# Interactive analysis of multi-layer linguistic corpora with ANNIS

Florian Zipser*, Thomas Krause*, Arne Neumann+
*Humboldt-Universität zu Berlin, +Universität Potsdam

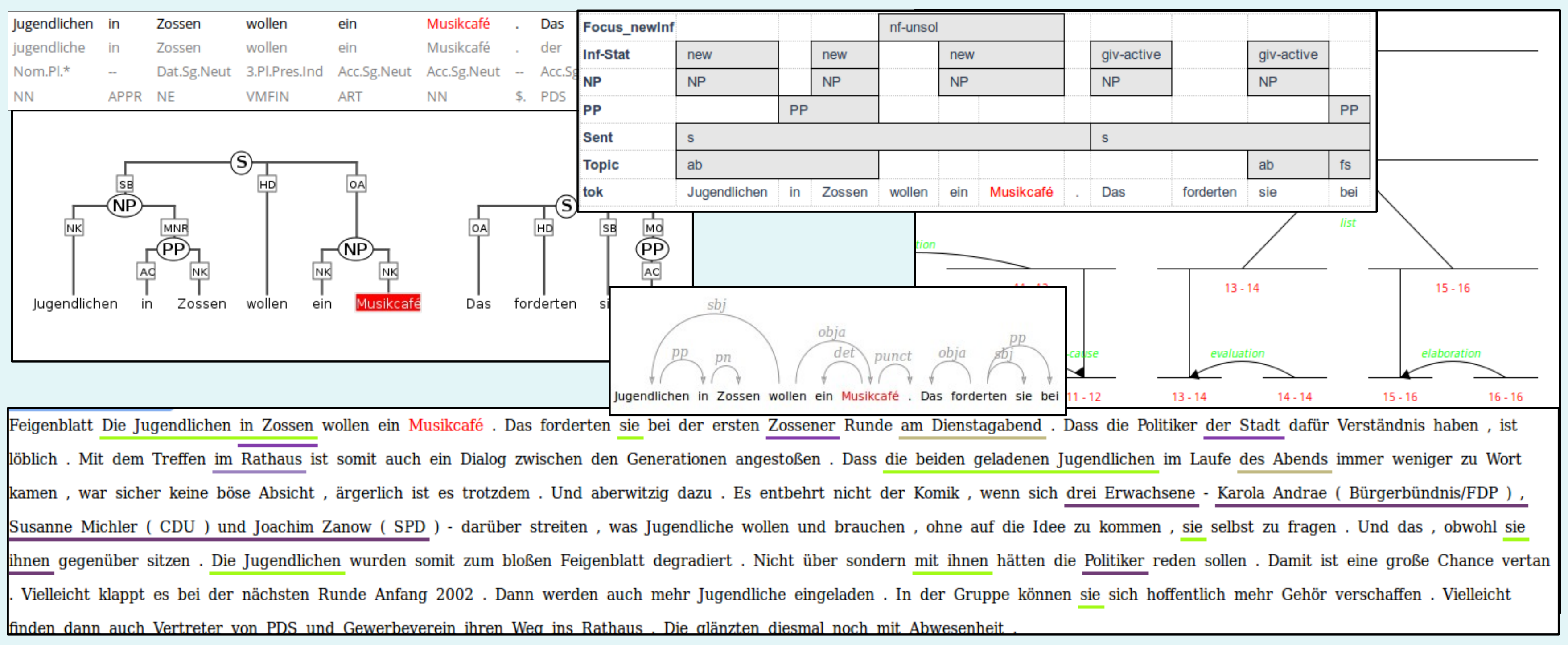
http://annis-tools.org

## Goal

- Higher usability through simplified query syntax (AQL)
- Improvements of AQL for new operators: Equality and Inequality for values, OR Operator and explicit variable naming
- Simple statisitcal evaluations direclty in ANNIS with frequency analysis

## Background

- linguistic phenomena can be spread over different annotation layers, only the combination of annotations will find them
- several multi-layer corpora have been developed for this purpose: TüBa-D/Z (Telljohann et al. 2009), PCC (Stede & Neumann 2014) or Falko (Reznicek et al. 2012)
- annotation layers differ in content and structure: token annotations, constituency trees, dependency trees, rhetorical structure annotation, spoken data, etc.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Jugendlichen | in | Zossen | wollen | ein | Musikcafé | . | Das |
| jugendliche | in | Zossen | wollen | ein | Musikcafé | . | der |
| Nom.Pl.* | -- | Dat.Sg.Neut | 3.Pl.Pres.Ind | Acc.Sg.Neut | Acc.Sg.Neut | -- | Acc.Sg |
| NN | APPR | NE | VMFIN | ART | NN | $. | PDS |

| Focus_newInf | | | | nf-unsol | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Inf-Stat | new | | new | | new | | | giv-active | | giv-active | |
| NP | NP | | NP | | NP | | | NP | | NP | |
| PP | | PP | | | | | | | | | PP |
| Sent | s | | | | | | | s | | | |
| Topic | ab | | | | | | | | | ab | fs |
| tok | Jugendlichen | in | Zossen | wollen | ein | Musikcafé | . | Das | forderten | sie | bei |

Feigenblatt Die Jugendlichen in Zossen wollen ein Musikcafé . Das forderten sie bei der ersten Zossener Runde am Dienstagabend . Dass die Politiker der Stadt dafür Verständnis haben , ist löblich . Mit dem Treffen im Rathaus ist somit auch ein Dialog zwischen den Generationen angestoßen . Dass die beiden geladenen Jugendlichen im Laufe des Abends immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem . Und aberwitzig dazu . Es entbehrt nicht der Komik , wenn sich drei Erwachsene - Karola Andrae ( Bürgerbündnis/FDP ) , Susanne Michler ( CDU ) und Joachim Zanow ( SPD ) - darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl sie ihnen gegenüber sitzen . Die Jugendlichen wurden somit zum bloßen Feigenblatt degradiert . Nicht über sondern mit ihnen hätten die Politiker reden sollen . Damit ist eine große Chance vertan . Vielleicht klappt es bei der nächsten Runde Anfang 2002 . Dann werden auch mehr Jugendliche eingeladen . In der Gruppe können sie sich hoffentlich mehr Gehör verschaffen . Vielleicht finden dann auch Vertreter von PDS und Gewerbeverein ihren Weg ins Rathaus . Die glänzten diesmal noch mit Abwesenheit .

→ **We need a unified search and visualization system**

## ANNIS

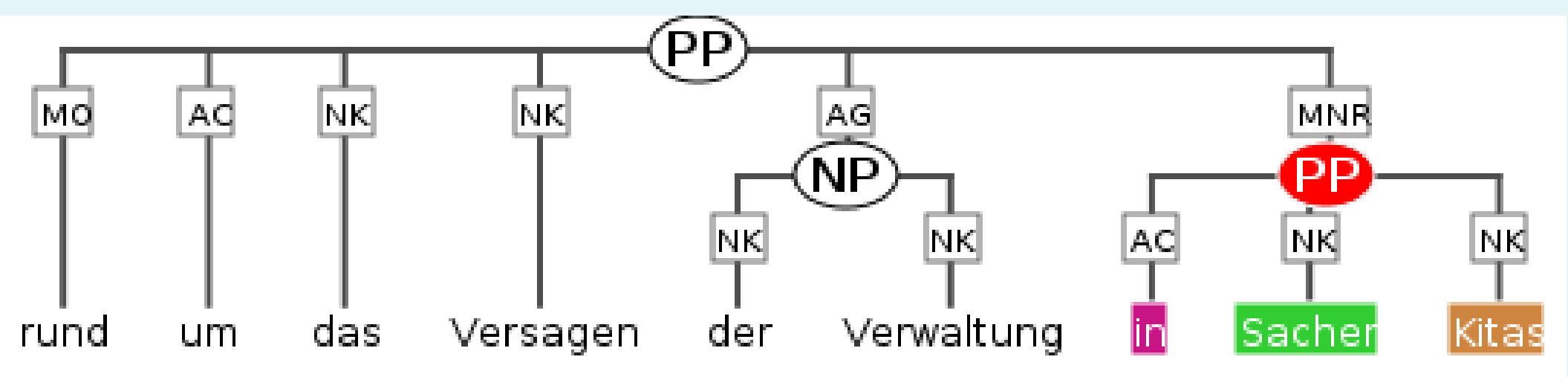
- not limited to a specific type of annotation
- not limited to a single corpus
- same query language for different corpora (AQL)
- specialized and configurable visualizations for different annotation layers
- support for existing corpora from many formats via SaltNPepper (Zipser & Romary 2010)
- export for further statistical evaluation: CSV, plain text or ARFF (WEKA (Hall et al. 2009))

## Query language improvements

- **simplified query syntax (better readability)**
  *Prepositional phrase with a complex preposition (preposition + noun + noun)*

  - so far*:

```
cat="PP" & pos & pos="NN" & pos="NN"
& #1 >[func="AC"] #2
& #2 . #3
& #3 . #4
& #1 >[func="NK"] #3
& #1 >[func="NK"] #4
```



  - simplified*:

```
cat="PP" >[func="AC"] pos . pos="NN" . pos="NN"
& #1 >[func="NK"] #3
& #1 >[func="NK"] #4
```

- **search for equal/different value**
  *two same/different part-of-speech tags, one directly following the other*

```
pos . pos & #1 == #2

pos . pos & #1 != #2
```

| | |
|---|---|
| ersten | Zossener |
| erster | Zossener |
| Pos.Dat.Sg.Fem | Pos.*.*.* |
| ADJA | ADJA |

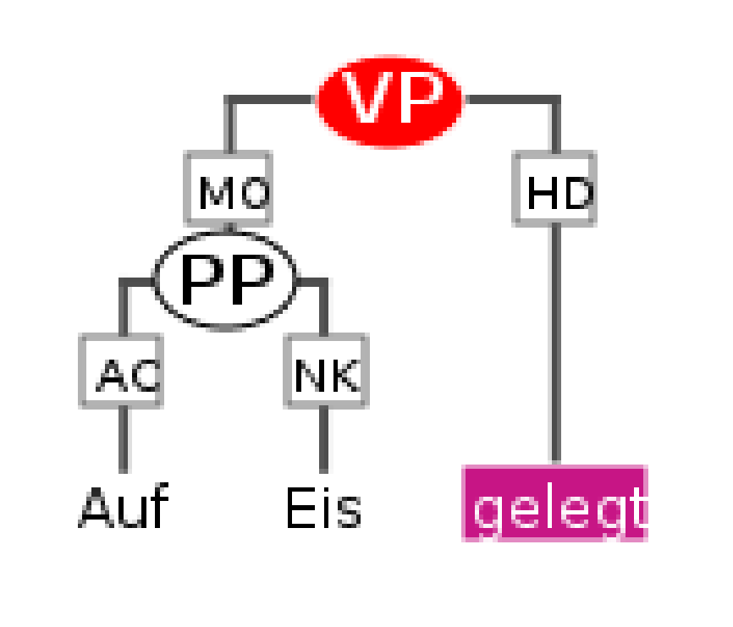
* queries use the TIGER annotation scheme

## Query language improvements

- **OR operator and variable naming**

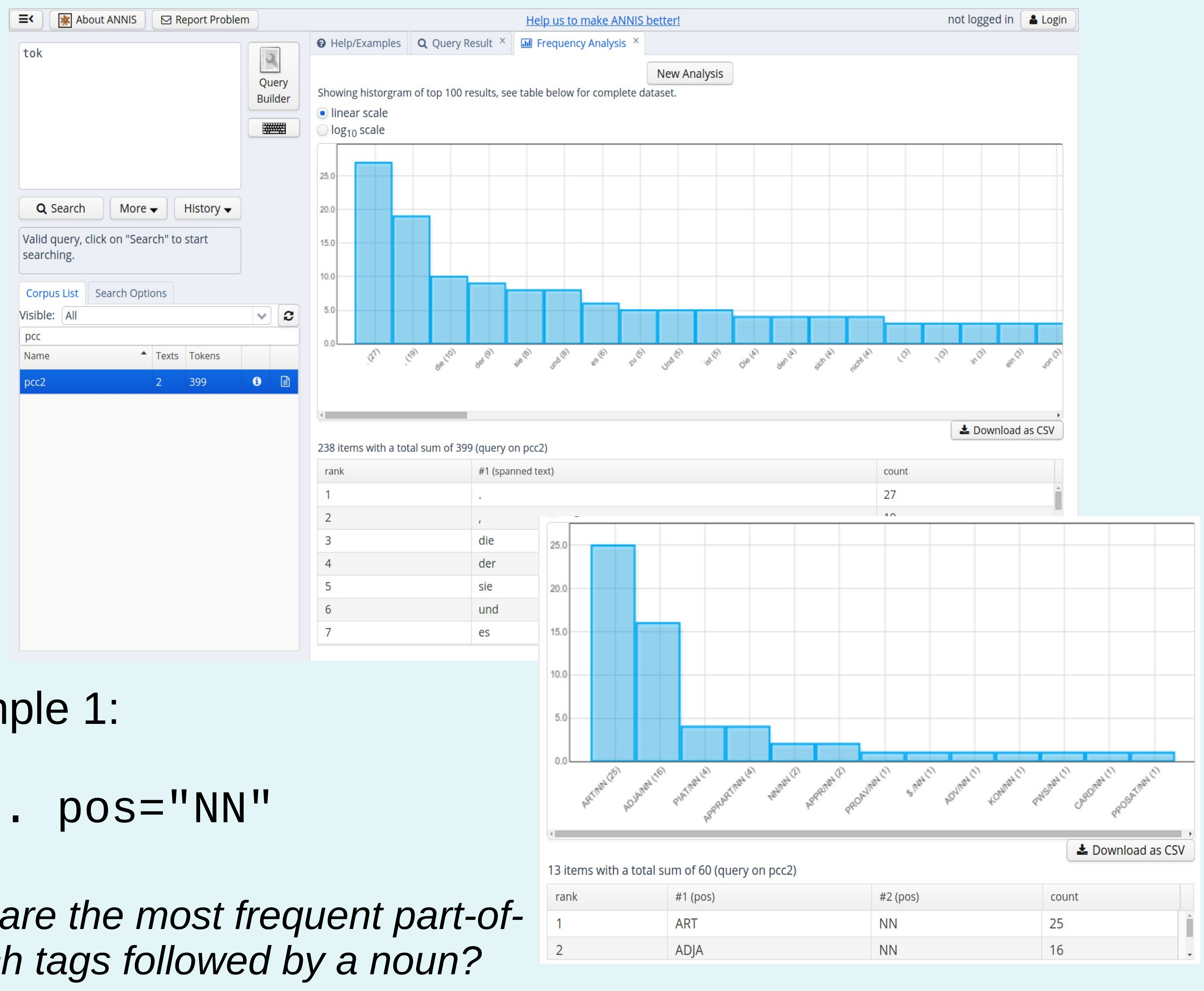
*A verbal phrase, which dominates either an adjective or the token „gelegt“ which is often misclassified by automatic POS-taggers*

```
a#cat="VP"
&(b#pos="ADJA" | b#"gelegt")
& #a > #b
```



## Frequency analysis

- **frequency analysis directly in ANNIS**



tok

Valid query, click on "Search" to start searching.

pcc

| Name | Texts | Tokens |
|---|---|---|
| pcc2 | 2 | 399 |

Showing historgram of top 100 results, see table below for complete dataset.

238 items with a total sum of 399 (query on pcc2)

| rank | #1 (spanned text) | count |
|---|---|---|
| 1 | . | 27 |
| 2 | , | |
| 3 | die | |
| 4 | der | |
| 5 | sie | |
| 6 | und | |
| 7 | es | |

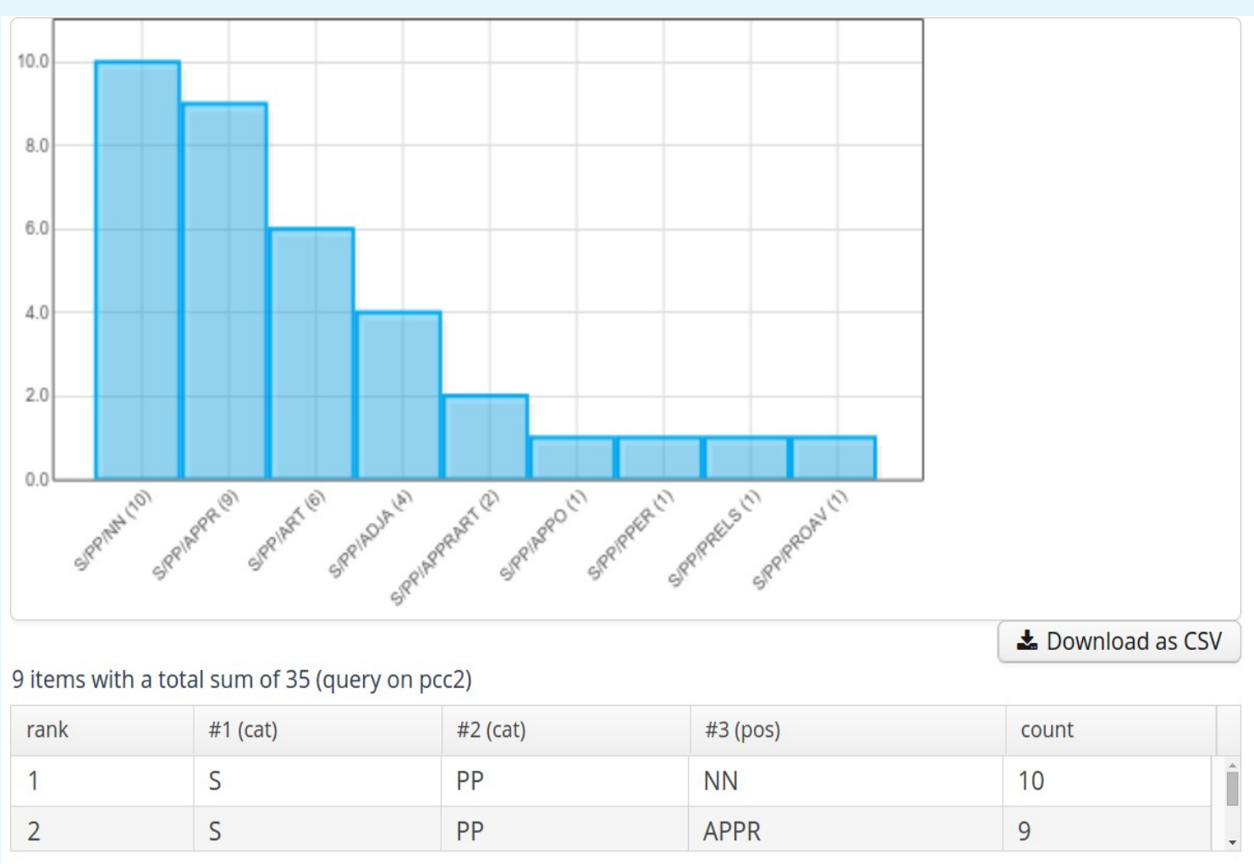
13 items with a total sum of 60 (query on pcc2)

| rank | #1 (pos) | #2 (pos) | count |
|---|---|---|---|
| 1 | ART | NN | 25 |
| 2 | ADJA | NN | 16 |

Example 1:

```
pos . pos="NN"
```

*What are the most frequent part-of-speech tags followed by a noun?*



9 items with a total sum of 35 (query on pcc2)

| rank | #1 (cat) | #2 (cat) | #3 (pos) | count |
|---|---|---|---|---|
| 1 | S | PP | NN | 10 |
| 2 | S | PP | APPR | 9 |

Example 2:

```
cat="S" > cat="PP" > pos
```

*What are the most frequent part-of-speech tags in a prepositional phrase which is in a sentence?*

## Outlook

- meta data in frequency analysis
- syntax highlighting for AQL
- implement a new and much faster/scalable graph database backend as a replacement for the relational database PostgreSQL

## References

- **M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. 2009**. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
- **M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K.Schmidt,H. Hirschmann, T. Andreas. 2012**. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.
  https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01
- **M. Stede, A. Neumann 2014.** Potsdam Commentary Corpus 2.0:
  Annotation for Discourse Research. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Island.
- **H. Telljohann, E. W. Hinrichs, S. Kübler, H. Zinsmeister, K. Beck. 2009.** Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Universität Tübingen Seminar für Sprachwissenschaft.
- **F. Zipser, L. Romary. 2010.** A Model Oriented Approach to the Mapping of Annotation Formats using Standards. Workshop Language Resource & Language Technology Standards, LREC 2010. Malta, 7-18. http://u.hu-berlin.de/saltnpepper

Web: http://annis-tools.org
Email: annis-admin@ling.uni-potsdam.de