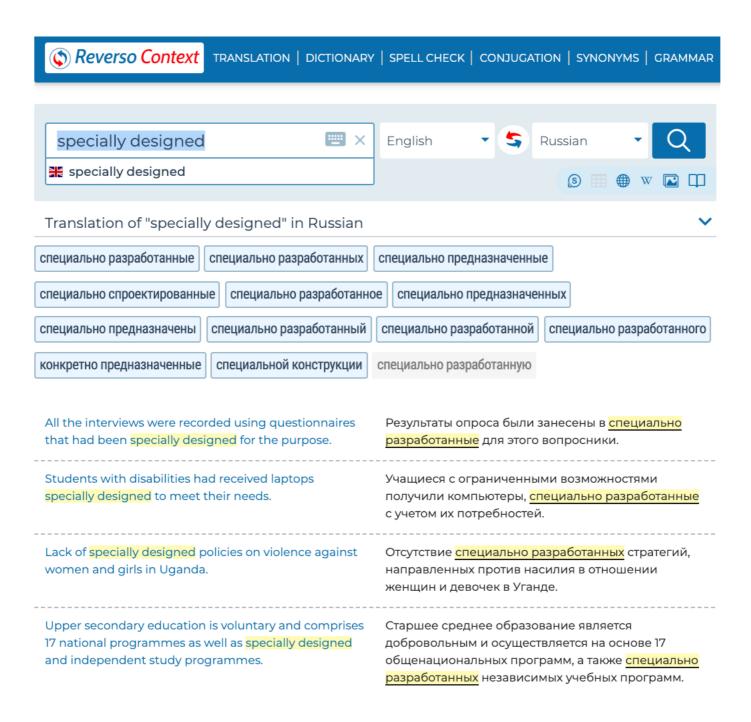
Проект: Контекстуальный англо-русский словарь

Январь 2020 г.

Пример для подражания - 1



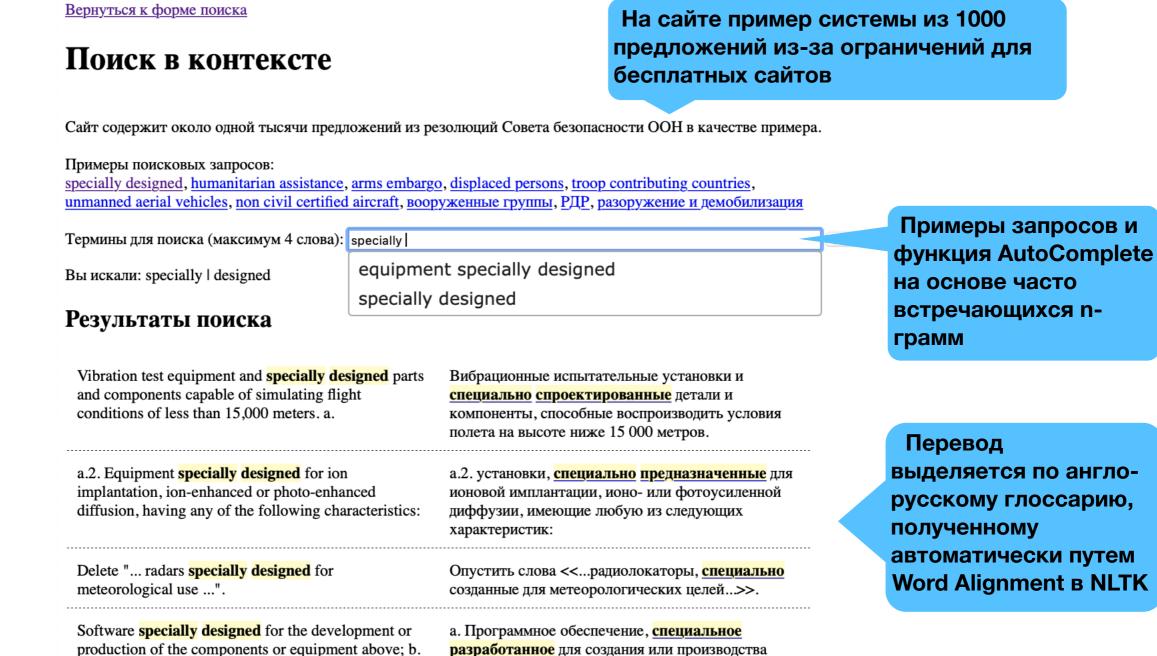
Пример для подражания - 2



Что получилось — contextdict.com

Not Secure — www.contextdict.com

вышеуказанных компонентов или установок;



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):

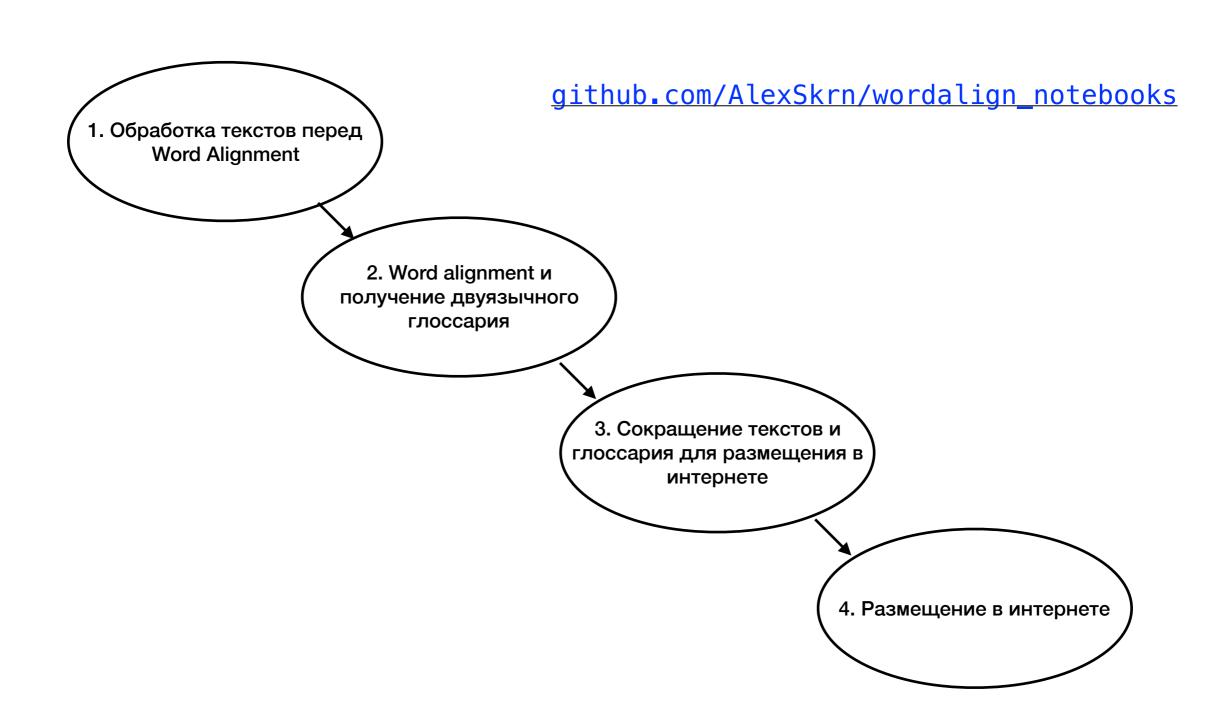
UNv1.0.en-ru.en - английский текст (3,45 GB)

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

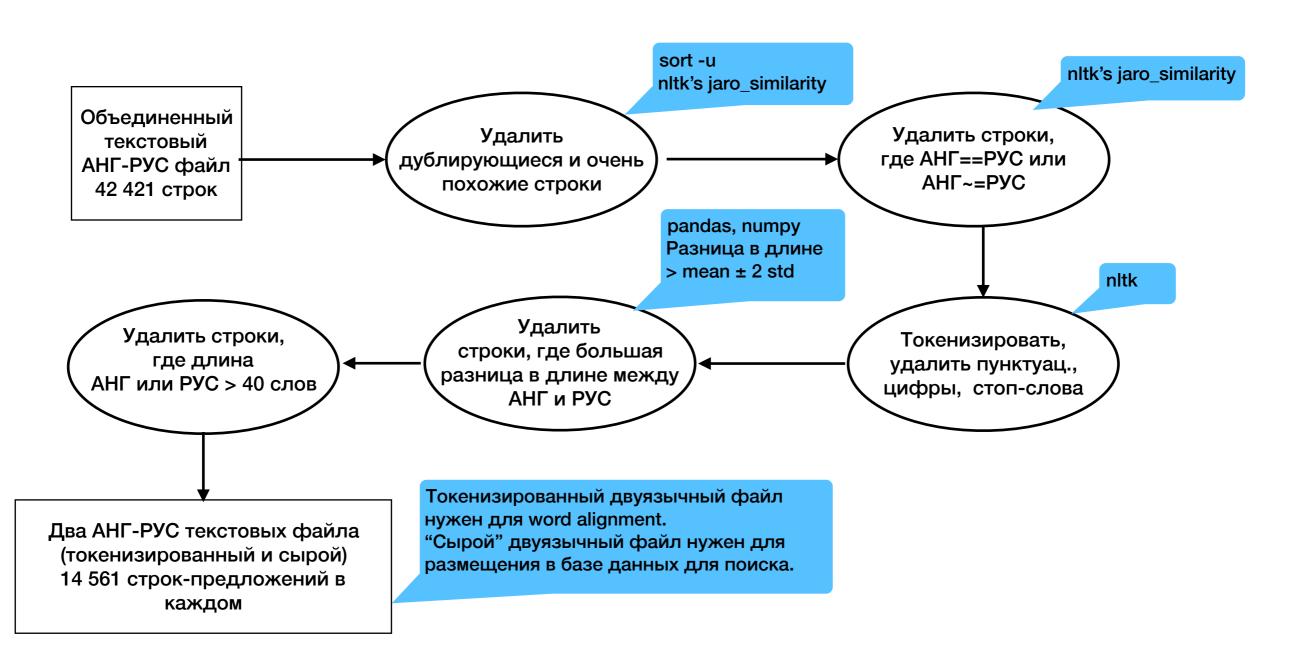
UNv1.0.en-ru.ids - индекс

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

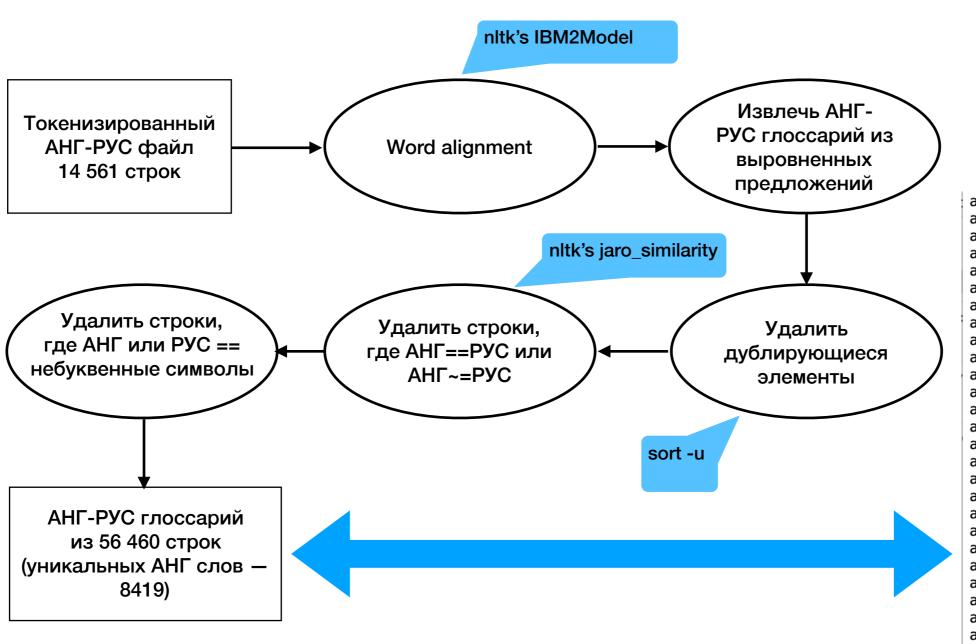
Этапы



1. Обработка текстов перед Word Alignment



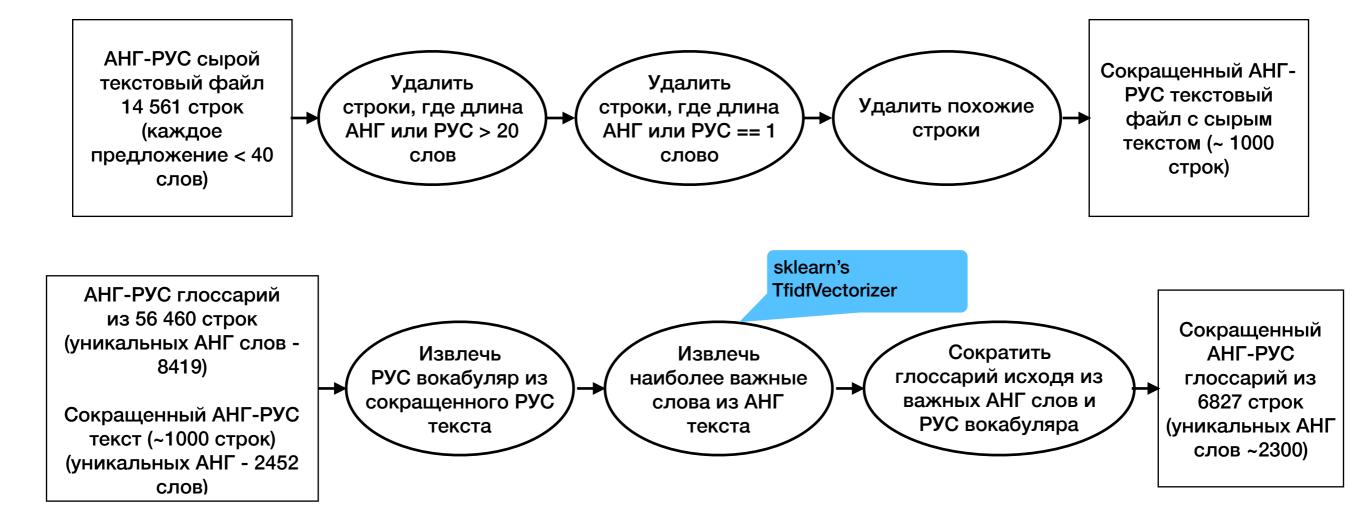
2. Word alignment и получение двуязычного глоссария



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

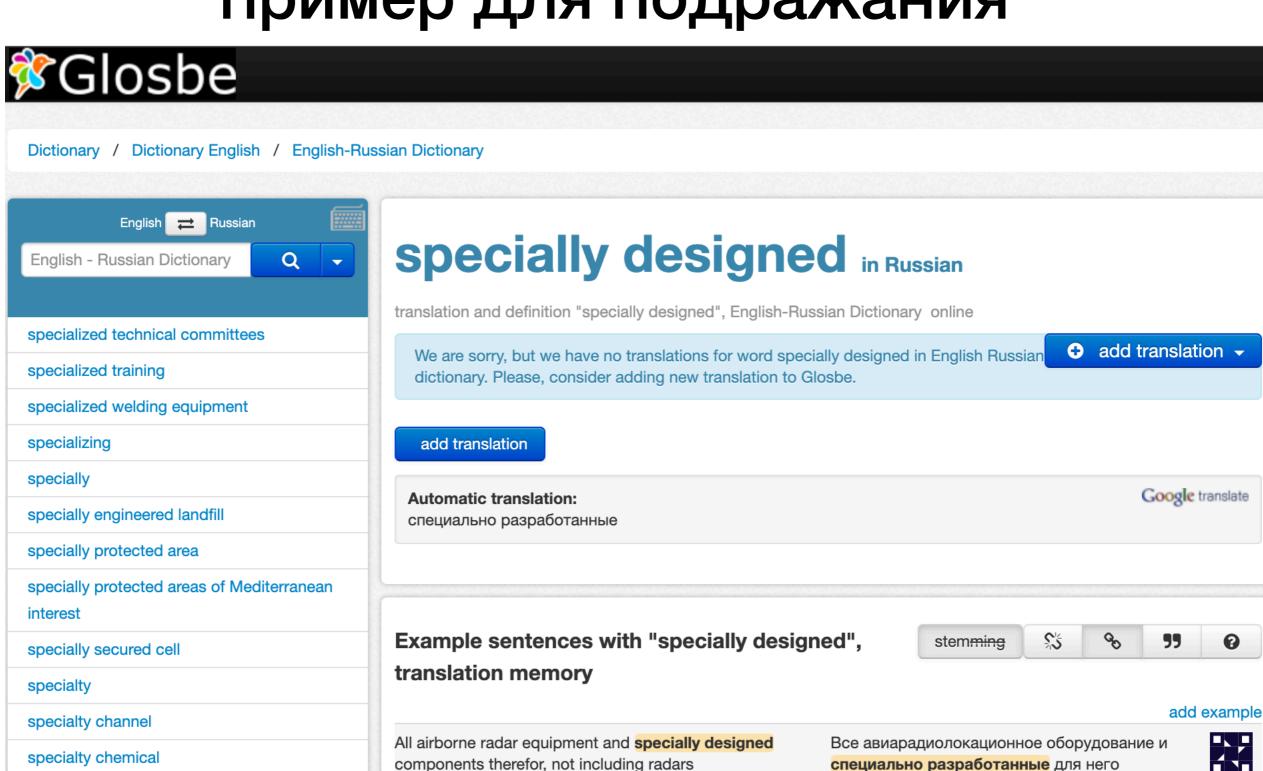
компоненты, за исключением радиолокаторов,

или работы в режиме 3, режиме C и режиме S

гражданского авиадиспетчерского оборудования.

специально созданные для метеорологических целей

Дополнительный пример для подражания



specially designed for meteorological use or Mode 3,

band:

Mode C, and Mode S civilian air traffic control equipment

specially designed to operate only in the 960-1215 MHz

specialty crop

specialty welding machine