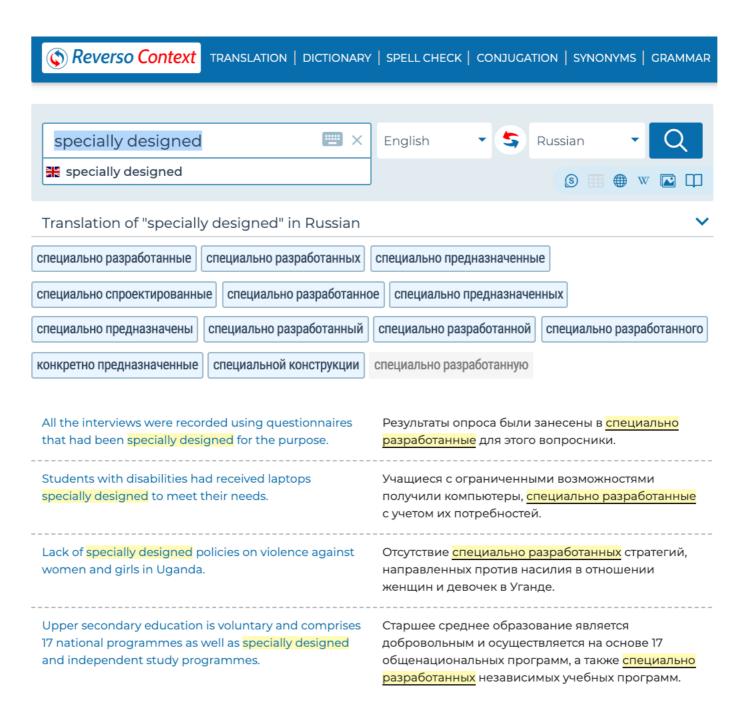
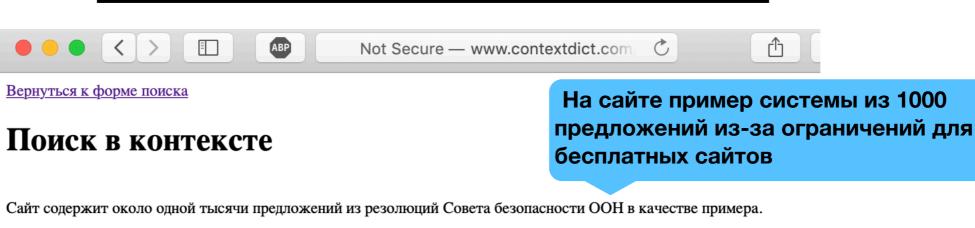
## Проект: Контекстный англорусский словарь

22 январь 2020 г.

## Пример для подражания – context.reverso.net



#### Что получилось www.contextdict.com



#### Примеры поисковых запросов:

specially designed, humanitarian assistance, arms embargo, displaced persons, troop contributing countries, unmanned aerial vehicles, non civil certified aircraft, вооруженные группы, РДР, разоружение и демобилизация

Термины для поиска (максимум 4 слова): specially

Вы искали: specially I designed

equipment specially designed specially designed

#### Результаты поиска

Vibration test equipment and **specially designed** parts and components capable of simulating flight conditions of less than 15,000 meters. a.

Вибрационные испытательные установки и специально спроектированные детали и компоненты, способные воспроизводить условия полета на высоте ниже 15 000 метров.

a.2. Equipment specially designed for ion implantation, ion-enhanced or photo-enhanced diffusion, having any of the following characteristics: а.2. установки, **специально** предназначенные для ионовой имплантации, ионо- или фотоусиленной диффузии, имеющие любую из следующих характеристик:

Delete "... radars **specially designed** for meteorological use ...".

Опустить слова <<...радиолокаторы, специально созданные для метеорологических целей...>>.

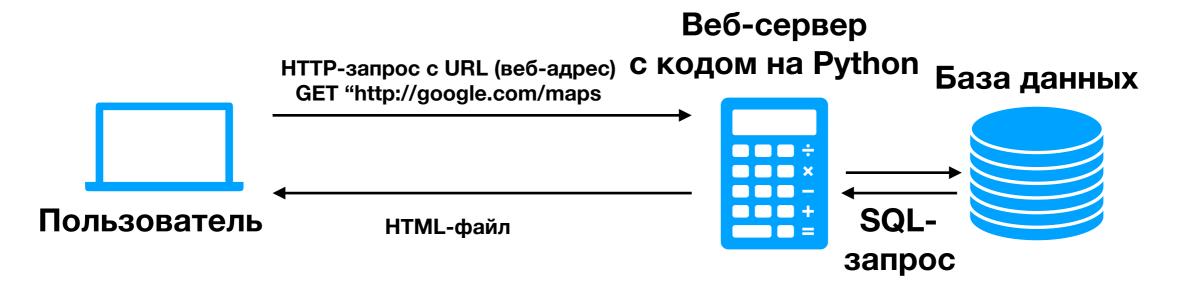
Software **specially designed** for the development or production of the components or equipment above; b.

а. Программное обеспечение, специальное разработанное для создания или производства вышеуказанных компонентов или установок;

Примеры запросов и функция AutoComplete на основе часто встречающихся пграмм

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

#### Как это работает



#### Таблицы в базе данных



### Источник параллельного корпуса

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):

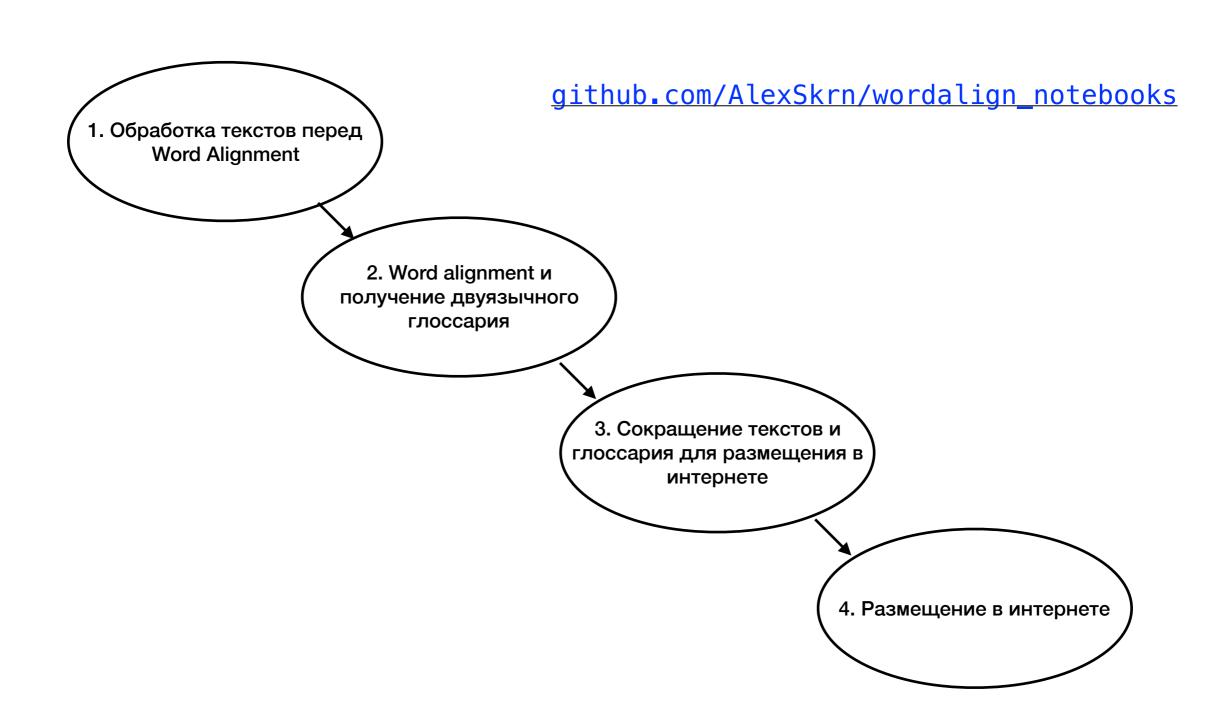
UNv1.0.en-ru.en - английский текст (3,45 GB)

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

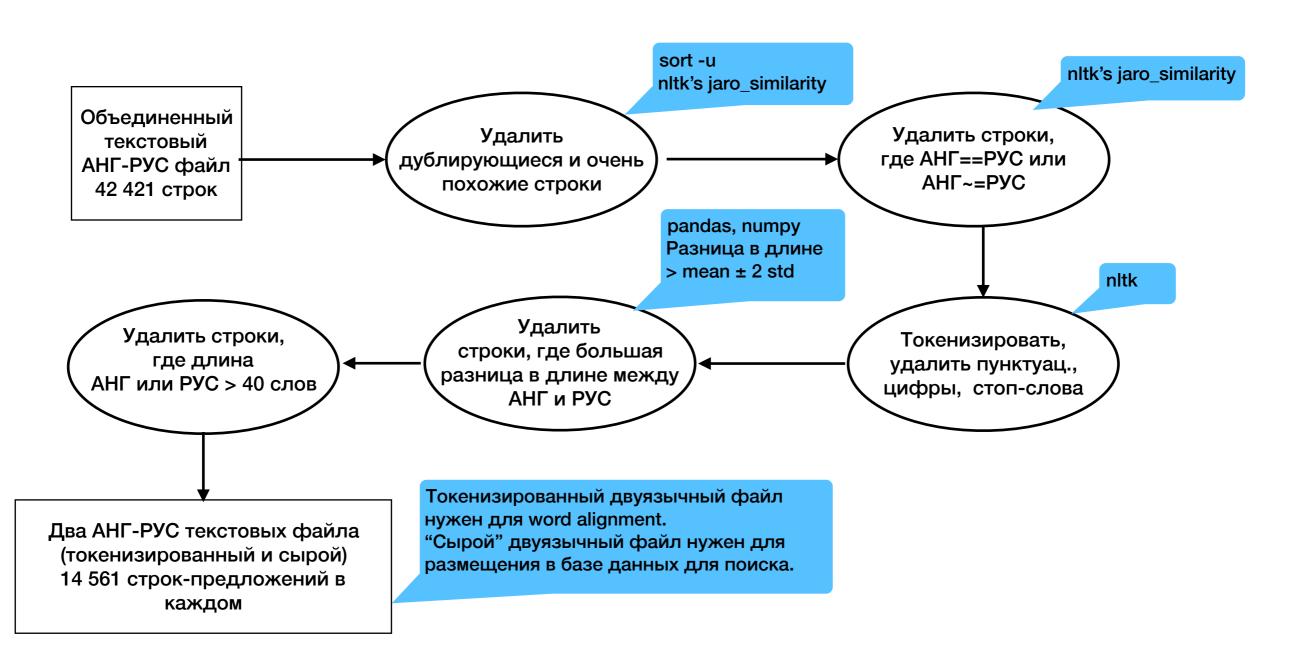
UNv1.0.en-ru.ids - индекс

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

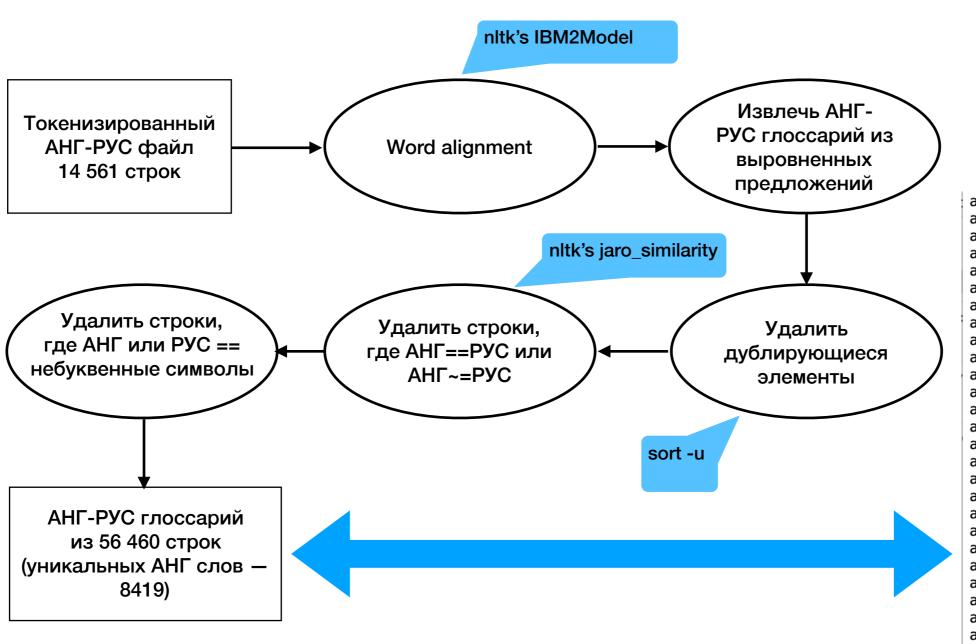
### Этапы



# 1. Обработка текстов перед Word Alignment



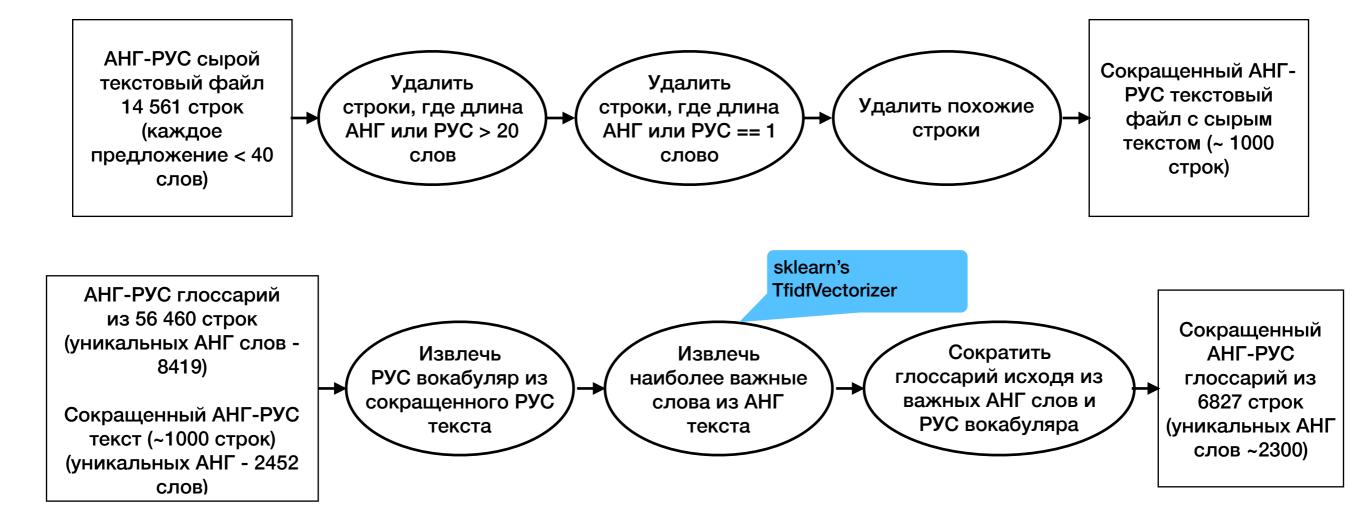
# 2. Word alignment и получение двуязычного глоссария



abuja абудже абуджийским abuja abuja абуджийских abuja абуджийского абуджийском abuja абуджийскому abuja abuja продлено abuse ангольские abuse ангольской abuse глумления abuse миссиях abuse надругательства abuse надругательствам abuse надругательствами abuse надругательство abuse посягательств abuse признан abuse сексуальной abused злоупотребляли abused соучастники abuses дискриминацией abuses достоверную abuses жестокого abuses злодеяний abuses злоупотреблений abuses злоупотребления abuses злоупотреблениям злоупотреблениях abuses abuses надежную

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

Хостинговый провайдер устанавливает ограничение в 10,000 строк базы данных для бесплатных веб-сайтов. Поэтому для размещения данных и приложения на веб-сайте мне нужно было сократить текст до примерно 1 тысячи строк-предложений, а глоссарий до примерно 9 тысяч строк



#### 4. Размещение в интернете

Адрес сайта: www.contextdict.com

- Веб-фреймворк для написания сайта Flask
- База данных PostgreSQL, менеджер базы данных питоновский модуль рееwee
- Хостинговый провайдер Heroku

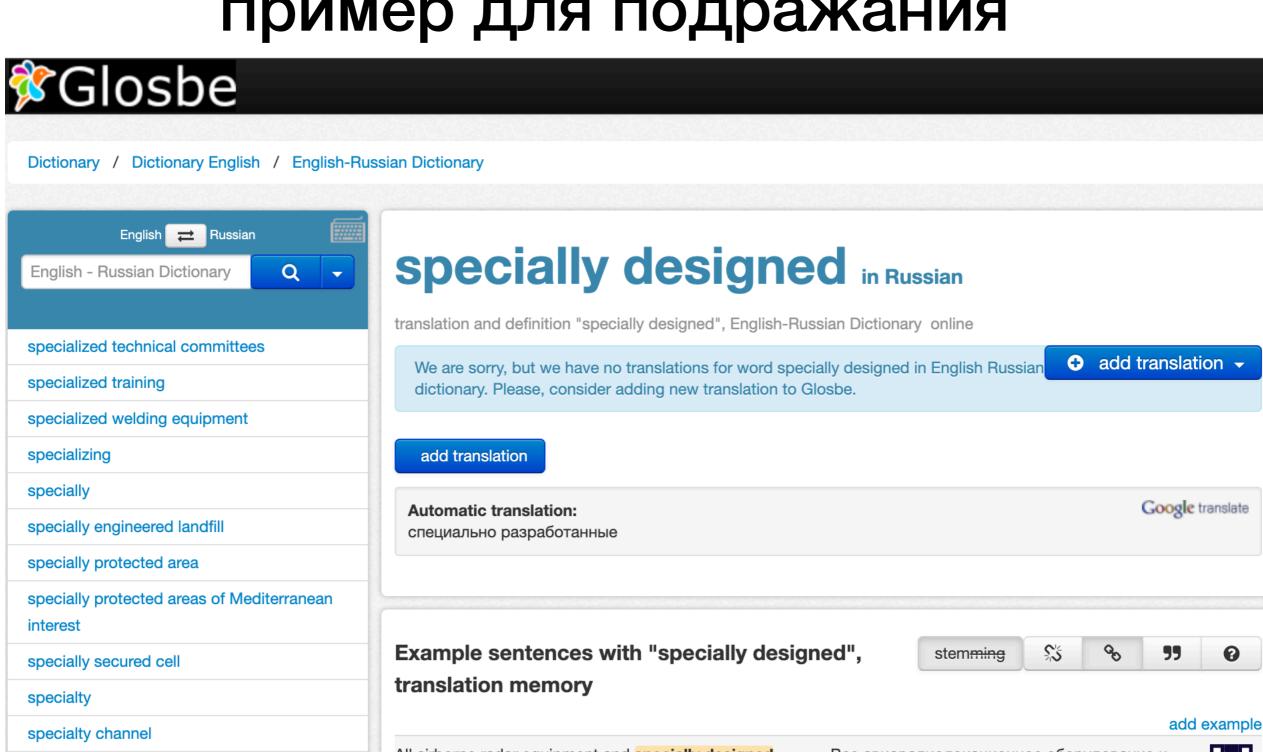
#### Следующие шаги

- Попробовать вместо глоссария использовать напрямую так называемый формат Giza, т.е. строки типа '1-1 2-2 ...' для каждой пары предложений (это означает, что первому слову в исходном предложении соответствует первое слово в переводе и т.д.)
- Попробовать использовать другой инструмент для Word Alignment

# Дополнительный пример для подражания



## Дополнительный пример для подражания



All airborne radar equipment and specially designed components therefor, not including radars specially designed for meteorological use or Mode 3, Mode C, and Mode S civilian air traffic control equipment specially designed to operate only in the 960-1215 MHz band:

specialty chemical

specialty welding machine

specialty crop

Все авиарадиолокационное оборудование и специально разработанные для него компоненты, за исключением радиолокаторов, специально созданные для метеорологических целей или работы в режиме 3, режиме С и режиме S гражданского авиадиспетчерского оборудования.