

Встреча 08.02.2018

Аня рассказала о проделанной работе: проба синтаксической разметки русинтаксом и перевод из .conll в .xml.

Настя рассказала о частотных ошибках: ошибки в коллокациях, эллипсис, нарушение устойчивых конструкций, дискурсивные ошибки. + можно учесть наличие нестандартно оформленных цитат и ссылок (пустые скобки, пустые кавычки). Описание ссылок формализовано (фамилия, год, страница), так что их можно найти. Надо добавить базовые вещи: частотные списки, гапаксы, встречающиеся в тексте работы, которые встречаются в проверяемых текстах, но не корпусах доменов.

Саше доделывать корпус из киберленинки + вордтувек.

Организационный вопрос по задачам Маши и Ани и их пересечению должен решиться в ближайшие дни.

Надо придумать регулярное выражение для определения ссылок.

Нужно подписать соглашение о неразглашении.