Mittwoch
08.03.2017
17:30 – 18:00
B3 1, 0.14

Segmentation and topic annotation of German newspaper editorials

Manfred Stede
Universität Potsdam
stede@uni-potsdam.de

AG2

Attempts to reliably identify aboutness topics in authentic language have shown that this task is notoriously difficult (Cook/Bildhauer 2013). We recently presented an annotation study (Stede/Mamprin 2016) where the annotator agreement shows some improvement over the state of the art, and we released a new annotation layer of aboutness topic on the Potsdam Commentary Corpus (Stede/Neumann 2014).

In the present work, we use that data for an initial qualitative study, which looks at the relationships between topics and subjects. A crucial factor here is segmentation, which in our approach consists of a largely structure-driven "generic" discourse segmentation, followed by a task-specific one (here: information structure) that filters for certain segment types. E.g., for subordinate clauses, we follow Matic et al (2014) in distinguishing between *d-subordination* (one complex proposition with the matrix clause) and *ad-subordination* (two distinct propositions).

Studying 20 texts with 316 discourse segments and 169 aboutness topics, we find that 119 (70%) coincide with subjects. The reasons for disjoint topics/subjects sometimes are structural (40%), while for the majority, there appears to be an underlying pragmatic choice (e.g., change of discourse topic). Almost half of the discourse segments in the data are *thetic* (topicless), and we provide a classification of the reasons.

References: • Cook, P. and Bildhauer, F. (2013): Annotating information structure. The case of "topic". In: *Dialogue and Discourse* 4(2):118–141. • Matic, D., van Gijn, R. and van Valin, R. (2014): Overview. In: van Gijn, R., Hammond, J., Matic, D., van Putten, S. and Galucio, A.V. (eds.): *Information structure and reference tracking in complex sentences*. Amsterdam: John Benjamins. • Stede, M. and Neumann, A. (2014): Potsdam Commentary Corpus 2.0: Annotation for discourse research. In: *Proc. of LREC*, Reykjavik. • Stede, M. and Mamprin, S. (2016): Information structure in the Potsdam Commentary Corpus: Topics. In: *Proc. of LREC*, Portoroz.