Digital Transformation: Jointly Facing a Grand Challenge

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

- Corpus linguistics → empirical research
- Development of multilayer architectures: each type of information is added on an individual annotation layer (see Fig. 1)

File Edit View Transcription Tier Event Timeline Format CLARIN Help du siehst äh das Rad vor dir [00: 29 [00:: 30 [00:30 | 31 [00: 32 [00: 33 [00: 34 [00:3 | 35 [00:: 36 [00: 37 [00:: 38 [00: 39 [00: 40 [00 instructor [diplomatic transcription] siehst äh |das |Rad |vor du okay ja siehst du okay instructor [normalized text] das Rad vor jа wheel before you word translation the yeah okay uhm see you Yeah, okay instructor [clause translation] Do you see the wheel in front of you? Okay instructor [lemma] sehen Rad du vor ja PPER VVFIN ART NN APPR PPER ADV NE instructor [part of speech] instructor [utterance] utt utt 0.4 [break] instructee [dipl] $\mathbf{m}\mathbf{h}\mathbf{m}$

↑ 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).

More ▼

BeMaTaC_L1_3.0_instructe(12 11.199

22

BeMaTaC_L1_3.0_instructo 12 11.199

History ▼

Texts Tokens

12 11.199

12.522

12.522

25.934

0

0

Query

අත

linear scale

break=/(1|2).*/

Q Search

148 matches

in 12 documents

BeMaTaC_L1_3.0

BeMaTaC_L2_3.0

BeMaTaC_L2_3.0_instructe 5

BeMaTaC_L2_3.0_instructo 5

NoSta-D-1.4-bematac

Corpus List | Search Options

Visible: CorpusSelection

instructee Inormi

Quantitative Analyses in ANNIS:

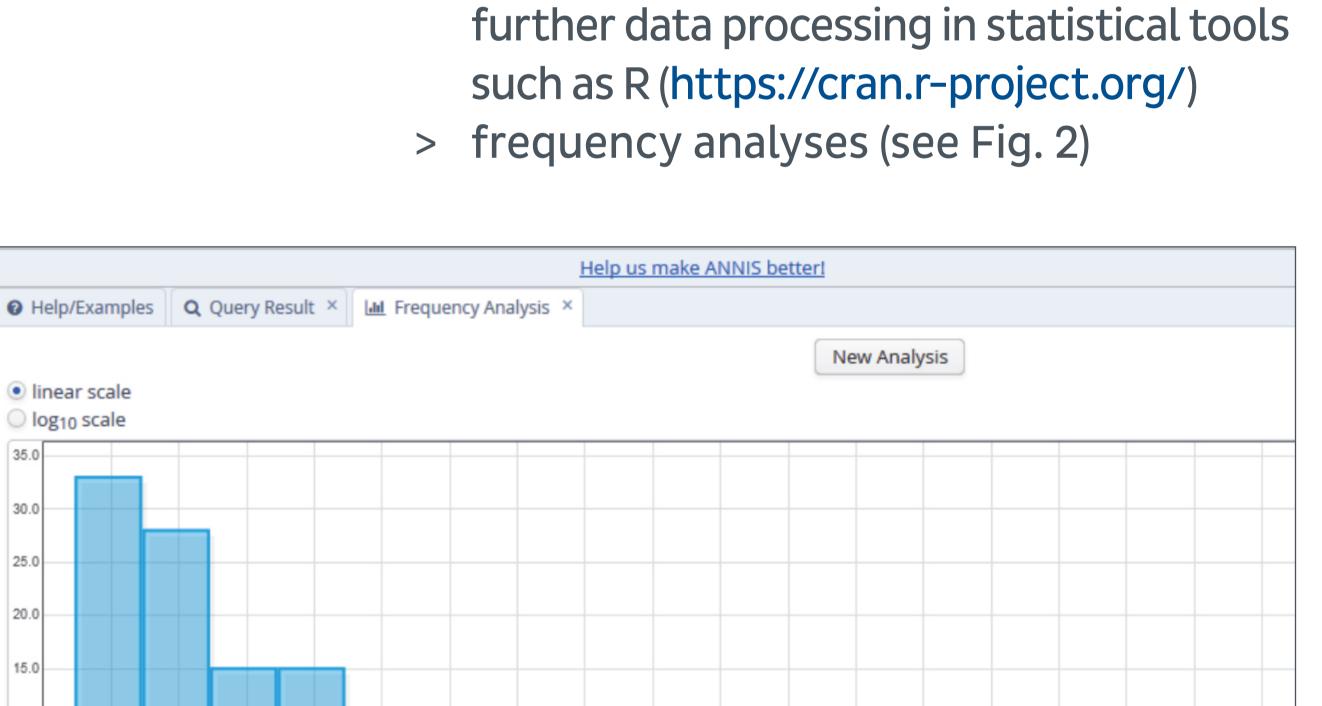
ANNIS allows for ...

- > systematic searches on all imported annotation layers and their
- > match export in several data formats for

Perspectives for Data Exploration and

- KWiC (key word in context) view for each search match
- such as R (https://cran.r-project.org/)

combinations



Corpora as Research Resources:

- > Corpus: Digitally available collection of authentic texts
- method for theory formation
- Corpora are often publicly available and digitally referenceable

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

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

Berlin University Alliance



21 items with a total sum of 148 (query on BeMaTaC_L1_3.0)

#1|break

1.0

1.1

1.2

1.3





count

33

28

15

15

