Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

Кафедра интеллектуальных информационных технологий

**Отчет по лабораторной работе №4**

**по курсу «Естественно-языковой интерфейс интеллектуальных систем»**

**на тему: “Разработка системы машинного перевода документов”**

|  |  |
| --- | --- |
| Выполнили студенты группы 721702: | Гурбович А.И. Каленик П.Р. |
| Проверил: | Крапивин Ю.Б. |

**МИНСК**  
**2020**

**Цель работы:**  
Освоить на практике основные принципы машинного перевода документов.

**Условие:**  
Вариант: 8. Направление перевода: Англо-русский. Научные статьи по медицине, Критика предметов изобразительного искусства. Используемое СМП: MyMemory API.

ссылка на репозиторий с кодом и тестовыми файлами <https://github.com/Pazhiloipavuk/eyaziisLab4>

**Информация о тестовой коллекции документов:**  
Тестовые документы сформированы на английском и русском языках.

Пример тестового документа на русском языке:

* Имя файла: art.txt
* Тип файла: текстовый документ
* Размер: 4КБ
* Содержание:

Художественная критика – это субдисциплина (часть) искусствознания, а также вид литературного творчества, оценка и истолкование произведений изобразительного искусства, явлений жизни в них отраженных, а также обстоятельств художественного процесса и художественного быта.

Художественная критика рассматривается как одна из областей искусствоведения наряду с теорией и историей искусства. В отличие от истории искусства, художественная критика освещает по преимуществу процессы происходящие в современной художественной жизни, рассматривает современную продукцию изобразительного искусства. Однако не столько материал, сколько подход, продиктованный актуальными общественными задачами и живым восприятием публики, отличает художественную критику от истории изобразительного искусства. Даже когда речь идет о классическом наследии, вводимом в культурный оборот современности, можно и нужно вести речь о художественной критике. Скажем, рецензия на выставку в Государственном Эрмитаже будет явлением художественной критики, а не истории искусства.

Критика [от лат. critikē] – искусство разбирать, судить. Критические суждения возникли одновременно с рождением искусства, первоначально как мнения наиболее уважаемых и умудренных зрителей. В эпоху античности художественная критика выделяется в отрасль литературного труда. Чрезвычайно ценные описания некоторых шедевров античного изобразительного искусства входят в произведения сатирика Лукиана (II век до н. э.).

Историки искусства, опираются в своих исследованиях на описания Лукиана «Дискобола» (Мирона), «Семейства кентавров» (Зевксида). Описания Лукиана отличаются точностью и конкретностью, они дают возможность наглядно представить себе произведение и раскрыть его характер. В античную эпоху умение описать статую или картину становится необходимым элементом в процессе подготовки оратора (специалиста в области произнесения публичных речей). Наиболее известным представителем риторической художественной критики был живший в Риме греческий писатель Филострат (Старший) – конец II в. н. э. Большой интерес представляет его книга «Картины». Филострат уже в полном смысле слова художественный критик. Известно его высказывание о мотивах критического творчества. В Предисловии к книге «Картины» он пишет: «Я хочу поведать о тех произведениях живописи, о которых была как – то у меня беседа с молодежью. Я ее вел с целью объяснить им эти картины и внушить интерес к вещам, достойным внимания». Филострат с большим критическим талантом раскрывал и комментировал сюжеты произведений изобразительного искусства, анализировал их форму и художественное мастерство автора. Его последователем и продолжателем был его внук – Филострат Младший (III в.н.э.), также оставивший ряд описаний произведений античной живописи.

Пример тестового документа на английском языке:

* Имя файла: covid.txt
* Тип документа: текстовый документ
* Размер: 4КБ
* Содержание:

The gold-standard method for diagnosing COVID-19 is the polymerase chain reaction (PCR) test, which detects the coronavirus’s genetic material in a nose or throat swab. Now, a survey of more than 15,000 people has singled out the people most likely to receive false negatives on the test.

Caitlin Dugdale at Massachusetts General Hospital in Boston and her colleagues looked at PCR test results from about 15,000 people who showed COVID-19 symptoms or were thought for other reasons to be infected with SARS-CoV-2. Nearly 2,700 individuals tested negative and had a second PCR test done within 2 weeks.

Among those who received a second test, 60 — or 2.2% — tested positive. Of these, 60% had their initial test either one day or less before symptom onset or more than 7 days after it, suggesting that the PCR test is most likely to yield a false negative in people tested early or late in the course of infection.

People with COVID-19 symptoms who test negative should be retested, especially in areas where the virus is widespread, the researchers say.

A far-reaching study of SARS-CoV-2 transmission in China’s Hunan Province found that the encounters that were most likely to spread the coronavirus were those between members of the same household.

Kaiyuan Sun at the National Institutes of Health in Bethesda, Maryland, Hongjie Yu at Fudan University in Shanghai, China, and their colleagues analysed data from 1,178 people in Hunan who were infected with SARS-CoV-2 and more than 15,000 close contacts of the infected people. The team found that contacts between people who live together posed the greatest risk of transmission, followed by contacts between members of an extended family. The transmission risk was lower still for social contacts and community encounters, such as those on public transport. Every extra day of contact raised transmission risk by 10%, the team found.

The analysis suggests that Hunan’s lockdown actually increased the risk of viral spread within households, whose members spent more time than normal at home together during lockdown. But social and community transmission fell during the same period.

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

Входная информация:

Текстовый документ, на русском или английском языке.

Выходная информация:

Переведённый текст, также при желании, пользователь может сохранить результат в файл “название\_оригинального\_документа + translated.txt”

**Описание алгоритма:**

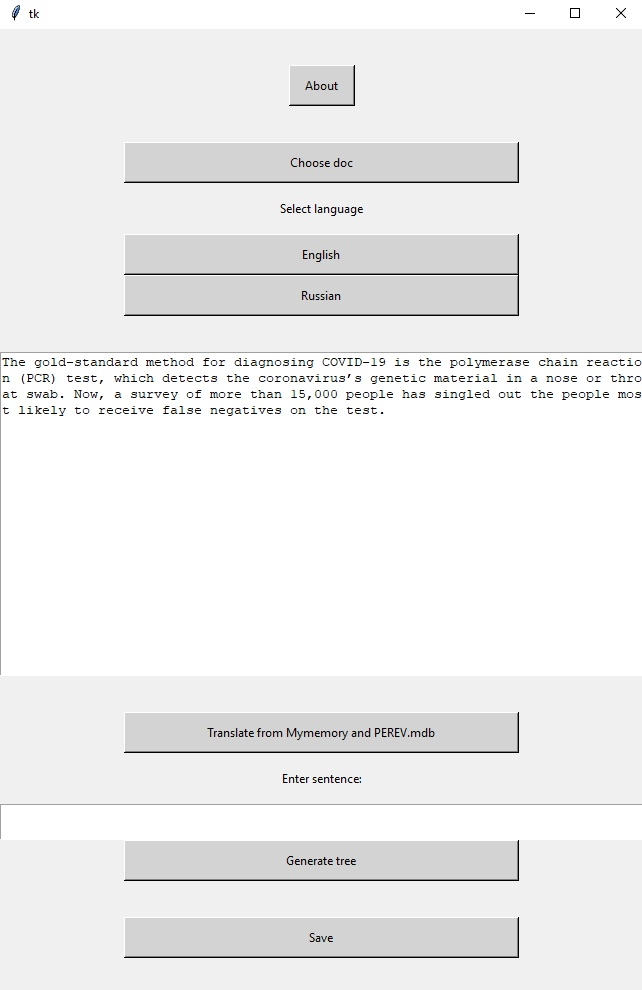
1. Пользователь выбирает документ, из которого будет взят текст.
2. Пользователь выбирает язык.
3. Приложение достаёт из таблицы Dict все данные и собирает их в массив, каждый элемент, которого представляет пару [(русское\_слово, английское\_слово)].
4. Приложение проходит по всем словам текста и, если встречает слово, которое есть массиве, полученном на предыдущем этапе, заменяет его на эквивалент на другом языке.
5. Приложение отправляет запрос на API MyMemory. В запрос входит частично переведённый текст и язык, выбранный пользователем на втором этапе.
6. MyMemory возвращает полностью переведённый текст и приложение отображает переведённый текст.
7. Пользователь может сохранить перевод в документ с именем “название\_оригинального\_документа + translated.txt”

**Результаты тестирования системы:**

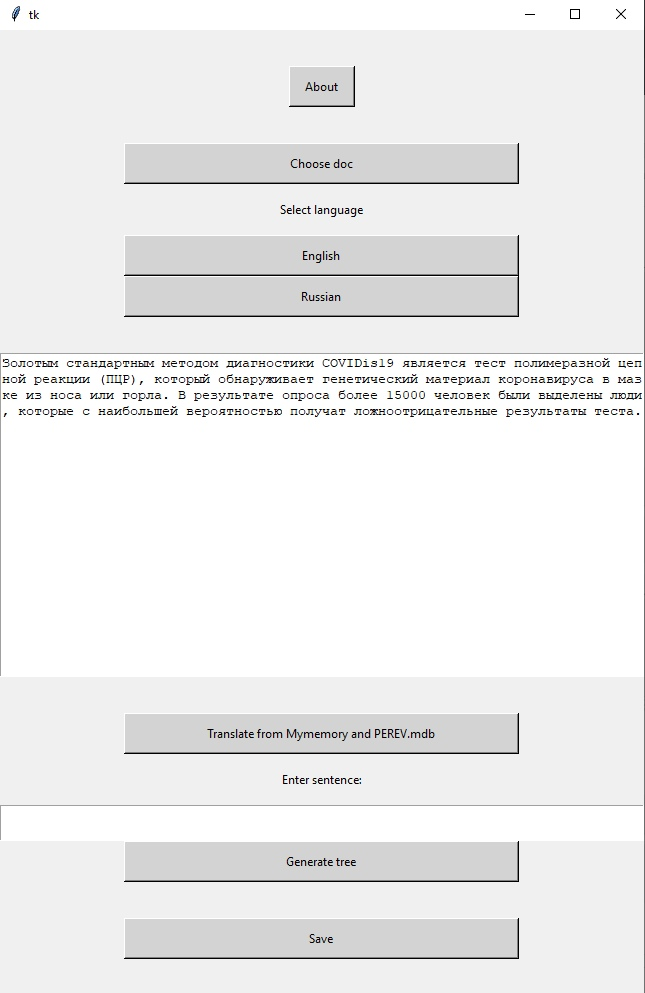
Стартовый экран:



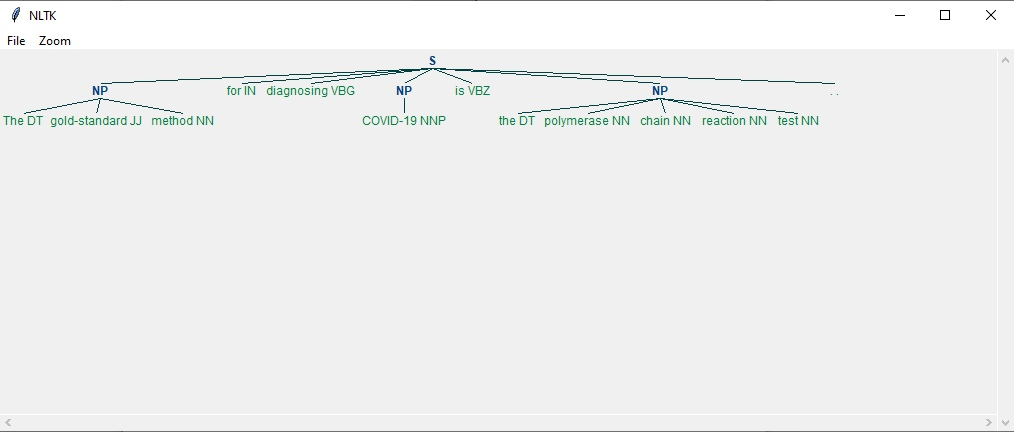
Тестовый документ covid.txt:



Переведённый текст:



Дерево синтаксического разбора:



**Описание использованных компонент:**  
pyodbc **-** библиотека для быстрого доступа к базе данных через ODBC.  
requests — это HTTP-библиотека, написанная на Python (под лицензией Apache2). Библиотека requests дает вам возможность посылать HTTP/1.1-запросы, используя Python.

**Вывод:**

В результате тестирования было выявлено, что система достаточно хорошо справляется с переводом текста используя MyMemory API.

Так как система имеет свою БД слов и выражений, то есть возможность расширить эту функциональность. Например, добавлять новые слова/выражения из переведенного текста в ручном или автоматическом режиме.