### Primary Data

- German interviews
- Professional host vs. guests of varying eloquence
- Ongoing collection (100+ interviews)
- Raw audio (~10 min each)
- Edited transcripts
- „Semi-planned“ speech

### Manual Output

- Modified Transcripts
  - 20 Interviews
  - ~3 hours of audio
  - ~36,000 written tokens
  - Part-of-Speech (TIGER/STTS)
  - Discourse
  - Information Status (RefLex)

### Automatic Output

- Silver Standard
  - Phonetics
  - Prosody
  - Part-of-Speech
  - Syntax
  - Coreference

### Confidence Estimation

- Tool-internal score
- Error-detection
- Agreement between different tools

---

### Data

- Non-static collection of German radio interviews (currently 100 interviews of about 10 minutes each, starting from May 2014)
- Multi-modal data: audio recordings and written transcripts
- Both types of available primary data are equal in status: transcript reflects decisions of transcriber

### Confidence Estimation

- Provided as additional (meta-)annotation layers so they can easily be used in exploration tools like ICARUS
- Exploiting annotation redundancies
  - horizontally (multiple annotations of same type)
  - vertically (related annotations of different types)
- Multiple sources for confidence estimation
  - exposure of tool-internal scores
  - counting relative number of tool-specific fallback decisions, e.g. root attachment of leftover tokens
  - error or inconsistency detection
  - agreement scores between multiple tools
  - evaluation results on appropriate gold-standard subset

- Ideally at least one confidence estimation per “real” annotation
- No standardization defined for confidence values yet
- no general interpretation – only tool-internal comparability
- no comparability for confidence values from different sources
- Confidence information can then be used to create excerpts of the data suitable for a given research question or application

### References


---

This work was funded by the German Research Foundation (DFG) via the SFB 732, project INF.