

---

## Supporting the Cognitive Process in Annotation Tasks

---

Nina Seemann, Michaela Geierhos, Marie-Luis Merten, Doris Tophinke,  
Marcel Wever, Eyke Hüllermeier  
*Paderborn University*

{seemann|geierhos}@hni.upb.de, mlmerten@mail.upb.de,  
{doris.tophinke|marcel.wever|eyke}@upb.de

Tag  
Datum  
Zeit  
Raum

Annotation tools typically use the common text analysis pipeline where (i) tokenization takes place, (ii) End-of-Sentences are detected, (iii) Part-of-Speech (POS) tags are assigned, and (iv) syntactic annotations are applied. But this does not work for non-standard data where rules or pre-trained models are not yet available for all steps, and boundaries for syntactic constructions are fluid. When annotating historical corpora, the previously mentioned sequence of steps has to be done manually in this strict order.

Therefore, we present an annotation tool that is guided by the cognitive annotation process that steps back from this pipeline. The need for such a tool showed up in our project InterGramm where we investigate Middle Low German (MLG) on morphological and syntactic level. We started our annotation task by using CorA (Bollmann et al., 2014), an established tool for historical data. Unfortunately, it does not support syntactic annotations, which we need to trace changing grammar rules from the 13th to 17th century. So we extended CorA, capturing uncertainties and ambiguities as well (Seemann et al., 2017). Experience has shown that being bound to start on token level before aggregating token sequences and aligning POS as well as construction tags appears to be difficult for human annotators. Naturally, linguists start with identifying syntactic patterns, then assign POS tags according to the corresponding context and decide during this process how many tokens belong to one lexeme. Furthermore, they prefer annotating in the direction of reading because it is easier to spot compound lexemes or syntactic constructions. Thus, we developed a new annotation tool with pattern learning support providing the annotators with suggestions inferred from previously studied MLG texts.

**References:** • M. Bollmann, F. Petran, S. Dipper and J. Krasselt (2014): CorA: A web-based annotation tool for historical and other non-standard language data. In: Proc. of the 8th LaTeCH, 86–90. • N. Seemann, M.-L. Merten, M. Geierhos, D. Tophinke and E. Hüllermeier (2017): Annotation Challenges for Reconstructing the Structural Elaboration of Middle Low German. In: Proc. of the Joint SIGHUM LaTeCH, 40–45.