Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge

Kira Droganova and Nikita Medyankin

School of Linguistics, Faculty of Humanities, Higher School of Economics, Moscow, Russia

nikita.medyankin@gmail.com, kira.droganova@gmail.com

Pipeline workflow

Plain Text

Segmentation module

Mystem
- Tokenization
- Lemmatization
- Morphology

TreeTagger
- Case/number disambiguation

Postprocessing
- Undeclined nouns
- Multiword correction
- Tokenization fixes

MaltParser
- Syntactic annotation

CONLL Text
Evaluation

**Morphology**
Part of speech accuracy  97%
Part of speech and features accuracy  88%

**Syntax**
Labeled attachment score (LAS)  85%
Unlabeled attachment score (UAS)  90.7%

**Overall Pipeline Quality**
Labeled attachment score (LAS)  76.7%
Unlabeled attachment score (UAS)  84.1%
Web interface and Command line tool

http://web-corpora.net/wsgi3/ru-syntax/

https://github.com/tiefling-cat/ru-syntax