



# ANNIS3: Towards Generic Corpus Search and Visualization

Thomas Krause, Benjamin Weißenfels, Amir Zeldes and Florian Zipser

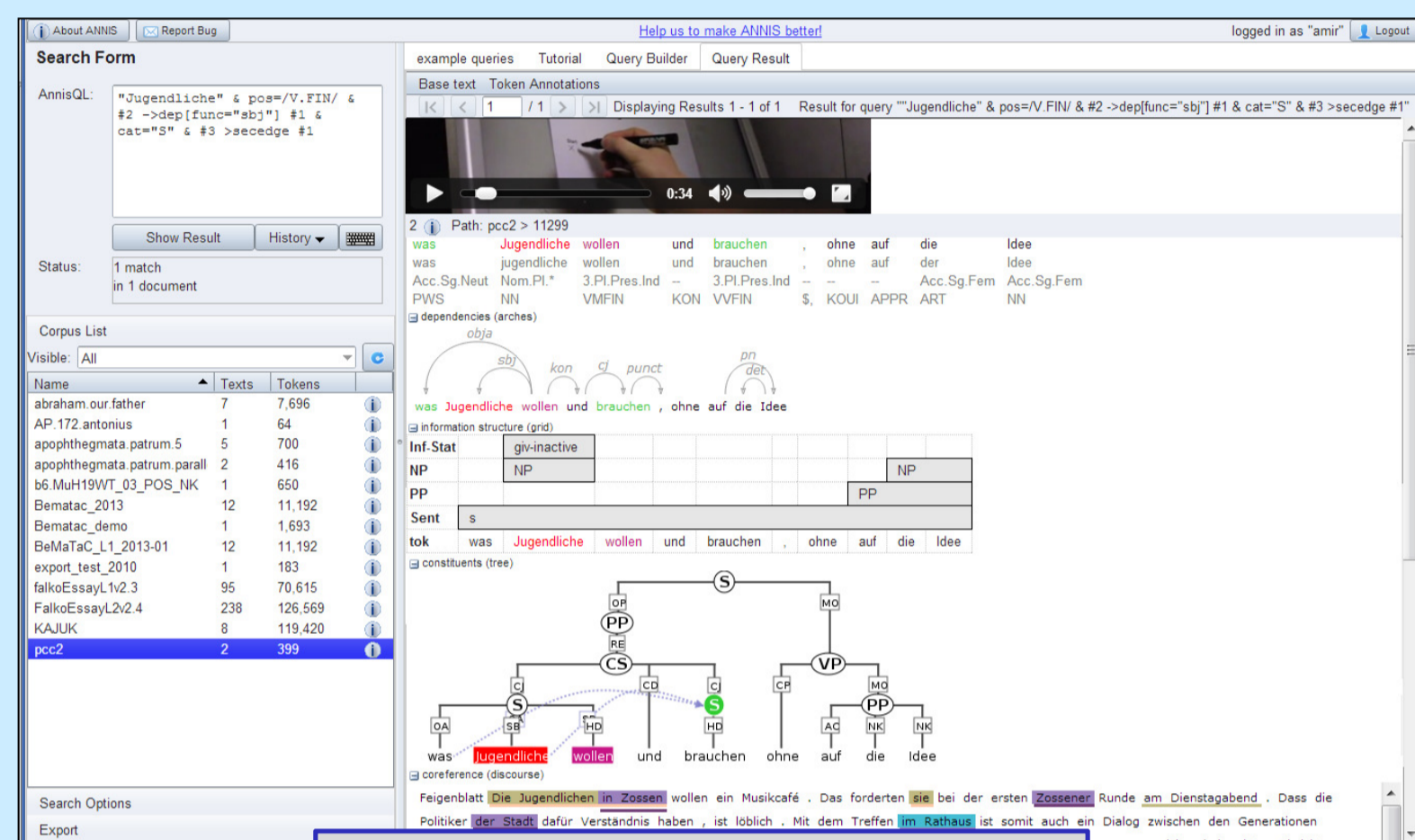


Humboldt-Universität zu Berlin

LAUDATIO

## 1. Challenges

- **Different types of annotations to query and visualize**
  - (Semi-)automatic: multiple taggers, constituency/dependency parsers ...
  - Manual: coreference, information structure, rhetorical structure ...
- **Different types of corpus data to support**
  - **Historical corpora:** diplomatic and normalized text, manuscript structure
  - **Multimodal corpora:** aligned audio/video, multiple overlapping speakers
  - **Parallel corpora:** representing conflicting word & sentence alignment
  - **Learner corpora:** conflicting base text and target hypotheses (Reznicek et al. 2013)
  - And other multilayer corpora: any and every annotation may repeat/conflict with other structures



ANNIS3 Interface

Many corpora violate **assumption of one continuous stream of segments** (multiple languages, speakers, corrected texts...)

- Combinatorial explosion of types: unrealistic to design a system for each
- Reusing the architecture for unique search and visualization applications
- Simplifying the query language (**AQL**) to deal with new structures

## 2. Unified data model and query language

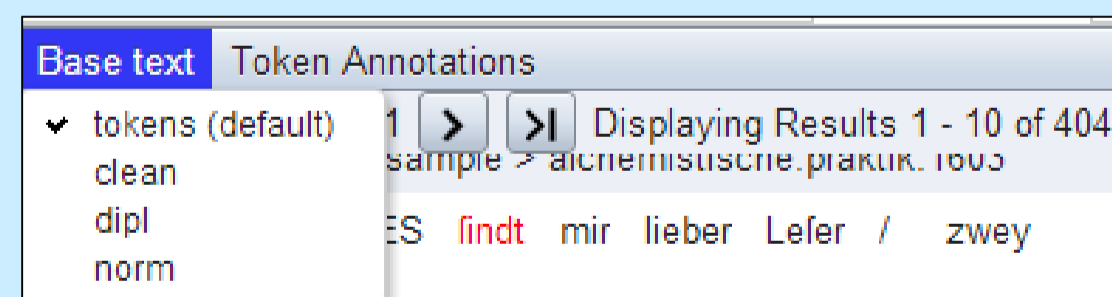
### Dealing with multiple source formats



- Annotations come from multiple formats
  - Convert multiple formats with SaltNPepper (Zipser & Romary 2010)
  - Salt model represents merged data in ANNIS (Zeldes et al. 2009)
  - Reconcile conflicting **segmentations**
- Archive data in PAULA XML (Dipper 2005), a standoff XML format for multi-layer corpora

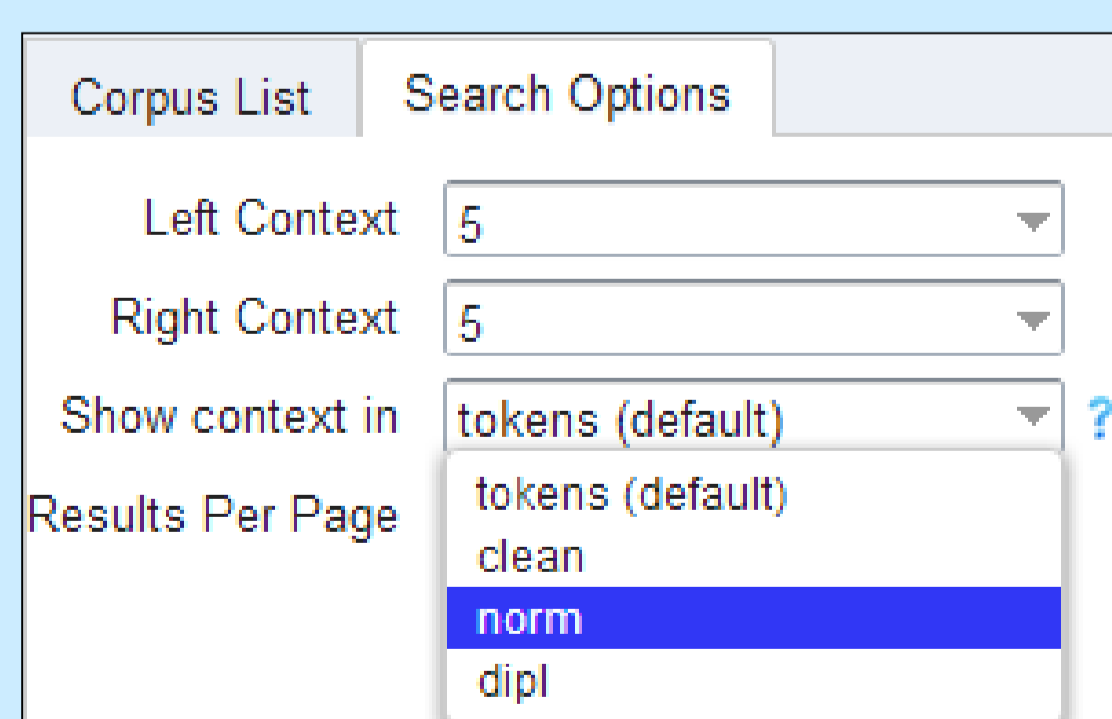
### Segmentations in the new ANNIS3 data model

- Deal with multiple alternative base texts: **one segmentation each**
- Any annotation layer can be a segmentation:
  - Diplomatic/normalized word forms
  - Broad and narrow phonetic transcription
  - Data from different speakers



- Segmentations can be selected as:

- The base text for concordance KWIC views (Key-Word in Context)
- The unit for defining the desired context (e.g.  $\pm 5$  normalized word forms)
- Search criteria for proximity and adjacency in the ANNIS Query Language (**AQL**), using **typed precedence operators**:

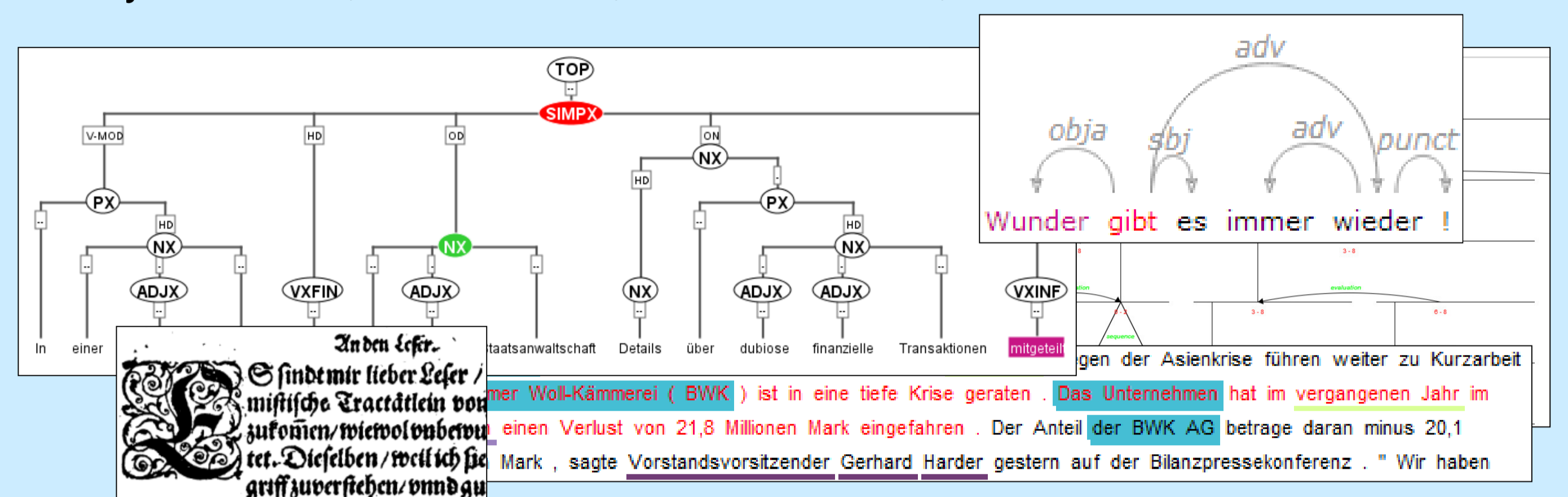


Search consecutive utterances of a speaker (even if others intervene):  
"ja" .instructor "ja" //the instructor says ja twice

Find differently spelled words within 10 diplomatic units in a manuscript:  
/s.\*/ .dipl,1,10 /r.\*/ & //words in s- and f- in 1-10 dipl  
lemma == lemma & //two identical lemmas  
#1 == #3 & //1st word covers 1st lemma  
#2 == #4 //2nd word covers 2nd lemma

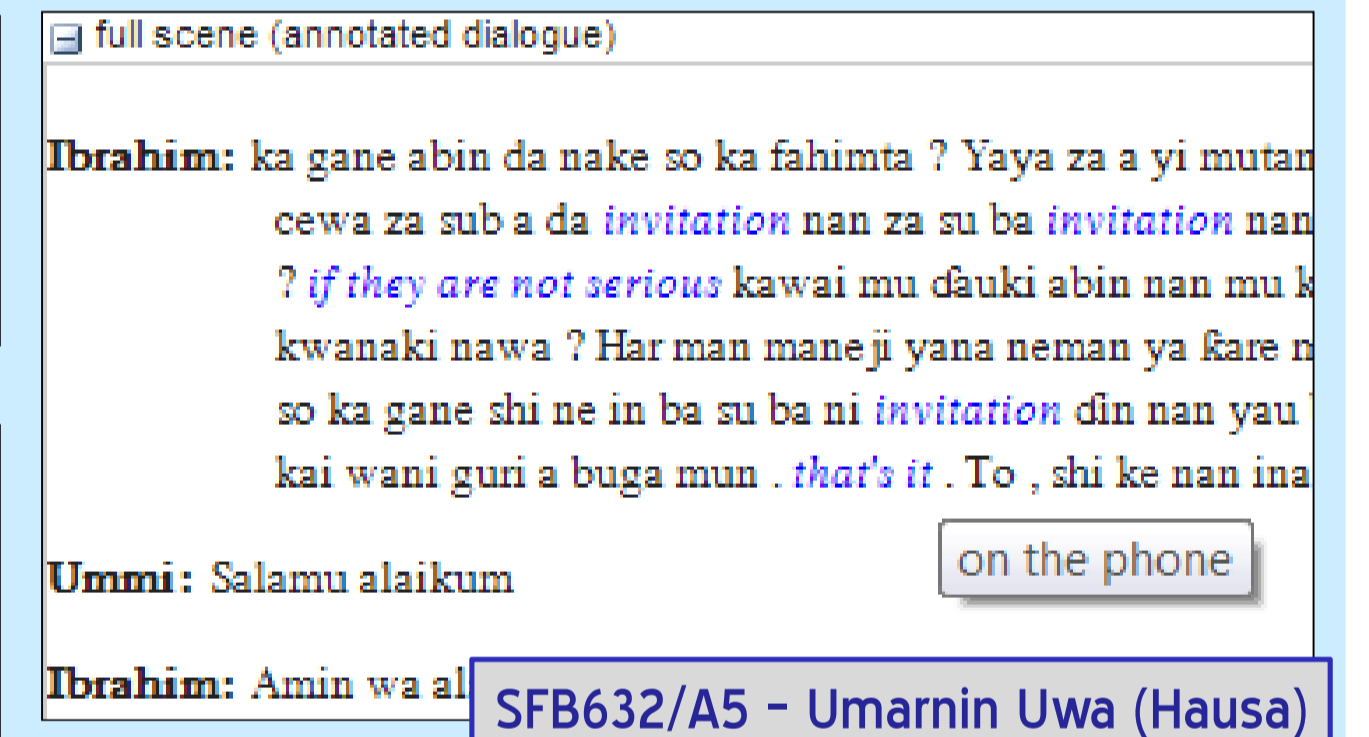
## 3. Reusable, configurable visualizations

- Dedicated visualization are needed for many common data types:
  - Syntax trees, coreference, rhetorical trees, etc.



- But many corpora have unique data types:
  - Alternative views of the base text in digital editions of manuscripts
  - Alternating, non-overlapping text in subtitle corpora, film transcripts
  - Conflicting annotations for different layers of Information structure ...  
→ Impossible to foresee all necessary visualization types
- Approach: Use **annotation triggered style sheets**
  - Expressiveness of HTML5 with flexibility of CSS3
  - Short development cycles from corpus to visualization
  - Implementation as **configuration file** and **CSS file**:

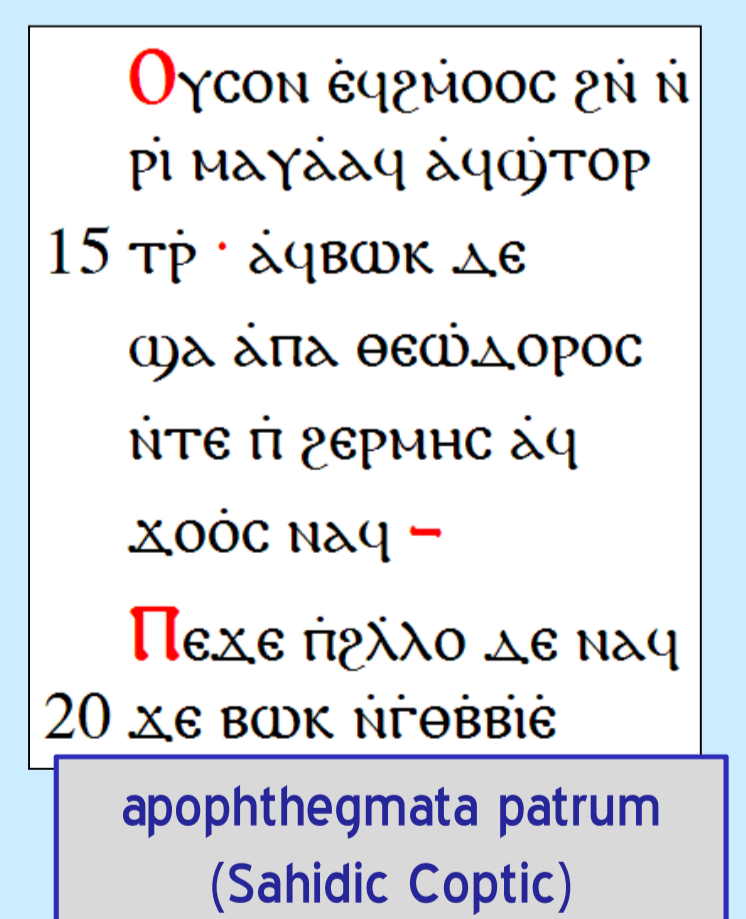
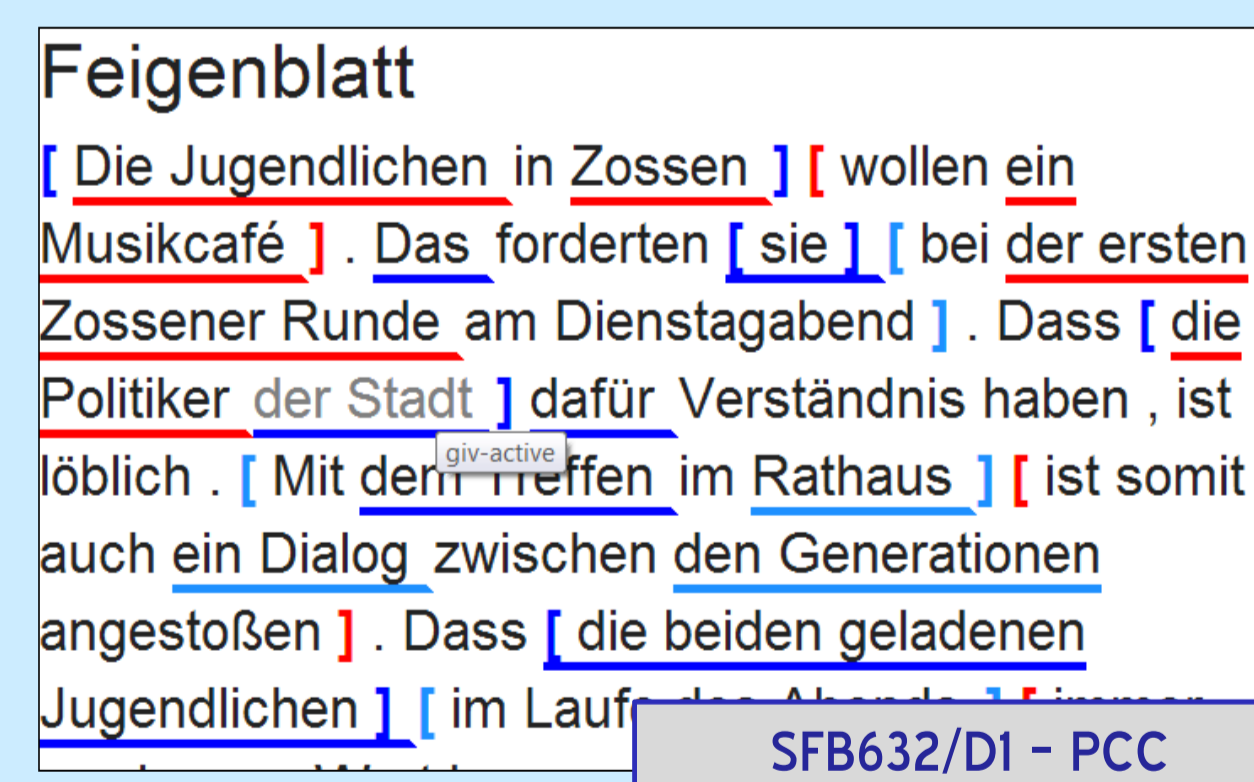
```
tok span; style="word" value
spkr div:spkr; style="spk" value
lang span:lang; style="lang" value
info t:title; style="info" value
```



```
.word:after{content: " ";}
div.spk{display: block; padding-top: 6px; padding-bottom: 6px; text-indent: -65px; padding-left: 65px}
div.spk:before{content: attr(speaker) " "; font-weight: bold}
.lang{color: blue; font-style: italic}
.info:hover{color: red}
```

## 4. Applications and future directions

- A variety of dedicated visualizations can be developed with little code
- Visualization of Information Structure in PCC (Stede 2004)
- Digital manuscript editions for Coptic (Projects KOMeT/SCRIPTORIUM):  
<http://coptic.pacific.edu/>



### Some planned extensions include:

- Adding matching javascript files for more **interactive visualizers**
- Visualizer-triggered searching (click on words, jump between **linked results**)
- **Aggregate visualizers** based on results from multiple documents/corpora

### References

Dipper, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, 39–50.  
 Reznicek, M./Lüdeling, A./Hirschmann, H. 2013. Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A./Ballier, N./Thompson, P. (eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 101–124.  
 Stede, M. 2004. The Potsdam Commentary Corpus. In Webber, B./Byron, D. K. (eds.) *Proceeding of the ACL-04 Workshop on Discourse Annotation*. Barcelona, 96–102.  
 Zeldes, A./Ritz, J./Lüdeling, A./Chiaros, C. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.  
 Zipser, F./Romary, L. 2010. A Model Oriented Approach to the Mapping of Annotation Formats using Standards. *Language Resource & Language Technology Standards, LREC 2010*. Malta, 7–18.