Mittwoch
08.03.2017
15:45 – 16:30
B4 1, Foyer

COReX und COReCO: A lexico-grammatical document annotation framework for large German corpora

Felix Bildhauer
Institut für Deutsche Sprache, Mannheim
bildhauer@ids-mannheim.de

Roland Schäfer
Freie Universität Berlin (DFG SCHA1916/1-1)
roland.schaefer@fu-berlin.de

As an unstructured collection of documents, very large modern corpora would be of little use for many linguists. The automatic creation of linguistically relevant meta data is therefore crucial. We present an open-source annotation framework developed for the German DECOW and DeReKo corpora that provides (1) topical and (2) grammatical text categorization. The classification methods are based on document-internal features because this is the most viable approach for automatic classification, and it has conceptual advantages. Classifying documents by internally defined *text types* rather than situationally defined categories such as *register* or *genre* (Lee 2001) avoids many and potentially insolvable conceptual problems of determining and operationalizing the true set of register and genre categories (see the less than satisfying results in Biber & Egbert 2016). We classify documents by (1) distributions of words and (2) distributions of grammatical features.

CL Poster

For the lexical classification, we use topic modeling algorithms (e. g., Latent Dirichlet Allocation; Blei et al. 2003) combined with supervised machine learning to annotate documents with a coarse-grained set of twenty easily interpretable *topic domains* (such as *Politics* or *Sports*). For the grammatical classification, we annotate each document with automatically extracted features similar to Biber's (1988) features. We generate over thirty per-document features such as the density of modal verbs, genitives, or passive constructions. Users have access to the raw distributions of these features and aggregated categories obtained by clustering.

References: • Biber, Douglas. 1988. Variation across speech and writing. Cambridge, MA: Cambridge university Press. • Biber, Douglas and Egbert, Jesse. 2016. Using Grammatical Features for Automatic Register Identification in an Unrestricted Corpus of Documents from the Open Web. J. Res. Des. & Stat. in Ling. & Comm. Sc. 2, 3–36. • Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. Latent dirichlet allocation. J. M. L. Res. 3, 993–1022. • Lee, David. 2001. Genres, registers, text types, domains, and styles: Claryfying the concepts and navigating a path through the BNC jungle. L. Learn. & Tech. 5(3), 37–72.