

Marcel Bollmann¹, Stefanie Dipper¹, Mario Frank^{1,2}, Julia Krasselt¹, Florian Petran¹, Tom Ruetten²

¹Ruhr-Universität Bochum ²Humboldt-Universität zu Berlin

ANNIS-Hist: Historische deutschsprachige Korpora in ANNIS

In den letzten Jahren gibt es mehr und mehr Initiativen mit dem Ziel, historische Sprachdaten für die Forschungsgemeinde aufzubereiten und zur Verfügung zu stellen. Für Sprachdaten des Deutschen gibt es beispielsweise vier DFG-geförderte Projekte zur Erstellung von Referenzkorpora des Altdeutschen, des Mittel- und Frühneuhochdeutschen sowie des Mittelniederdeutschen/Niederrheinischen.

In unserem Poster und unserer Demo wollen wir eine Reihe von neueren historischen Korpora des Deutschen vorstellen:

- das Referenzkorpora Altdeutsch (750-1050)
(Berlin/Frankfurt/Jena, <http://www.deutschdiachrondigital.de>),
- das Referenzkorpora Mittelhochdeutsch (1050-1350)
(Bonn/Bochum, <http://referenzkorpora-mhd.uni-bonn.de>)
- sowie das Anselm-Korpora (Frühneuhochdeutsch, 14.-16.Jh)
(Bochum, <http://www.linguistics.ruhr-uni-bochum.de/anselm>).

Die Korpora sind u.a. annotiert mit:

- Metadaten, die beispielsweise Information über den (vermuteten) Entstehungszeitraum und Sprachraum des jeweiligen Textes enthalten,
- der normalisierten Wortform, die abstrahiert von den dialektspezifischen Wortformen,
- dem Lemma gemäß gängiger Referenzwörterbücher
- der Wortart nach HiTS ("Historisches Tagset") bzw. STTS und
- morphologischer Information.

Die Korpora und ihre Annotationen werden der Forschungsgemeinde über das Korpusuchtool ANNIS (<http://www.sfb632.uni-potsdam.de/annis>) verfügbar gemacht. In unserer Demo wollen wir verschiedene Such-Interfaces vorstellen, die für die Abfrage der historischen Korpora genutzt werden können. Wir wollen zum einen zeigen, wie mit Hilfe einfacher Anfragen ein erster Überblick über die vorhandenen Texte und ihre Annotationen gewonnen werden kann. Zum anderen wollen wir vorführen, wie für komplexe Suchen nach linguistisch interessanten Phänomenen Annotationen der verschiedenen Ebenen kombiniert werden können.