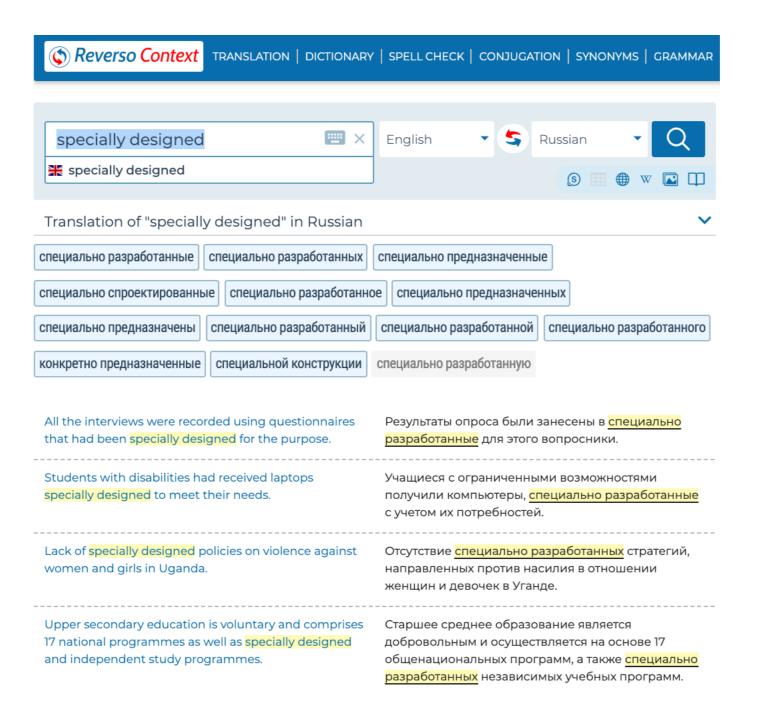
Предварительная версия: 23.12.2019

05.01.2020

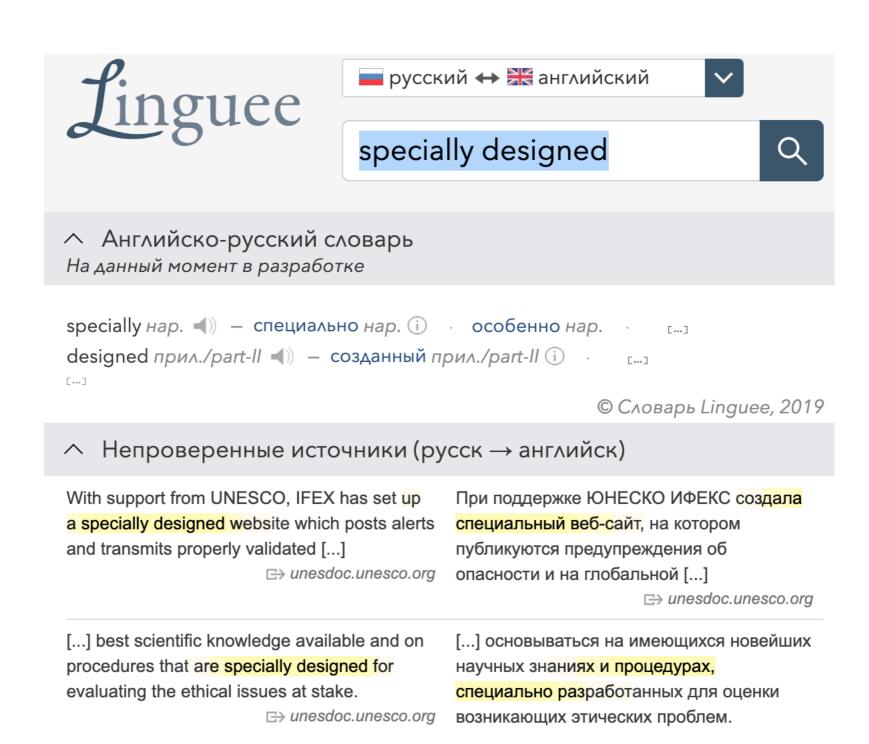
Проект: Контекстуальный англо-русский словарь

Январь 2020 г.

Пример для подражания - 1



Пример для подражания - 2



Что получилось

Адрес сайта

www.contextdict.com



Поиск в контексте

На сайте пример системы из 1000 предложений из-за ограничений для бесплатных сайтов

Сайт содержит около одной тысячи предложений из резолюций СБ ООН.

Примеры поисковых запросов:

specially designed, humanitarian assistance, arms embargo, displaced persons, troop contributing countries, decides remain seized, unmanned aerial vehicles, non civil certified aircraft

Примеры запросов и функция AutoComplete на основе часто встречающихся n-грамм

Найти!

Термины для поиска (максимум 4 слова): specially designed

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.

a.2. Equipment **specially designed** for ion implantation, ion-enhanced or photo-enhanced diffusion, having any of the following characteristics:

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

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

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

а.2. установки, специально предназначенные для ионовой имплантации, ионо- или фотоусиленной диффузии, имеющие любую из следующих характеристик:

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

а. Программное обеспечение, специальное разработанное для создания или производства вышему зачиных компочентов или установок:

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):

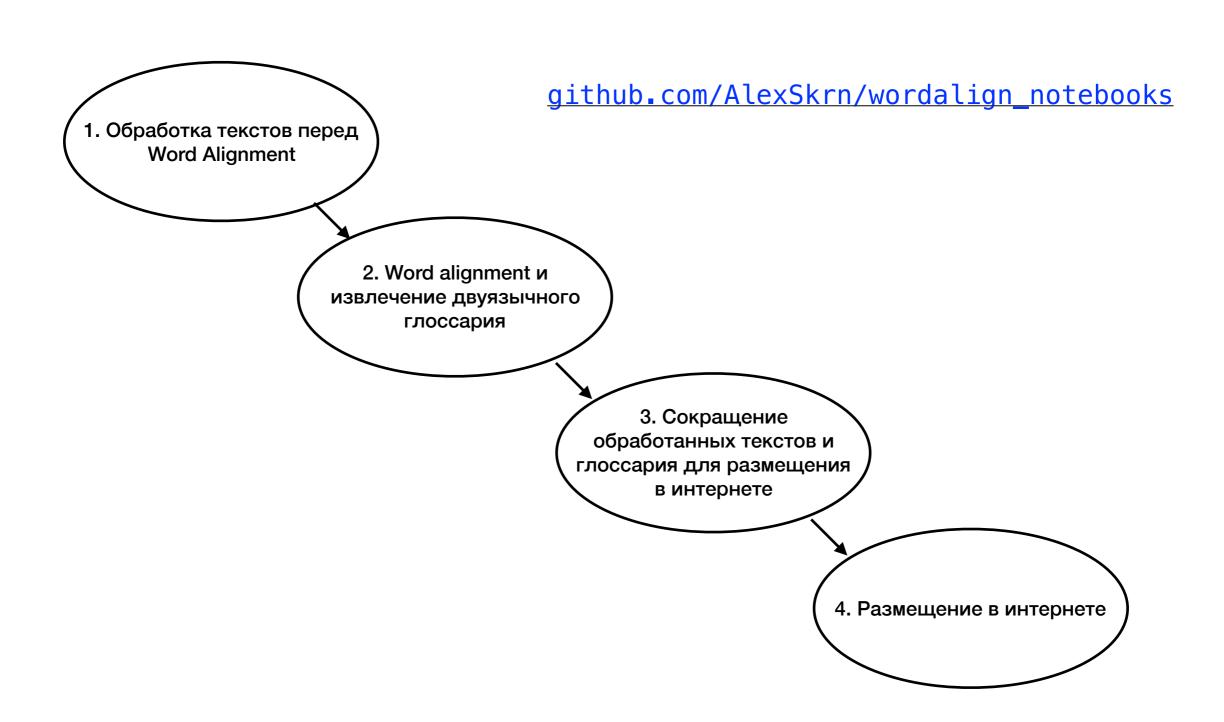
UNv1.0.en-ru.en - английский текст (3,45 GB)

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

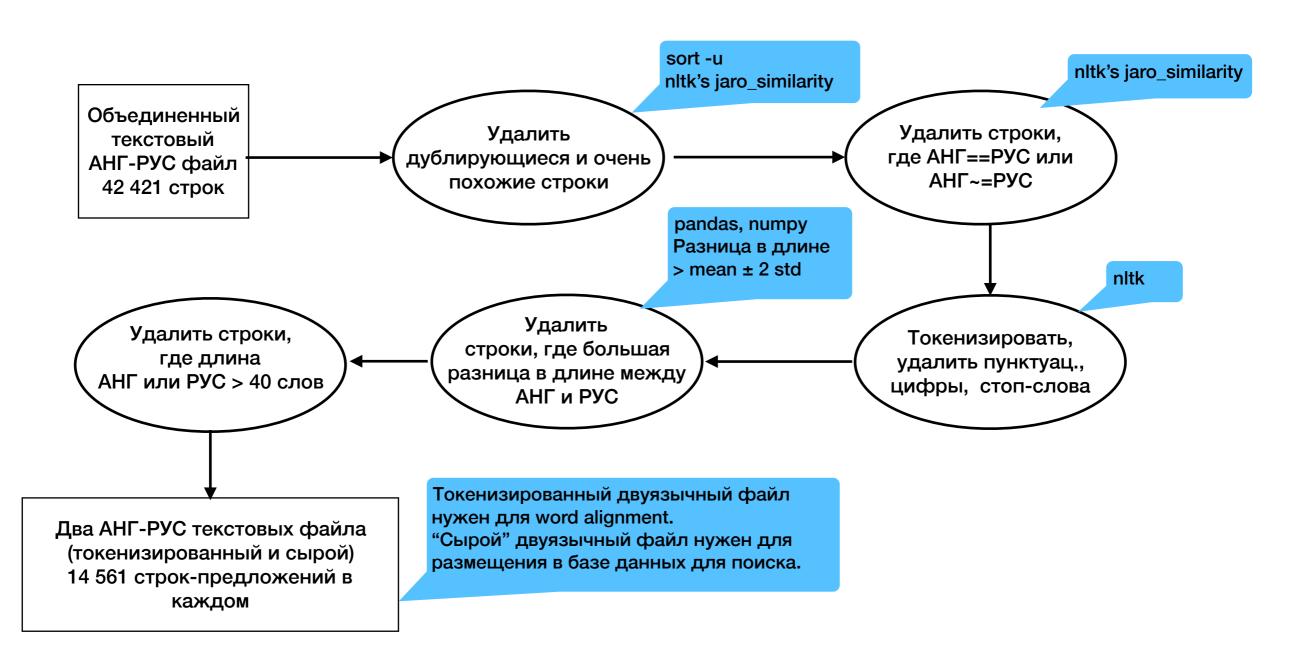
UNv1.0.en-ru.ids - индекс

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

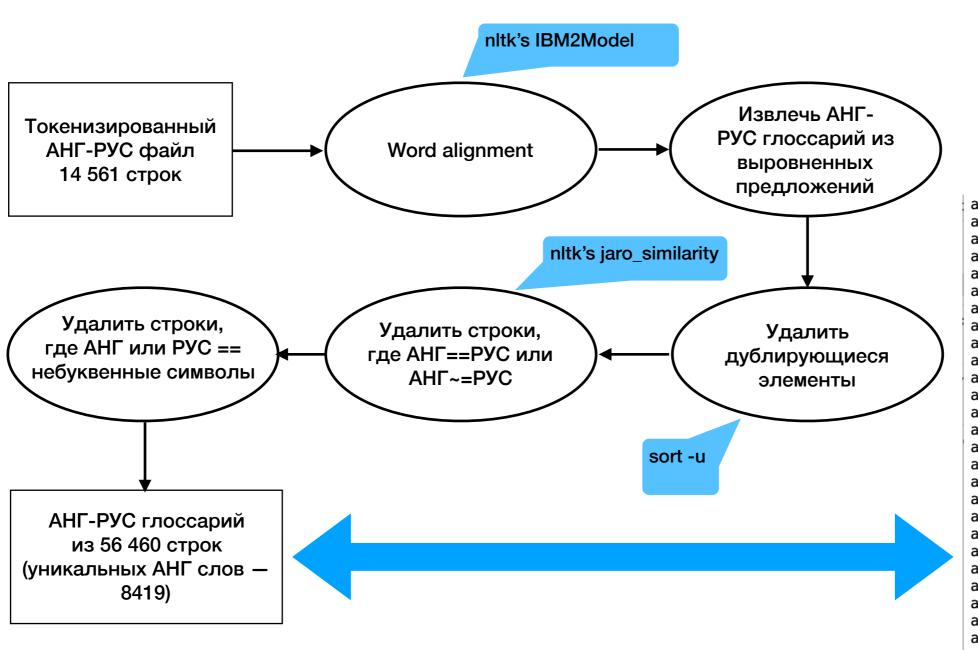
Этапы



1. Обработка текстов перед Word Alignment



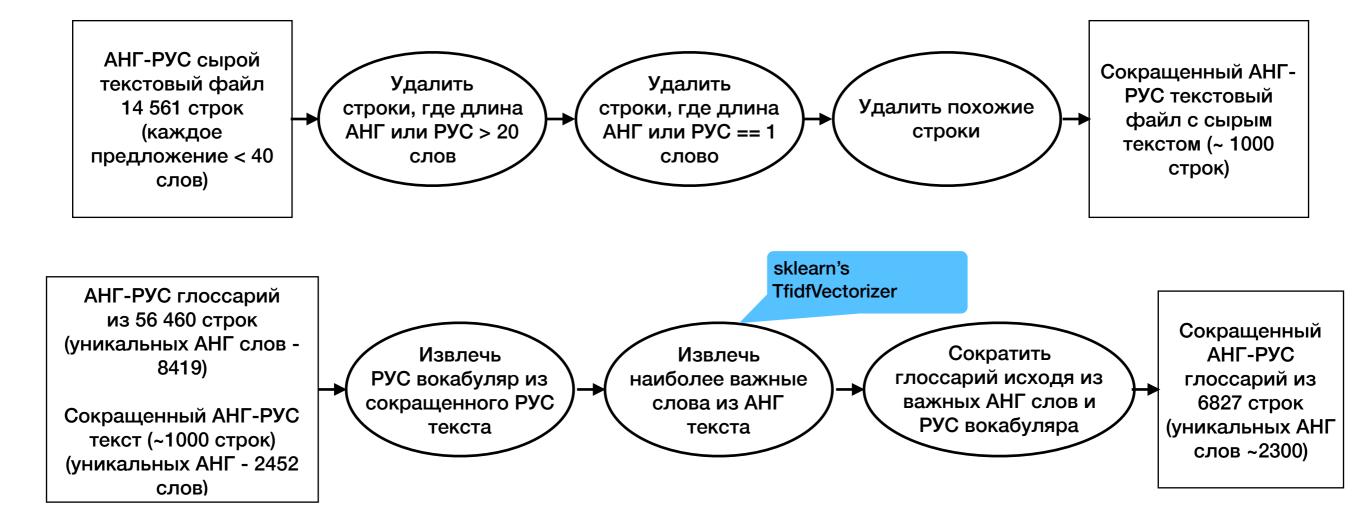
2. Word alignment и извлечение двуязычного глоссария



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 обработанных 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
- Провайдер доменного имени GoDaddy

Чего не хватает

- Поиска в направлении русский -> английский
- Функции AutoComplete в поле поиска 🔽
 - jquery (библиотека JavaScript) плюс список nграмм из текста как источник предлагаемых опций для AutoComplete

• ...

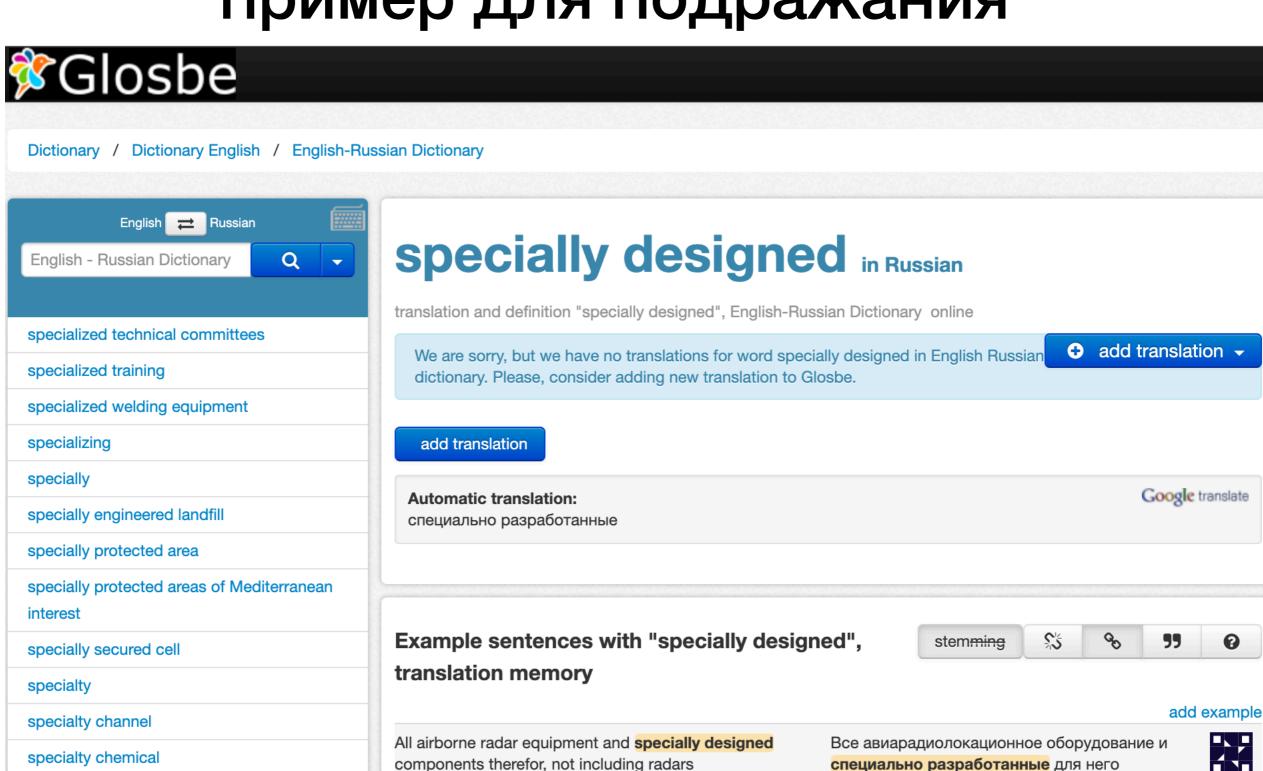
компоненты, за исключением радиолокаторов,

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