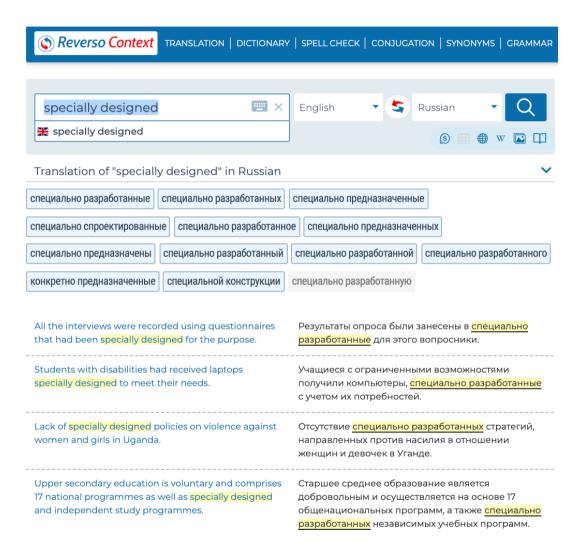
Предварительная версия: 23.12.2019

#### Проект: Контекстуальный англо-русский словарь

Декабрь 2019 г.

#### Пример для подражания - 1



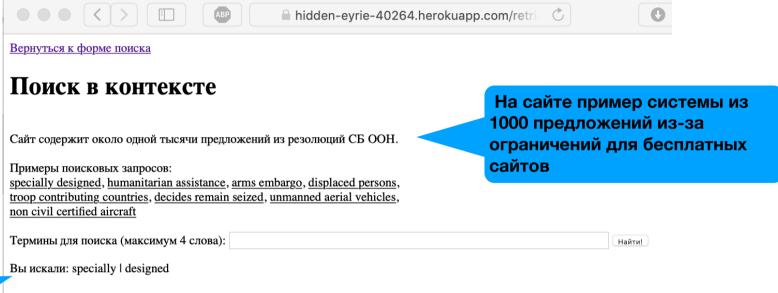
## Пример для подражания - 2 (в меньшей степени)



#### Что получилось

Адрес сайта

hidden-eyrie-40264.herokuapp.com



Поисковый запрос: "specially designed"

Коллокации из

предложений в

базе данных

#### Результаты поиска

Вибрационные испытательные установки и Vibration test equipment and specially designed parts <mark>специально спроектированные</mark> детали и and components capable of simulating flight компоненты, способные воспроизводить условия conditions of less than 15,000 meters. a. полета на высоте ниже 15 000 метров. а.2. установки, специально предназначенные для a.2. Equipment specially designed for ion ионовой имплантации, ионо- или фотоусиленной implantation, ion-enhanced or photo-enhanced диффузии, имеющие любую из следующих diffusion, having any of the following characteristics: характеристик: Delete "... radars specially designed for Опустить слова <<...радиолокаторы, специально meteorological use ...". созданные для метеорологических целей...>>. а. Программное обеспечение, специальное Software **specially** designed for the development or разработанное для создания или производства production of the components or equipment above; b. вышеуказанных компонентов или установок; а. Любой тип телекоммуникационного a. Any type of telecommunications equipment, оборудования, специально разработанного для specially designed to operate outside the temperature работы за пределами интервала температур от 218 . C. ... 010 IZ / EE0 (0) 4. 007 IZ /1040 (0)

Русский перевод выделяется по англорусскому глоссарию, полученному автоматически путем Word Alignment в NLTK

#### Источник текстов

https://conferences.unite.un.org/UNCorpus/

Англо-русские тексты ООН.

Три файла, в каждом 23 239 280 строк (текст уже разбит на предложения, каждая строка — предложение):

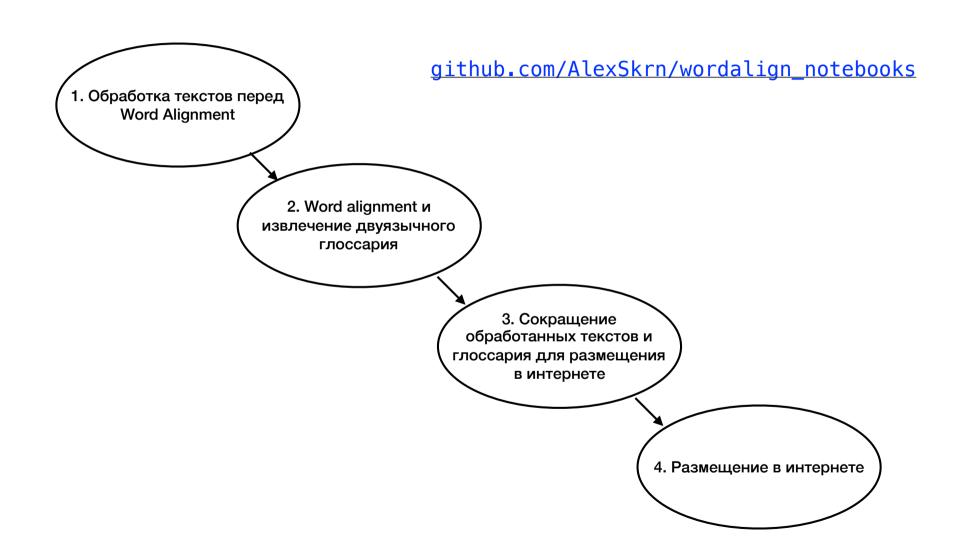
UNv1.0.en-ru.en - английский текст (3,45 GB)

UNv1.0.en-ru.ru - русский текст (6,95 GB)

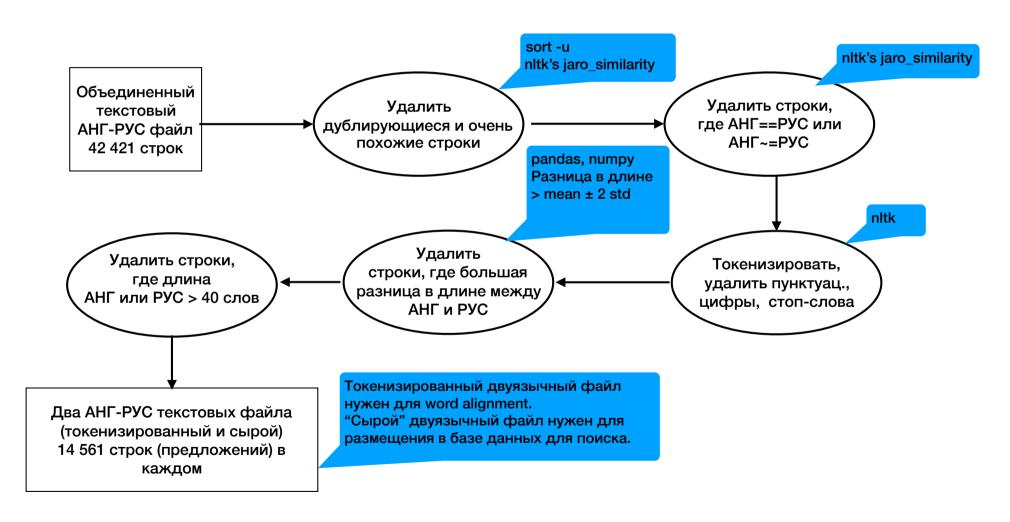
UNv1.0.en-ru.ids - индекс

Используя индекс, я отобрал только Резолюции Совета безопасности ОНН — 42 421 строк (предложений)

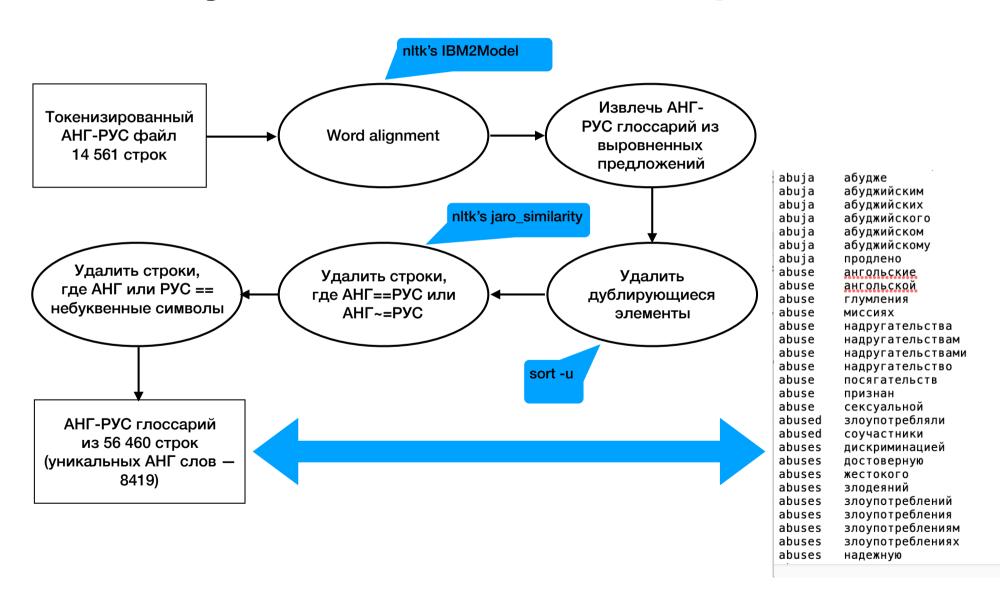
#### Этапы



# 1. Обработка текстов перед Word Alignment



### 2. Word alignment и извлечение двуязычного глоссария



### 3. Сокращение обработанных текстов и глоссария для размещения в интернете

Платформа Heroku имеет ограничение в 10,000 строк базы данных для бесплатных сайтов. Поэтому для размещения на сайте мне нужно было сократить текст до примерно 1 тысячи строк, а глоссарий до примерно 9 тысяч строк



# 4. Размещение в интернете

Адрес сайта: hidden-eyrie-40264.herokuapp.com

- Веб-фреймворк для написания сайта Flask
- База данных PostgreSQL, менеджер бд питоновский модуль peewee
- Примеры поисковых запросов на веб-сайте некоторые из наиболее частых n-граммов из английского текста
- Платформа для размещения сайта Heroku