31. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Postersession der Sektion Computerlinguistik, 5.3.2009

Search and Visualization of Richly Annotated Corpora with ANNIS2

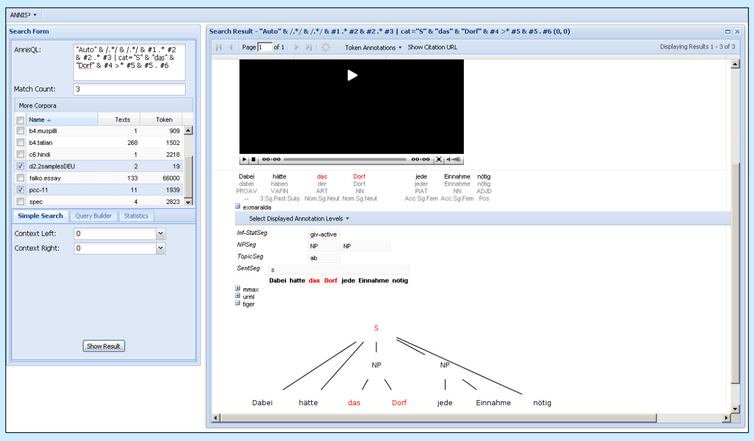
SFB 632 "Information Structure"
Project D1: Linguistic Database Annotation and Retrieval

Christian Chiarcos*, Thomas Krause+, Anke Lüdeling+, Julia Ritz*,
Viktor Rosenfeld+, Manfred Stede*, Amir Zeldes+ and Florian Zipser+

+ Humboldt-Universität zu Berlin, * Universität Potsdam

Background

- ANNIS2 is a versatile **web browser-based search and visualization architecture** for complex multilevel linguistic corpora
- Designed to provide search and visualization facilities for complex annotations such as **information structure** and **coreference** created in different subprojects of the SFB 632
- Developed for research involving **diverse annotations simultaneously**, allowing studies of the **interaction between different phenomena** (e.g. **syntax** or **prosody** and information structure)



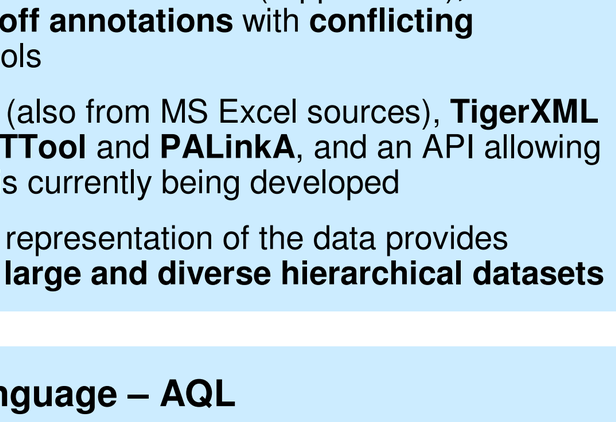
- The system imports data in the **PAULA XML** format (Dipper 2005), which allows multiple independent **standoff annotations** with **conflicting hierarchies,** created in different tools
- Converters exist for **EXMARaLDA** (also from MS Excel sources), **TigerXML** (Annotate/Synpathy), **MMAX2**, **RSTTool** and **PALinkA**, and an API allowing greater import/export functionality is currently being developed
- An advanced **relational database** representation of the data provides infrastructure for **fast searches in large and diverse hierarchical datasets**

Search with ANNIS Query Language – AQL

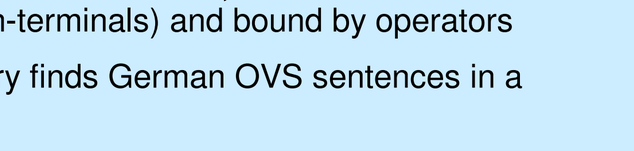
- Based on a simple model of richly **annotated nodes** and **labelled edges**
- Similar to Tiger Query Language and NITE-QL, search nodes are declared (annotations, terminals and non-terminals) and bound by operators
- For example, the following query finds German OVS sentences in a syntactically annotated corpus:

```
node & pos="VVFIN" & cat="S" & node &      two nodes, a verb and an S-node
#3 >[func="OA"] #1 &                       S dominates node 1 with label OA
#3 >[func="SB"] #4 &                       S dominates node 4 with label SB
#3 > #2 &                                  S dominates the verb
#1 .* #2 &                                 node 1 precedes the verb
#2 .* #4                                   the verb precedes node 4
```

- A **graphical Query Builder** makes formulating complex queries easier
- **Operators** define the possible **overlap** and **adjacency** relations between annotation spans, as well as **recursive and labelled hierarchical** relations between nodes
- Users can search in **multiple corpora simultaneously**



- Full support for **Regular Expressions** (RegEx) in tokens, annotations and edge label values
- Full **Unicode support** in both search and visualization, including RegEx

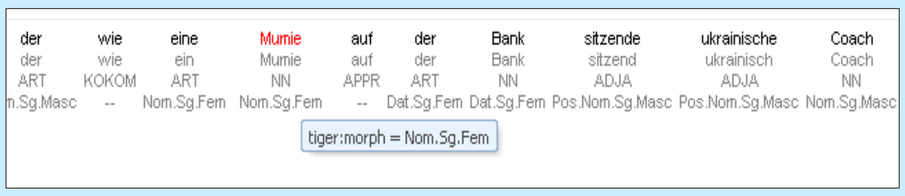


- Search result nodes and their annotations can be exported to **WEKA** (Witten & Frank 2005) for machine learning

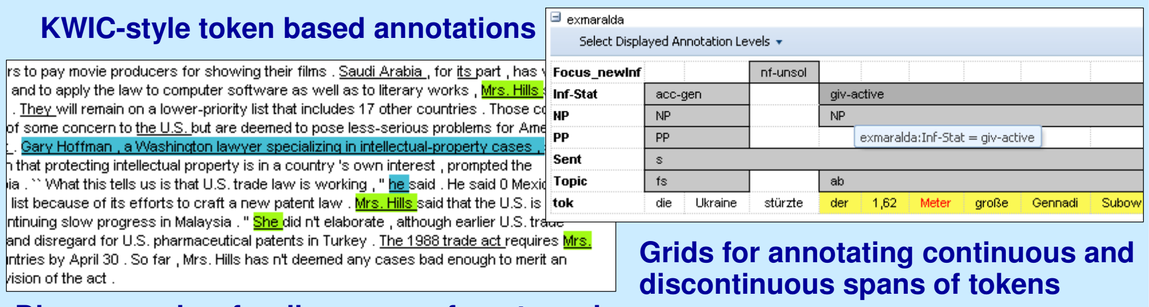
Visualization Modules

- Diverse data from heterogeneous corpora requires multiple visualizations
- Simultaneous querying and visualization of different annotations enables analysis of interdependency between different types of data

ANNIS2 uses extensible Java plug-in based modules for different types of data. Currently supported are:

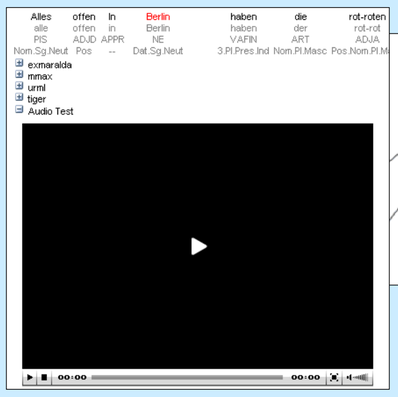


KWIC-style token based annotations

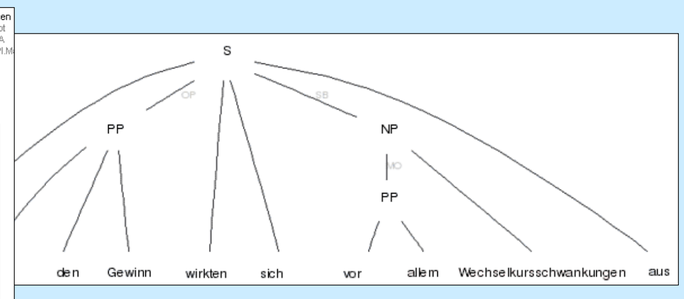


Grids for annotating continuous and discontinuous spans of tokens

Discourse view for discourse referents and coreferent expressions (anaphors/antecedents)



Audio/Video data aligned with token spans



Syntax trees with edge labels and crossing edges

Outlook

For upcoming versions of ANNIS, we will be concentrating on the following expansions:

- Support for the annotation of **subtoken units**
- Expanding AQL with **negation**, with and without implicit existence of nodes
- Visualization and search for **parallel corpora** using alignment operators
- Integration of flexible **statistical functionality** using aggregate functions and user defined mathematical functions
- More **export** / **import** filters to allow modification and update of corpora and subcorpora
- Supporting new data types from more annotation tools, such as **Serengeti** coreference data
- Development of specialized visualizations for supported data types (for example **rhetorical structure** document trees based on **RST** annotations)
- Support for annotations calling external APIs, for example **lexica** or **cartographic resources** such as Google Earth's KML for geographical annotations (e.g. to visualize distributions of dialect features)
- A **shopping-cart** style interface for collecting interesting results and exporting them to a file
- Enabling the embedding of specialized **non-Unicode conformant fonts** for historical data