Felix Bildhauer (IDS Mannheim), Roland Schäfer (FU Berlin)

Describing corpora, comparing corpora

Large corpora built from mixed sources are used in many different sub-fields of linguistics, such as descriptive approaches; empirical work within formal morphological, syntactic, and semantic frameworks; computational linguistics; and even cognitively oriented linguistics. For many languages, researchers can choose between several linguistically annotated corpora. However, questions arise as to the validity and generalisability of findings on the basis of the particular corpus chosen. For the same phenomenon, findings may differ across different corpora, with a possible reason being the different types of language found in different corpora. In order to make sense of divergent findings and to enable a better assessment of the validity of results obtained only from a single corpus, we need to be able to characterise single corpora and compare different corpora using informative criteria. The "types of language found in a corpus" can be characterised in many ways. For example, external meta data at the document level (such as socio-demographic information) or the communicative function of texts (*register*, *text type*, *genre,* etc.) can be utilised to describe the composition of a corpus. If such meta data are not available, corpus creators sometimes derive or reconstruct meta data from the text in the document itself (using counts of linguistic features, etc.). Especially in the case of very large corpora, such data often have to be generated (semi-)automatically. Douglas Biber's attempts at inferring register categories semi-automatically are probably the most prominent such methods in corpus linguistic circles.
In this tutorial, we will take a closer look at corpus meta data suitable for corpus comparison and their use in linguistic case studies, with an emphasis on automatically created meta data.

Topics include:

- different kinds of document-level meta data
- reliability of automatically generated meta data
- usage of such meta data in linguistic corpus studies
- (dis-)advantages of feature aggregation (such as Factor Analysis in Biber's approach)
- aspects of statistical modelling

The course language is English. However, many examples and illustrations will be taken from corpora of German. As part of this tutorial, there will be hands-on exercises in order to give participants a chance to get started working with meta data available for German corpora.