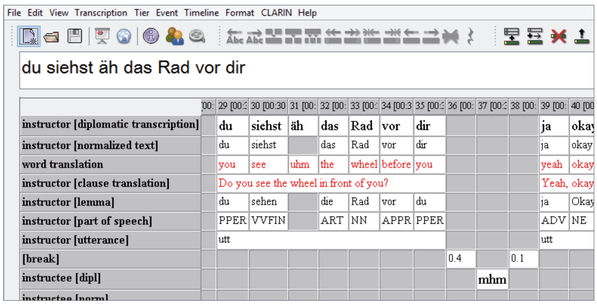
Digital Resources for Multilayer Annotation, Corpus Data Management and Publication, and Data Analysis at the Institute for German Language and Linguistics

Sustainability of Research Data:
In the past, data collected for linguistic research has often been ...
- > tailored only to the research question of the collecting group
- > shaped by implicit hypotheses
- > unavailable for replication or further analysis with a different research focus and/or by other researchers

Digital Resources Can Help...
- > to make analyses transparent, replicable, and extendable
- > to combine qualitative and quantitative aspects of a phenomenon
- > Providing researchers with structural evidence (not only anecdotic evidence)
- > Sharing research data between different research groups

Corpora as Research Resources:
- > Corpus: Digitally available collection of authentic texts
- > Corpus linguistics → empirical research method for theory formation
- > Corpora are often publicly available and digitally referenceable
- > Development of multilayer architectures: each type of information is added on an individual annotation layer (see Fig. 1)

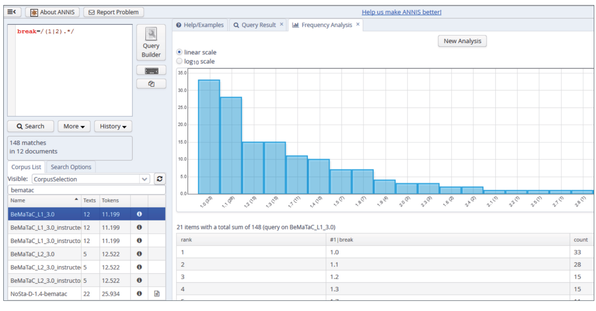
Explicit Data Modeling:
Corpora allow for...
- > transparent, explicit analyses
- > complex analyses with different types of information on independent levels of description
- > different depth of interpretation: corpus annotations are usually descriptions of primary data whose content can be processed and interpreted further



↑ Fig. 1: Excerpt from BeMaTaC corpus (hu-berlin.de/bematac) – spoken German dialog with orthographic transcription, normalization, and annotation of lemma forms, parts of speech, utterance chunks, and pauses. Annotations are processed and stored by the EXMARaLDA annotation tool (www.exmaralda.org).

Perspectives for Data Exploration and Quantitative Analyses in ANNIS:
ANNIS allows for ...
- > systematic searches on all imported annotation layers and their combinations
- > KWiC (key word in context) view for each search match
- > match export in several data formats for further data processing in statistical tools such as R (https://cran.r-project.org/)
- > frequency analyses (see Fig. 2)

Public Availability via Search Engine ANNIS:
- > ANNIS: Search Engine for complex annotated corpora (http://corpus-tools.org/annis/)
- > Online accessibility (internet browser) – Different instances for different corpus collections at different universities worldwide
- > Corpus collections at Humboldt University:
- > General collection of corpora hosted: https://korpling.german.hu-berlin.de/annis3/
- > Corpora of Old High German: https://korpling.german.hu-berlin.de/annis3/ddd
- > Learner Corpora: https://korpling.german.hu-berlin.de/annis3/falko



← Fig. 2: ANNIS search tool with query for pauses (length between 1.0–2.9 sec.) and frequency distribution of pause length type