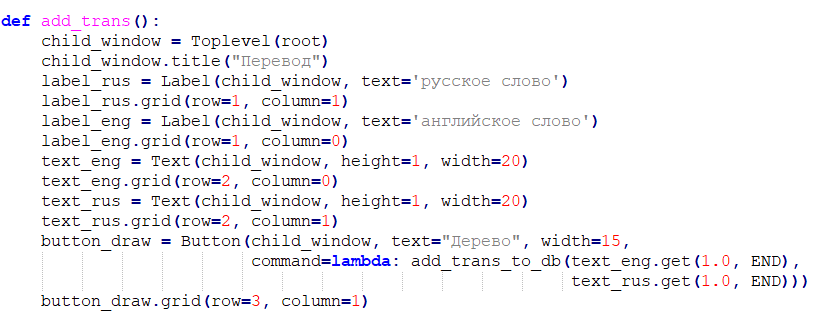
Информация выводится на экран (пронумерованный по предложениям переведенный текст), также при желании, пользователь может построить дерево, дополнить базу данных и сохранить результат базы данных в файл translate.txt.

**Описание алгоритмов:**

Перевод текст посредством **Googletrans** и добавление его в DB(sqlite):



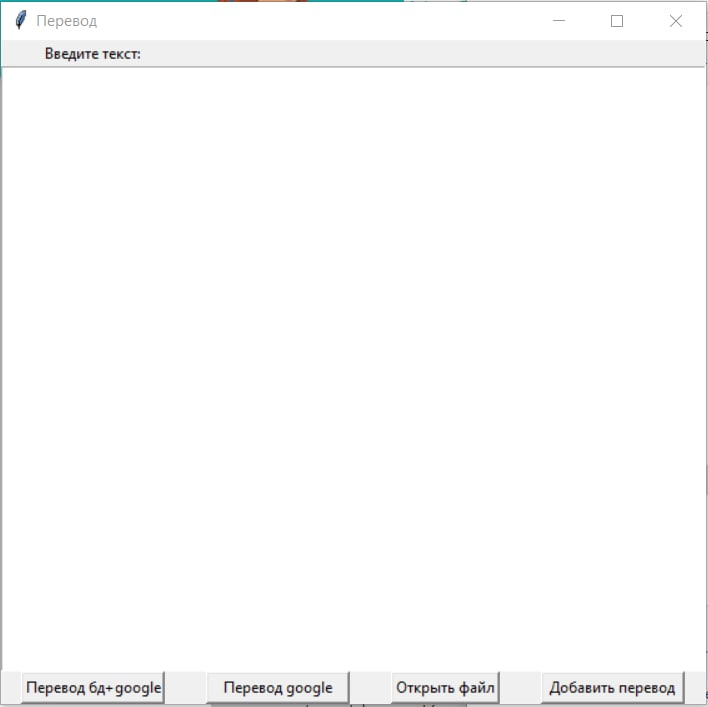
Добавление слова в словарь:



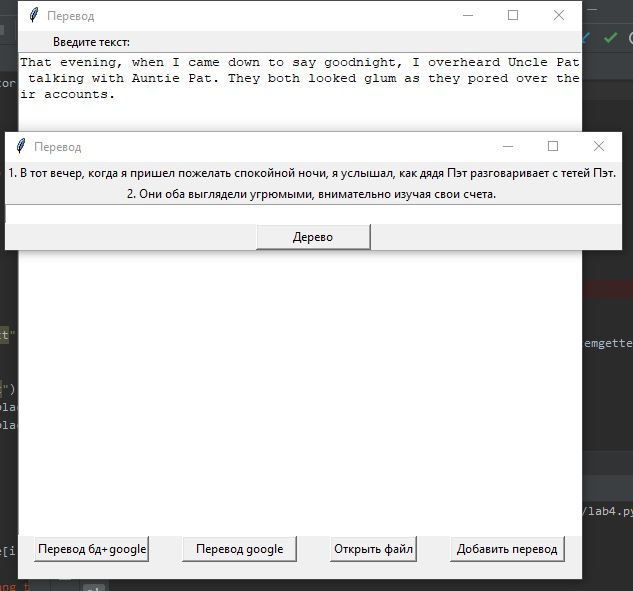
**Результаты тестирования системы**

Текст проверяется на наличие слов/выражений из БД межъязыковых соответствий лексических. Найденные в БД слова и/или выражения переводятся по словарю. Остальной текст переводится средствами систем машинного перевода, используя API Google Translate

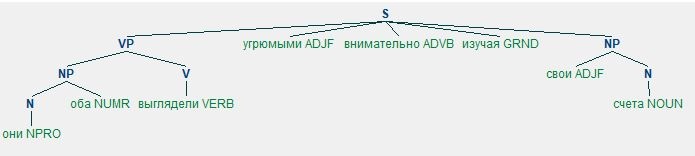
При нажатии на кнопку «Открыть файл», приложение предлагает выбрать исходный файл.



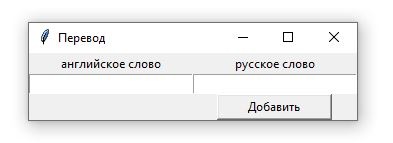
Также есть возможность ввести текст вручную, без считывания из файла. В приложении создан функционал перевода текста только используя API Google Translate либо используя базу данных + API Google Translate. Далее в новом окне переведенные предложения нумеруются и можно ввести номер для построения дерева предложения.



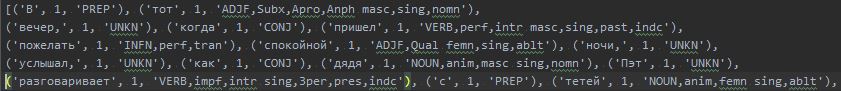
Пример построения дерева для второго предложения:



Также есть функционал для добавления слов в базу данных.



Пример части сформированной базы данных:



По итогу создастся файл с базой данных translate.txt с выше указанным фрагментом базы данных.

**Описание** **и особенности применения готовых к использованию компонент**

1. **NLTK** (Natural Language Toolkit) – ведущая платформа для создания NLP-программ на **Python**. У нее есть легкие в использовании интерфейсы для многих языковых корпусов, а также библиотеки для обработки текстов для классификации, токенизации, стемминга, разметки, фильтрации и семантических рассуждений.
2. **Tkinter** – это кроссплатформенная библиотека для разработки графического интерфейса на языке **Python** (начиная с **Python** 3.0 переименована в **tkinter**). **Tkinter** расшифровывается как Tk interface, и является интерфейсом к Tcl/Tk.
3. **SQLite** – это C библиотека, реализующая легковесную дисковую базу данных (БД), не требующую отдельного серверного процесса и позволяющую получить доступ к БД с использованием языка запросов SQL.
4. **Googletrans** - это бесплатная и неограниченная библиотека **Python**, в которой реализован API Google Translate. Он использует API Google Translate Ajax для вызова методов обнаружения и перевода.
5. **Pymorphy2 –** это морфологический анализатор для русского языка, написанный на языке Python и использующий словари из OpenCorpora. Он работает под Python 2.6, 2.7, 3.2 и 3.3 и распространяется по лицензии MIT.

**Вывод:**

В данной лабораторной работе мы освоили на практике основные принципы машинного перевода документов. Наш текст проверяется на наличие слов/выражений из БД межъязыковых соответствий лексических единиц в соответствии с вариантом. Найденные в БД слова и/или выражения переводятся по словарю, остальной текст переводится средствами систем машинного перевода.

Машинный перевод [художественных текстов](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5_%D1%81%D1%82%D0%B8%D0%BB%D0%B8_%D1%80%D0%B5%D1%87%D0%B8#%D0%A5%D1%83%D0%B4%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D1%81%D1%82%D0%B8%D0%BB%D1%8C) практически всегда оказывается неудовлетворительного качества. Тем не менее для технических документов при наличии специализированных машинных словарей и некоторой настройке системы на особенности того или иного типа текстов возможно получение перевода приемлемого качества, который нуждается лишь в небольшой редакторской корректировке.Чем более [формализован](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F) стиль исходного документа, тем большего качества перевода можно ожидать. Самых лучших результатов при использовании машинного перевода можно достичь для текстов, написанных в техническом (различные описания и руководства) и [официально-деловом стиле](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5_%D1%81%D1%82%D0%B8%D0%BB%D0%B8_%D1%80%D0%B5%D1%87%D0%B8#%D0%9E%D1%84%D0%B8%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE-%D0%B4%D0%B5%D0%BB%D0%BE%D0%B2%D0%BE%D0%B9_%D1%81%D1%82%D0%B8%D0%BB%D1%8C).