Freitag
10.03.2017
13:00 – 13:30
B3 1, 0.12

Development and annotation of a newspaper corpus as part of a doctoral thesis on text structure and cohesion in news items from the 17th and 18th centuries

Katrin Goldschmidt
Rheinische Friedrich-Wilhelms-Universität Bonn
kat.goldschmidt@gmail.com

Historical news items in contrast to contemporary ones are less identifiable by typographic means. Moreover, some news items are syntactically and / or thematically linked to each other, i.e. by repetition of single event-related entities (person, location, time, action) in successive news items (Fritz & Straßner 1996). In order to examine how text and event structure contribute to the constitution and linking of historical news items, a corpus of historical newspapers has been developed, which enables either (text)linguistic or journalistic questions.

In the course of corpus development 7 German newspaper issues (1609-1767) were transcribed, subdivided in ca. 430 news items by three annotators, segmented into sentences (ca. 1,600 sentences), and ca. 30,000 token were tagged with parts of speech. The annotation of textual macro structures (such as sources, cited documents or comments) and event-related entities is based on a multilevel annotation scheme, that allows the annotation of complex spans (i.e. entities as discontinuous phrases) and relations between feature values (i.e. part-whole relations).

Investigating the hypothesis that news item boundaries are typically marked by punctation and typographic means, the presentation will provide an overview of the corpus and some evaluation possibilities with the analysis platform ANNIS (Krause & Zeldes 2014).

References: • Fritz, G. & E. Straßner (eds.) (1996): Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert. Tübingen: Narr. • Krause, T. & A. Zeldes (2014): ANNIS3: A New Architecture for Generic Corpus Query and Visualization. Literary and Linguistic Computing.