



# Interfacing linguistic analysis tools with a database for result management – workflows in sentence and text analysis

Kurt Eberle, Kerstin Eckart, Ulrich Heid

{eberle,eckartkn,heid}@ims.uni-stuttgart.de

## Three types of relations between analysis components – integrated environment for quality assurance in corpus based linguistic analysis

### Vertical relations

- Pipeline architecture of text processing
- 'High level' analysis, e.g. constituent trees, depends on results of 'lower' levels, e.g. morphological analysis
- Advantageous for corpus studies:
  - Shared interest in 'lower' levels
  - 'Higher' levels computed more efficiently from results of 'lower levels'
  - Reusability of intermediate results
- Prerequisites:
  - Analysis tools have to support pipeline architecture
  - Analyses are stored and administrated for later reuse

### Horizontal relations

- Different tools producing analyses of a particular level, e.g. dependency analyses
- Taking corresponding results of the same level into account
- Advantageous to facilitate quality assurance of the annotations: comparison of analysis results
- Prerequisites:
  - Analyses have to be identifiable with respect to their horizontal status, i.e. analysis level and representation format
  - Format conversions for compatibility, e.g. into an abstract exchange format such as GrAF [Ide/Suderman 2007]

### Temporal relations

- Analysis tools evolving over time
- Producing analyses for the same input but with different versions of a tool
- Advantageous for system development
  - Valuable clues to tool improvement or decline, or to specific changes of the knowledge base
  - Identification of side-effects by comparing earlier versions of the analysis
- Prerequisites:
  - Information about tool and component versions
  - Analyses have to be relatable to the tools or annotators producing them

### Relational database

- B3DB, implemented as a PostgreSQL database [Eckart et al. 2010]
- Type system identifies the horizontal status of an analysis
  - Relating analyses and tool versions
  - Displaying annotation level and representation format
- Workflow modelling identifies vertical status of an analysis
  - Relating input and output wrt the analysis level
  - Relating tool versions that evolve over time
- Flexible queries conducted via SQL

### Multi-level processing tool

- B3-analysis-tool, based on a research prototype of the German parser of the *lingenio* machine translation product *translate* [Eberle et al. 2008]
- Adapted to collaborative linguistic research
  - ⇒ pipeline where each annotation level can be extracted separately
- Modules for morphological, syntactic, semantic and text semantic/pragmatic analyses
- Stored analysis settings provide the complete knowledge needed by subsequent analysis steps of the pipeline
  - All levels contribute to a detailed analysis
  - Analyses are connected to each other by text and sentence identifiers

### Interface: generic handling of different levels

- Access to DB analysis frontend via dbanalyze commands specifying
  - Input, type of input and type of output
  - Optional parameters to fine-tune the corresponding analysis
- General form:

dbanalyze(  
analysis(InputID,InputAnalysisType),  
Language,TypeofAnalysis,Domain,  
AdditionalParameters).

• Creating a syntactic analysis from the morphological one of DE sentence 315 in file 3:  
dbanalyze(  
analysis(3,315,morph),de,syn,[],[]).

### Use case

Task-specific disambiguation of German *ung*-nominalizations:

*nach*-PPs in combination with nominalizations of *verba dicendi* – *Mitteilung* ('announcement'), *Anmerkung* ('remark') [Eberle et al. 2009]

- Two readings of the preposition *nach*:
  - temporal ('after') vs. content-referring ('according to')
- Two readings of the nominalization of a *verbum dicendi*, e.g. *Mitteilung*:
  - event reading: 'the act of making an announcement' vs.
  - object reading: 'the content of the announcement'

### Example

#### Primary data

Sentence from local news (file 3, sentence 315):

(3,315)*Er verblieb nach seiner Mitteilung in stationärer Krankenhausbehandlung.*

He remained in stationary hospital treatment after/according to his announcement.

#### First step: morphological analysis

dbanalyze(sent(3,315),de,morph,[],[]).

#### Creating further steps, e.g. syntactic analyses

directly from sentence: dbanalyze(sent(3,315),de,syn,[],[]).

via existing analyses: dbanalyze(analysis(3,315,morph),de,syn,[],[]).

#### Pronoun resolution, e.g. with two preceding sentences

dbanalyze(analysis(3,315,syn),de,res,[],[prec:2]).

### Future work

- Technical extension: interface enhancement to full database capabilities and a platform independent version of tool and interface
- Architectural extension: taking into account horizontal relations and further analysis levels, such as DRS represented semantic structures

### References

- SFB 732: <http://www.uni-stuttgart.de/linguistik/sfb732/> – Lingenio: <http://www.lingenio.de/English/Research.htm> – PostgreSQL: <http://www.postgresql.org/>
- [Eberle et al. 2008] Kurt Eberle, Ulrich Heid, Manuel Kountz and Kerstin Eckart. A Tool for Corpus Analysis using partial Disambiguation and Bootstrapping of the Lexicon. In: Storrer, Angelika, Alexander Geyken, Alexander Siebert and Kay-Michael Würzner (eds.): Text Resources and Lexical Knowledge (Berlin: Walter de Gruyter), 145-157. 2008.
- [Eberle et al. 2009] Kurt Eberle, Gertrud Faaß and Ulrich Heid. Proposition or Temporalangabe? Disambiguierung von -ung-Nominalisierungen von verba dicendi in nach-PPs. In: Christian Chiarcos et al., editor, Von der Form zur Bedeutung: Texte automatisch verarbeiten / Proceedings of GSCL 2009, pages 81 – 91, Tübingen. Gunter Narr Verlag. 2009.
- [Eckart et al. 2010] Kerstin Eckart, Kurt Eberle and Ulrich Heid. An Infrastructure for More Reliable Corpus Analysis. WSP2010 Workshop. LREC 2010. Malta. 2010.
- [Ide/Suderman 2007] Nancy Ide and Keith Suderman. GrAF: A graph-based format for linguistic annotations. In Proceedings of the Linguistic Annotation Workshop, 1-8. 2007.