**Introduction to Corpus Linguistics**

SoSe 2019

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

**Topics**

1. Introduction. Types of corpora. Types of linguistic annotation. Fundamental concepts.

2. How to create your own corpus and perform basic operations with the data in R.

3. Universal Dependencies corpora and udpipe software: lexical, morphological and syntactic annotation.

4. How to perform different corpus queries in R.

5. Building and analysing various frequency lists. Popular visualization tools.

6. Key words analysis: which words are over- and underrepresented in a corpus?

7. Identification and analysis of collocations: words that go together.

8. Text classification (basics of machine learning): which texts are more similar and which are different from a linguistic point of view?

Corpora we’re going to use:

* A reference corpus with different text types
* Trump’s twitter
* Song lyrics
* Recipes
* Texts from Pottermore
* Film subtitles
* Any suggestions? It’s up to you!

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 at the scheduled time due to international conferences: May 7, 14 and June 4. We’ll make arrangements to compensate them.

Exam: a term paper (about 10 pages). You should make your own corpus and perform the analyses: frequency lists, key words, collocations and visualizationd. The corpus data should be interpreted from a linguistic perspective.

Examples: comparison of speakers of two political parties, two authors, two film genres, linguistic analysis of your favourite writer, music band, recipe book, etc etc. – in any language you know well!