ANNIS – a Linguistic Database for Complex Multilevel Annotation

Stefanie Dipper, dipper@ling.uni-potsdam.de
Michael Goetze, goetze@ling.uni-potsdam.de
Manfred Stede, stede@ling.uni-potsdam.de
Institute of Linguistics
SFB 632 “Information Structure”
University of Potsdam, Germany

1 Background
The goal:
- Integrative models of information structure (IS)
- Typology of the different means of expressing IS
The problem:
- Many factors seem related to IS
- There is a multitude of terminologies and theories
The method:
- Investigations based on empirical data, annotated by appropriate information
The database ANNIS:
- represents the data in a uniform way
- offers different ways of visualizing the data
- offers retrieval facility

2 The Data
Source data:
- Transcriptions of recorded speech
- Monologues and dialogues
- Written texts
- Questionnaire data (question-answer pairs)
Annotated data:
- Phonetics/phonology (e.g. prosody)
- Morphology, syntax (e.g. word order)
- Semantics (e.g. quantifier scope)
- Discourse structure (e.g. rhetorical relations)
- Metadata
- Language properties (e.g. typological information)

3 Requirements
Annotations:
- multi-layered
- tiers vs. trees/|graphs
- conflicting hierarchies (e.g. prosody and syntax)
- intra-sentential and inter-sentential
- meta data (including typological information)
Visualization:
- different views on data
- customization: focus on/filter out specific features
- discourse view vs. dialogue view vs. tier-based view

4 Implementation
Exmaralda, RST Tool, MMAX, annotate

5 ANNIS Query Facilities
The ANNIS Query Language is designed for an intuitive formulation of
- complex, multilevel queries (e.g. by providing operators for treebank queries) and
- search across (sub-)corpora with diverse annotation schemes
Users are further supported by a highlighting of matching expressions, a query storage function and various export facilities.

6 Implementation
ANNIS is a JAVA Servlet Application, which is accessed via standard web browsers. It can be run both locally (e.g. on a laptop in field studies) or on a central web server.
Internally, data is represented as JAVA objects and all data is loaded into memory. In the near future, a DBMS will supplement the RAM-based approach.

7 XML Representation
Generic XML format:
- represents all kinds of linguistic annotation in a uniform way
- serves as the input to the visualization module
Strict stand-off architecture:
- conflicting hierarchies, competing annotations
- allows for different tokenization
Features may be defined in a type hierarchy:
- underspecified annotation

8 Visualization
- HTML
- specified by an XML formatting file
- features can be coloured, filtered out, customizable