ОПИСАНИЕ

Кросс-языковой парсер на основе UD заключает в себе использование аннотированных данных одного языка для анализа другого. При этом аннотации для целевого языка нет и оба языка должны быть близкородственными.

Пары языков:

X	Y (целевой)
Немецкий	Идиш
Русский	Белорусский
Норвежский	Фарерский

1. язык X с разметкой в UD
2. язык Y без разметки в UD
3. параллельный корпус языков X и Y
4. при помощи UD парсера размечаем предложения в языке X
5. пословное выравнивание предложений при помощи fast_align
6. программа, которая переводит разметку предложений языка X на язык Y (на входе результат работы двух предыдущих пунктов)
7. Конечным продуктом проекта являются три парсера на основе UD для языков Y.

ЭТАПЫ

- обзор литературы
- выбор языков X и Y
- создание\поиск параллельных корпусов

Под созданием параллельного корпуса подразумевается скроллинг коллекции текстов на языке Y и их перевод на язык X при помощи машинного перевода (Гугл и Яндекс).

- разметка UD парсером языка X
- выравнивание предложений при помощи fast_align
- написание программы перевода UD разметки из одного языка на другой
- создание золотого стандарта на языке Y
- оценка качества разметки на языке Y

ЗАДАЧИ:

- написать краулер
- разобраться в API Google Translate и API Yandex Translate.

- ...

ТЕСТИРОВАНИЕ:
надо тестировать работу программы, которая переводит разметку из языка X на язык Y.
КАЛЕНДАРЬ: ссылка на наш гугл-календарь
ОТВЕТСТВЕННОСТИ:
Саша — немного менеджмента + программа, которая переводит разметку.
Костя — на данный момент работа с API Google Translate (идиш -> немецкий) и подготовка данных.
Паша — на данный момент работа с API Yandex Translate (идиш -> немецкий) и подготовка данных.