**Corpus Linguistics**

SoSe 2018

Instructor: Dr. Natalia Levshina [natalevs@gmail.com](mailto:natalevs@gmail.com)

**Topics**

1. Introduction. Types of corpora. Types of research questions.

2. Extraction and analysis of linguistic data from a tokenized corpus. The Leipzig Corpora collection.

3. Extraction and analysis of linguistic data from a lemmatized and morphologically annotated corpus. The COCA corpus.

4. Extraction and analysis of linguistic data from a syntactically annotated corpus. The Universal Dependencies corpus.

5. Extraction and analysis of linguistic data from a parallel corpus. The ParTy corpus of film subtitles.

6. Extraction and analysis of linguistic data from a transcribed spoken corpus. The Santa Barbara Corpus of Spoken American English.

Additional topics:

* Analysis of ngrams.
* Converting texts to XML and reading XML files. Online recipes and the British National Corpus.
* Syntactic and morphological annotation of corpora. Stanford Parser and other tools.

Freely downloadable tools:

* R <https://cran.r-project.org/>
* RStudio <https://www.rstudio.com/>
* Notepad++ <https://notepad-plus-plus.org/>

Classes that do not take place: 1 May (holiday) and 26 June 2018 (a conference).