Carolin Odebrecht &
Florian Zipser
Humboldt-Universität zu Berlin

ANNIS workshop

2014-08-26

A brief introduction

- Search and Visualization in Multilayer Linguistic Corpora
  - Imports existing corpora
    - Corpora already have to be annotated, ANNIS only uses what's there
    - No NLP!

A brief introduction

- Search and Visualization in Multilayer Linguistic Corpora
  - Makes corpora searchable
    - One query language for all corpora (AQL)
    - Abstraction over linguistic data necessary
    - But: Corpora have different annotations → query has to match the annotations

A brief introduction

- Search and Visualization in Multilayer Linguistic Corpora
  - Displays corpora
    - Many visualizations available
    - Corresponding to type of annotation (syntactic trees, phrase trees (RST), grids, coreferences ...)

A brief introduction

- What ANNIS cannot do
  - Does not know how to speak natural language

    → so you have to learn AQL

A brief introduction

- What ANNIS cannot do
  - Does not know how to speak natural language

    → so you have to learn AQL
  - ANNIS does not know any semantics

    → „NN“, „NP“, „sentence“, „word“, „my favorite annotation“ … are just sequences of characters

A brief introduction

- What ANNIS cannot do
  - Does not know how to speak natural language

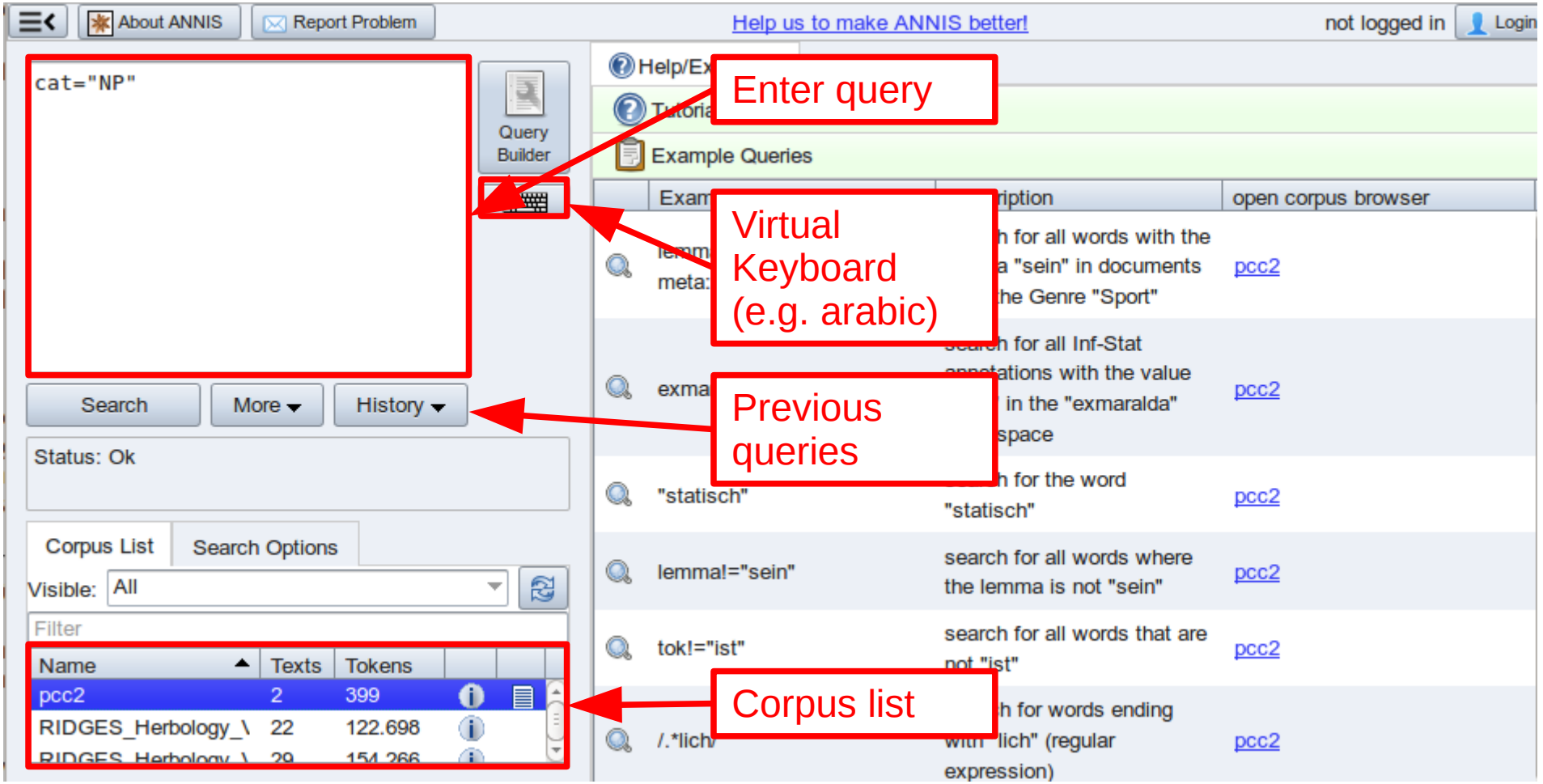
    → so you have to learn AQL
  - ANNIS does not know any semantics

    → „NN“, „NP“, „sentence“, „word“, „my favorite annotation“ … are just sequences of characters
  - You need to be exact

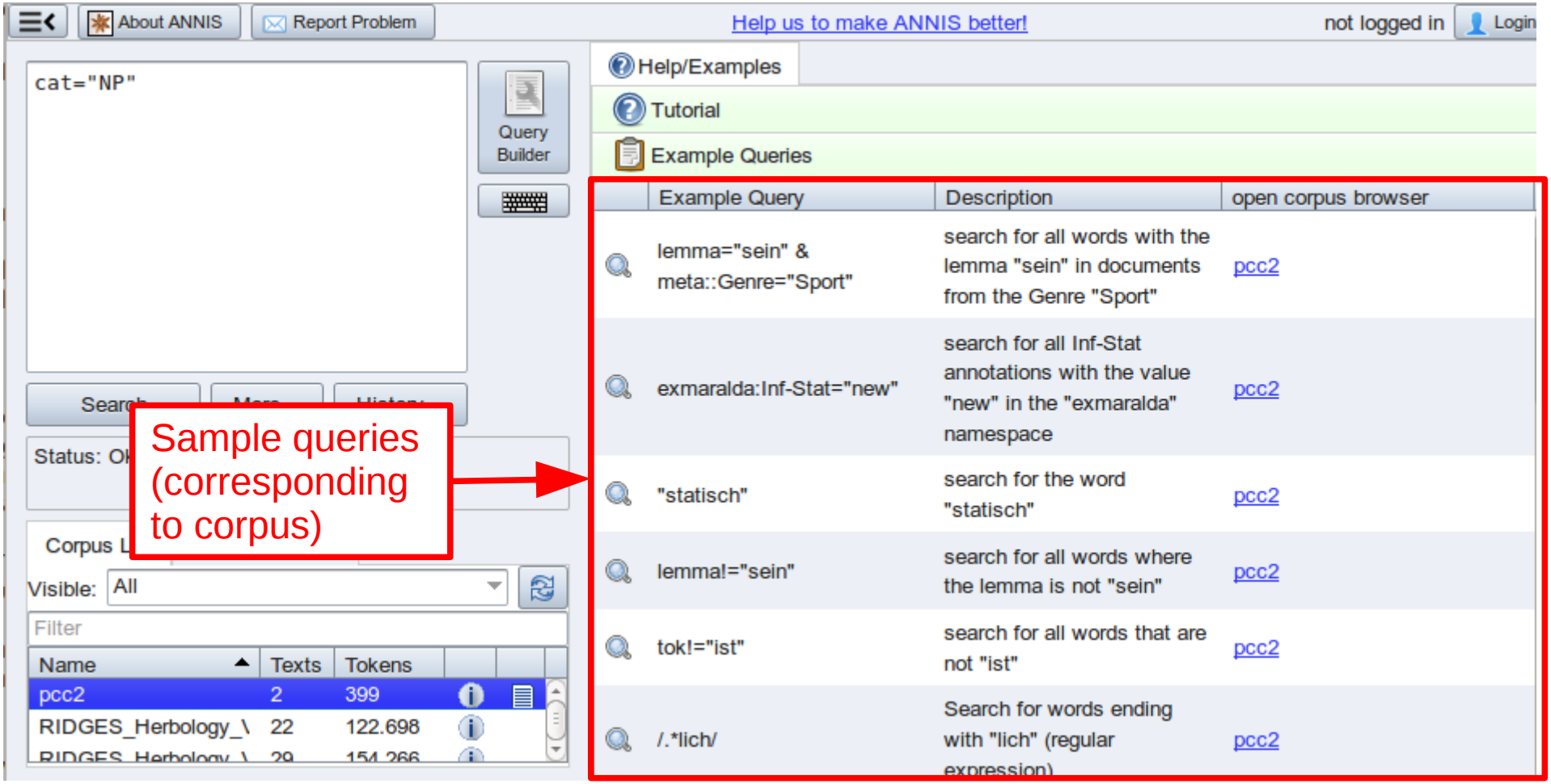
    → e.g. „POS“ != „pos“ and „NN“ != „NN “ (regard the blank)

ANNIS basics

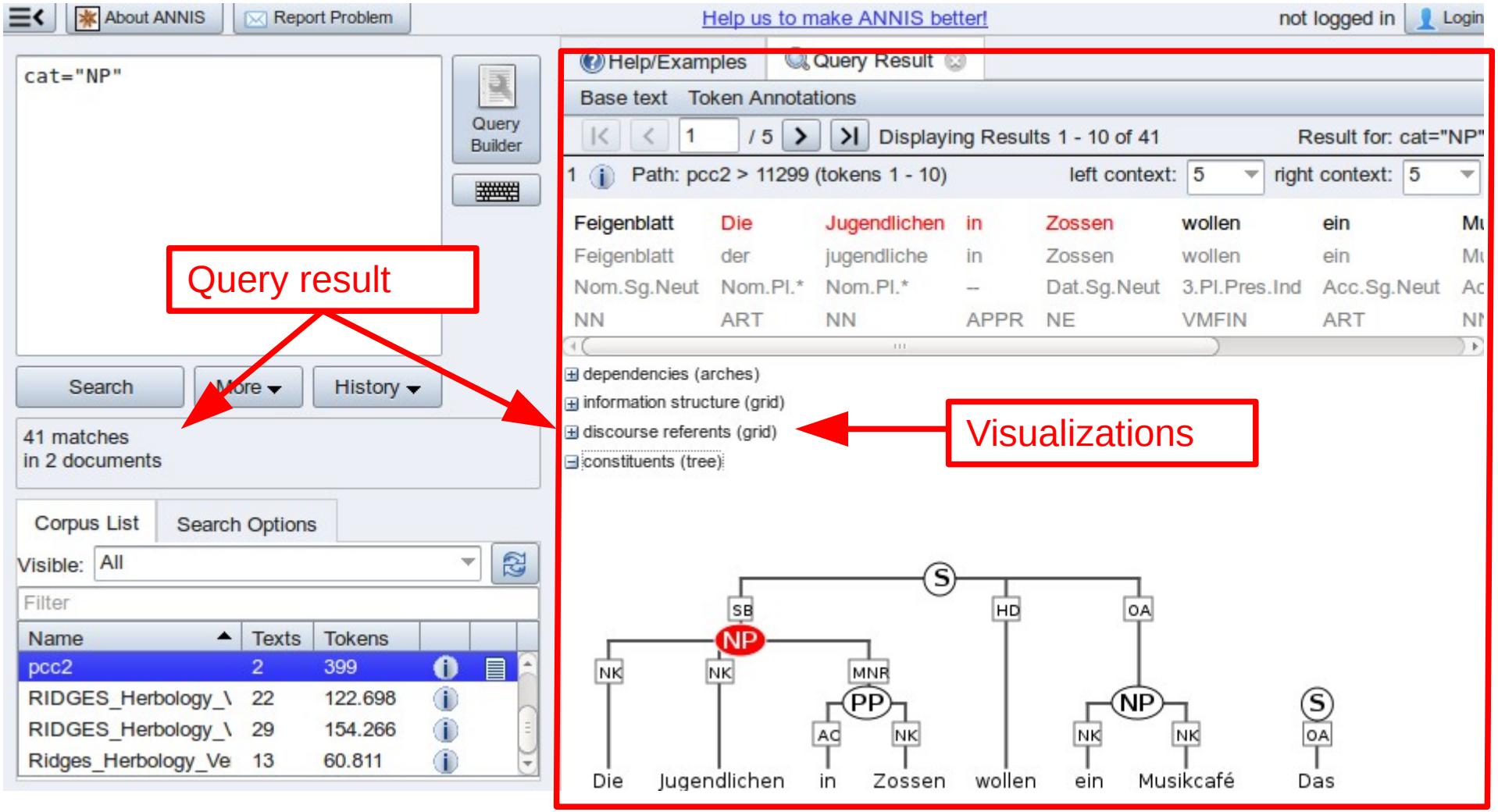
ANNIS: Search and Visualization in Multilayer Linguistic Corpora



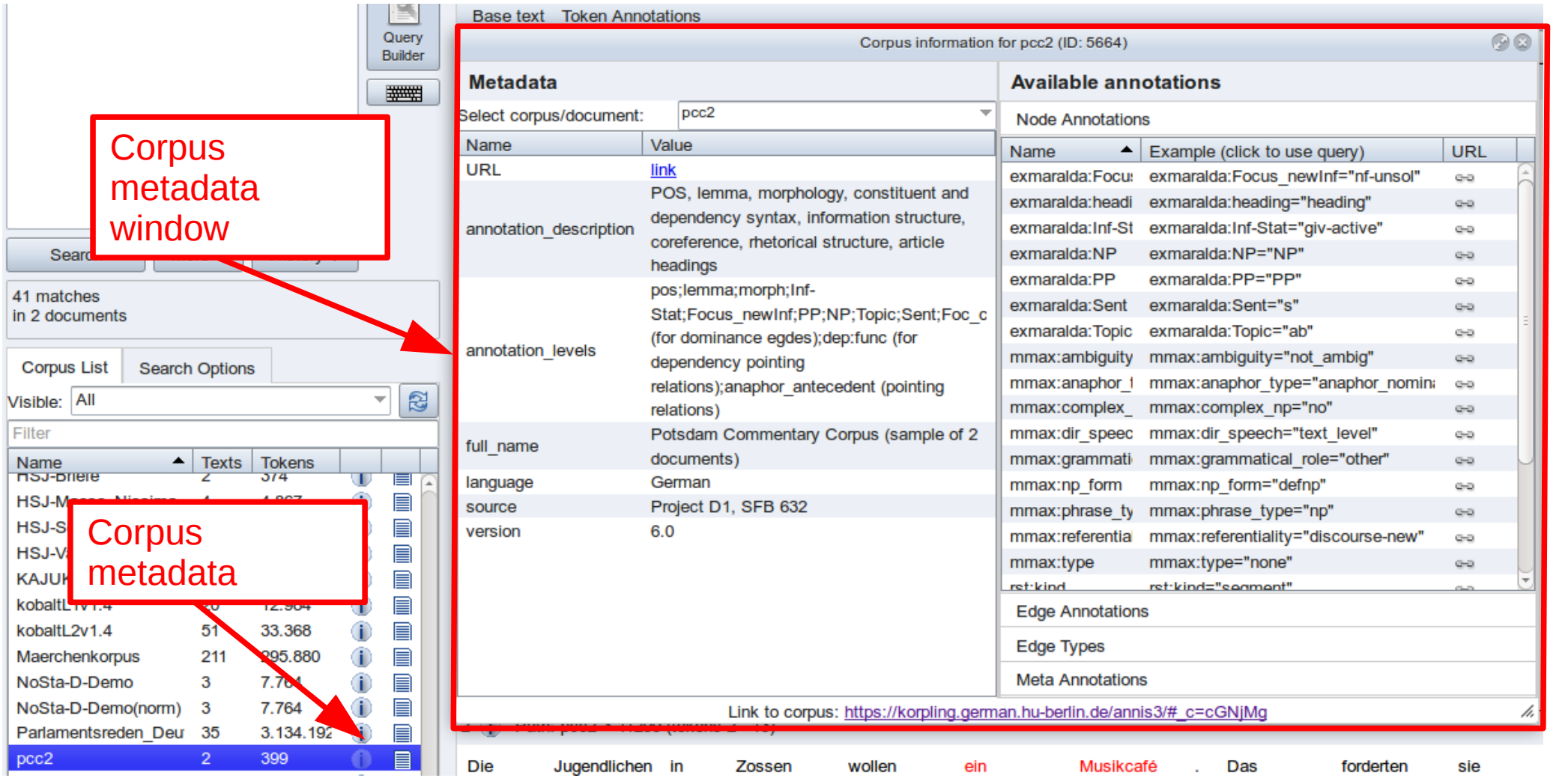
ANNIS: Search and Visualization in Multilayer Linguistic Corpora



ANNIS: Search and Visualization in Multilayer Linguistic Corpora



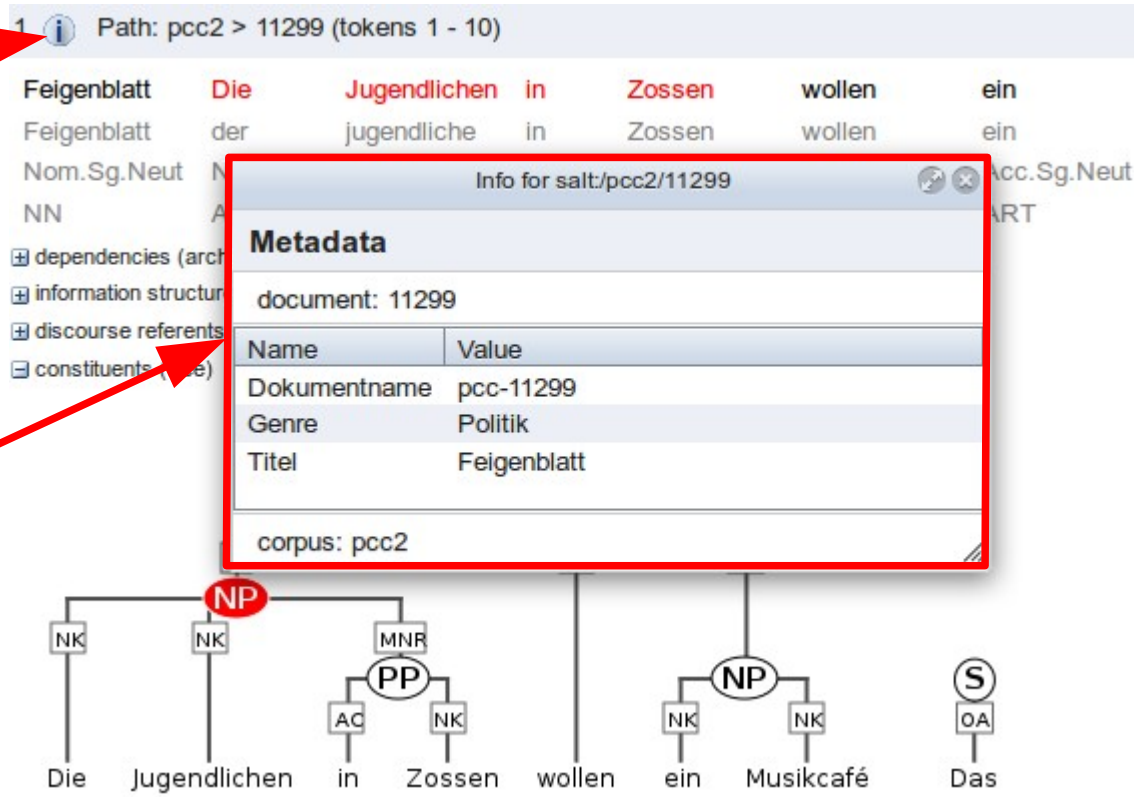
ANNIS: Search and Visualization in Multilayer Linguistic Corpora



ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Document metadata

Document metadata window



ANNIS basics

- Basic principles of AQL (ANNIS Query Language)
  - Attributes and values
    - Searching for exact character sequences
    - Searching for patterns
  - Combinatory search

Demo corpus

- Corpus for demonstration: pcc2 (a sub corpus of pcc)

  https://korpling.german.hu-berlin.de/annis3/#_c=cGNjMg

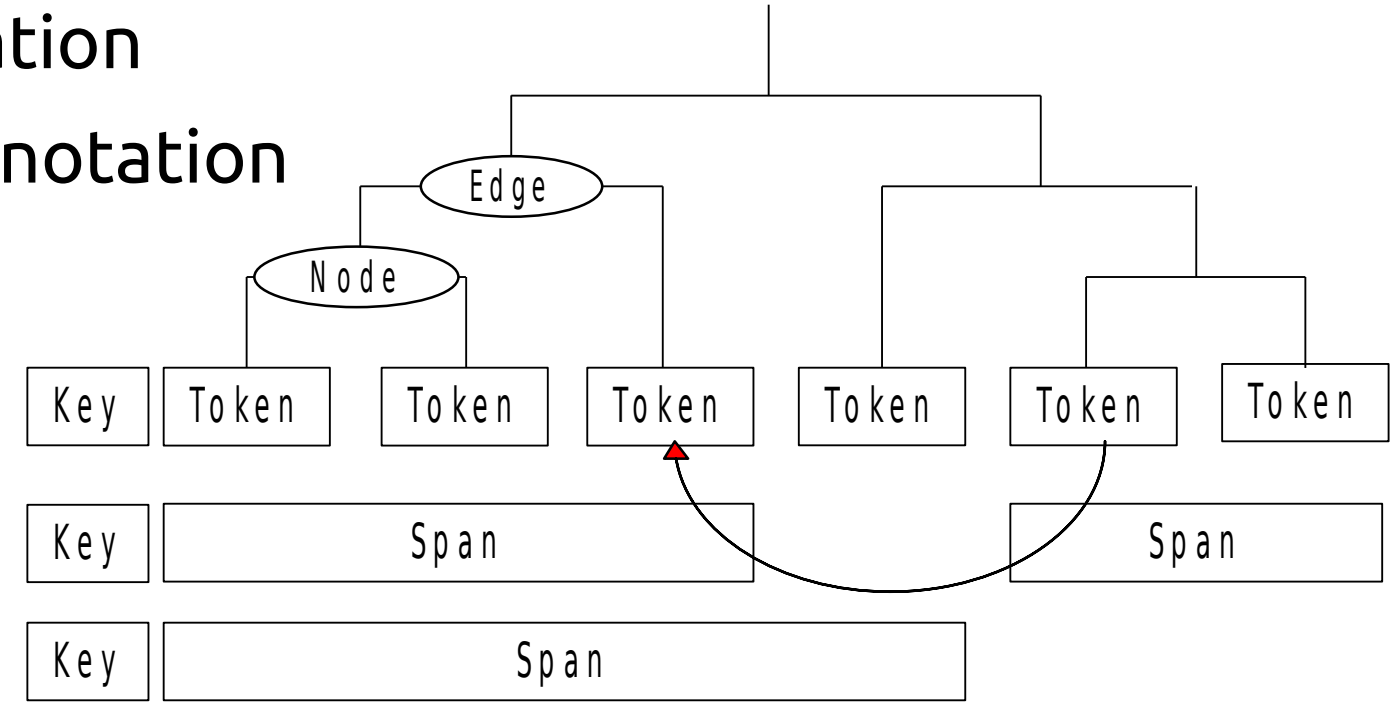
- Potsdam Commentary Corpus
  - German Newspaper commentaries

    'Märkische Allgemeine Zeitung'

    https://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html
  - Multiple annotations

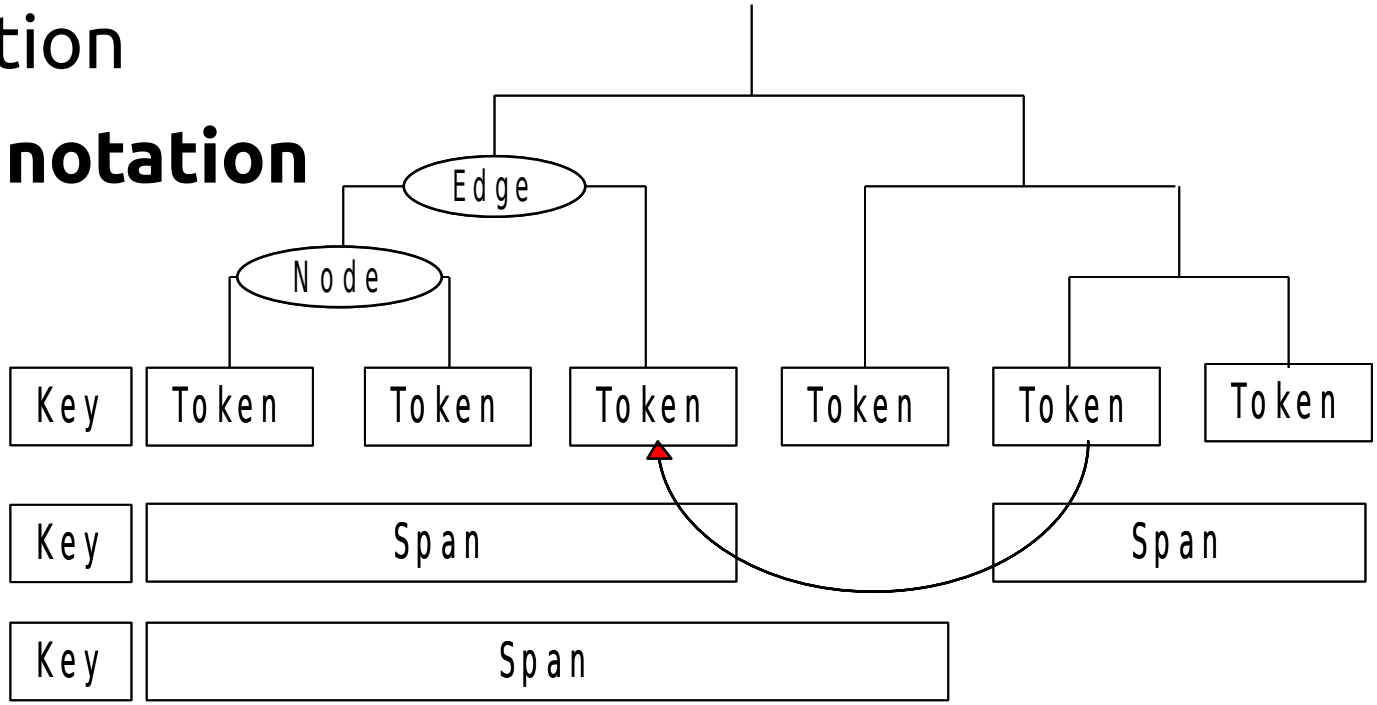
ANNIS basics

- Different types of annotations
  - Token annotation
  - Span annotation
  - Pointing relation
  - Hierarchy annotation (trees)



ANNIS basics

- Different types of annotations
  - **Token annotation**
  - **Span annotation**
  - Pointing relation
  - **Hierarchy annotation (trees)**



Exact word forms

- Token annotation
  - Exact sequence

searching for a word form

"Jugendlichen"

"jugendlichen"

Exact word forms

- Token annotation
  - Exact sequence

searching for a word form

"Jugendlichen" 3 hits

"jugendlichen" 0 hits

→ tok="jugendlichen"

Exact token annotation

- Token annotation
  - Exact sequence

    searching for an **exact** part of speech tag

    ```
    pos = "NN"
    ```

    attribute value

  - Attributes can have more than one value
  - Searching for all values of an attribute

Exact token annotation

- Token annotation
  - Exact sequence

searching for an exact part of speech tag

pos="NN"

pos="ADJA"

Exact token annotation

- Token annotation
  - Exact sequence

searching for an exact part of speech tag

`pos="NN"` 62 hits

`pos="ADJA"` 18 hits

searching for all values of an attribute

`pos` 399 hits

Exact span annotation

- Span annotation
  - Exact sequence

searching for sentences

```
Sent="s"
```

Exact span annotation

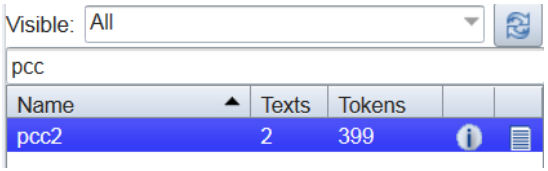
- Span annotation
  - Exact sequence

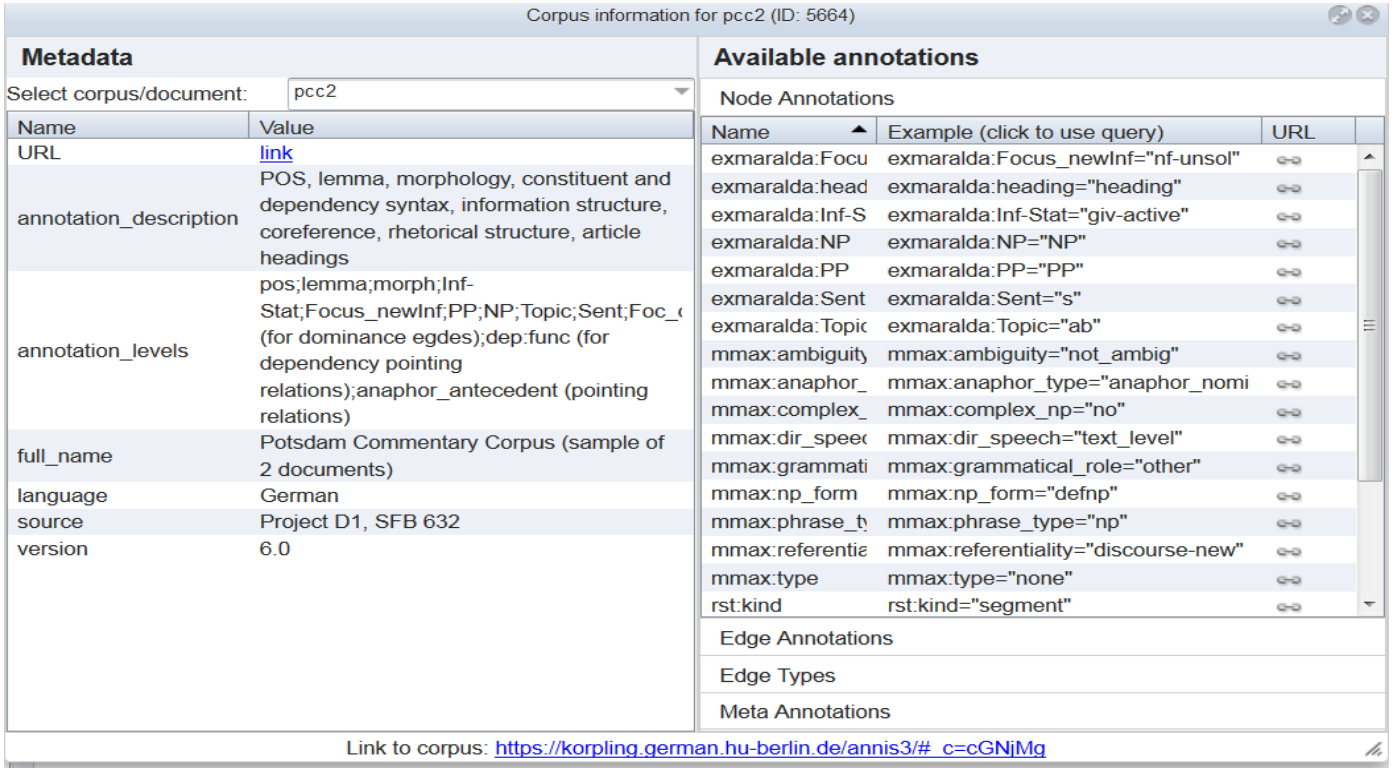
searching for sentences

Sent="s" 28 hits

Metadata

- Sent="s" 28 hits
  - necessary to know which annotations are in a corpus





Pattern

- Token annotation
  - Patterns
    - . matches any single character
    - * zero or more of the preceding element

searching for the beginning a of word

/Jugend.*/

/jugend.*/

Pattern

- Token annotation
  - Patterns
    - . matches any single character
    - * zero or more of the preceding element

searching for the beginning a of word

/Jugend.*/ 5 hits ("Jugendlichen" 3 hits)

Jugendlichen Jugendliche

/jugend.*/ 0 hits ("jugendlichen" 0 hits)

Pattern

- Token annotation
  - patterns

searching for **all** nouns

`pos=/N./` includes NN & NE

searching for **all** adjectives

`pos=/ADJ./` includes ADJA & ADJD

Pattern

- Token annotation
  - patterns

searching for **all** nouns

`pos=/N./` 73 hits (pos="NN" 62 hits)

searching for **all** adjectives

`pos=/ADJ./` 32 hits (pos="ADJA" 18 hits)

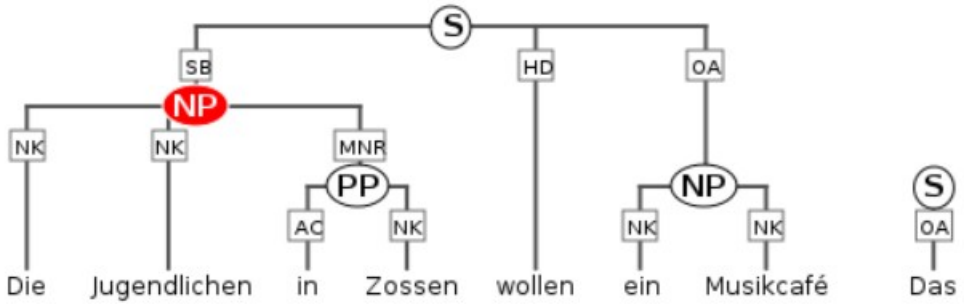
Relations between annotations

- Span annotation

  searching for all NPs

  cat="NP" 41 hits (pos="NN" 62 hits)

  e.g. *Die Jugendlichen in Zossen*



Relations between annotations

- Relations between attributes

  searching for all NPs which contain a preposition

  cat="NP" 41 hits

  pos="APPR" 19 hits

  e.g. *Die Jugendlichen in Zossen*

  → no relation between the two information!

Relations between annotations

- Relations between attributes

  searching for **all** NPs which contain a preposition

cat="NP"	#1
pos="APPR"	#2

  e.g. *Die Jugendlichen in Zossen*

  → NP includes APPR

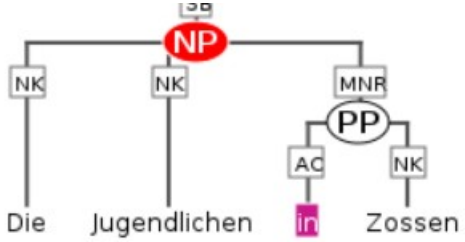
Relations between annotations

- Relations between attributes

  searching for **all** NPs which contain a preposition

```
cat="NP"    &
pos="APPR"        &
#1_i_#2
```

  e.g. *Die Jugendlichen in Zossen*



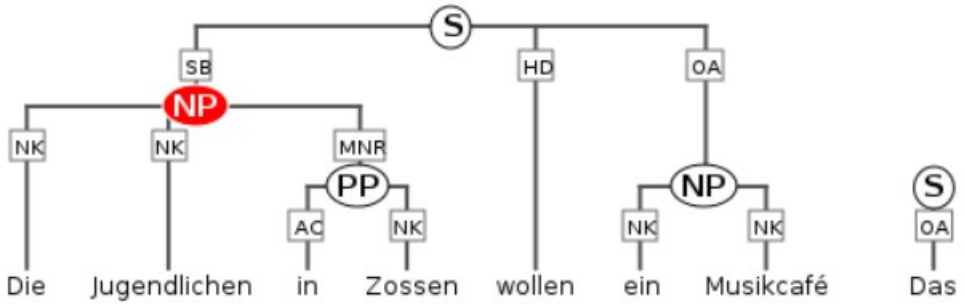
Hierarchy relations

- Relations between attributes

  searching for **all** NPs which are **objects**

  cat="NP"

  e.g. *Die Jugendlichen in Zossen* -->subject!

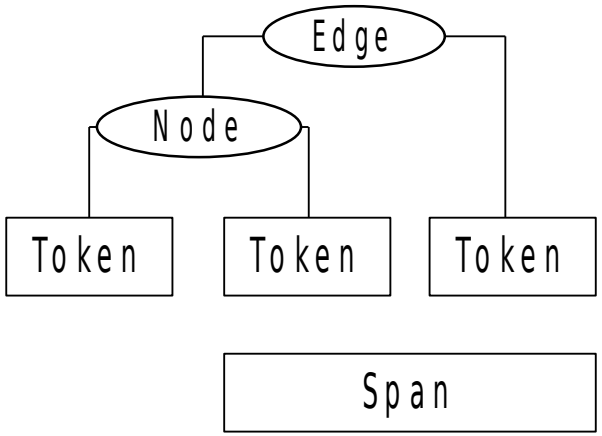


Hierarchy relations

- Relations between attributes

  searching all **NPs** which are **objects**

  - **NP →** node annotation
  - **OA →** edge annotation



Hierarchy relations

- Relations between attributes

  searching all NPs which are objects

  cat="NP"

  the syntactic function in the tree

  func="OA"

  → Note: At least there are two elements which relate in a way to each other!

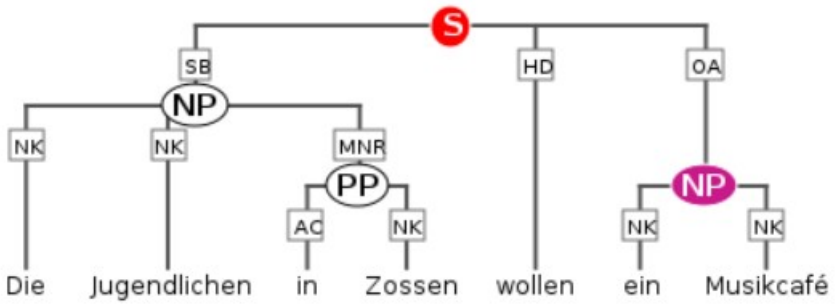
Hierarchy relations

- Relations between attributes

  searching all NPs which are objects

```
node & cat="NP" & #1 >[func="OA"] #2
```

  e.g. *ein Musikcafé* -->object!

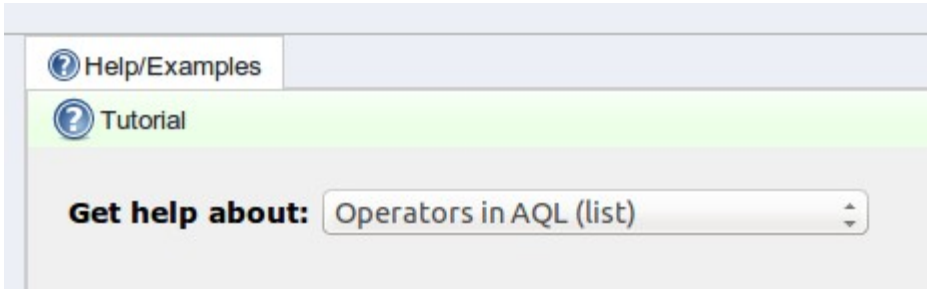


Used Relations

- Relations we used:

A _i_ B	A includes B
A > B	A dominates B
A >[func="OA"] B	A dominates B and B is an object

The full list of relations can be found in ANNIS



What's new in ANNIS

version 3.1.7

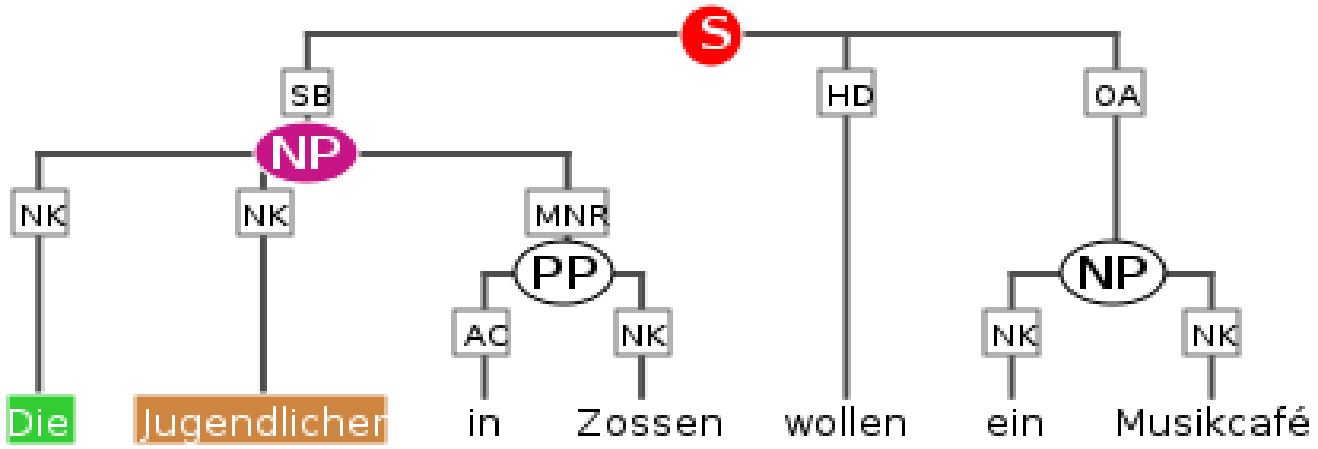
What's new in ANNIS

- Simplified syntax (AQL)
- Frequency analysis (Visualisierung)
- Expand match context (Visualisierung)
- Equality and Inequality (AQL)
- Variables (AQL)
- Complex OR expression (AQL)
- Document browser (Visualisierung)
- CSV export (Visualisierung)
- Tooltip for corpus names (Visualisierung)
- Report problem (Visualisierung)

Simplified syntax

- Question:

  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence

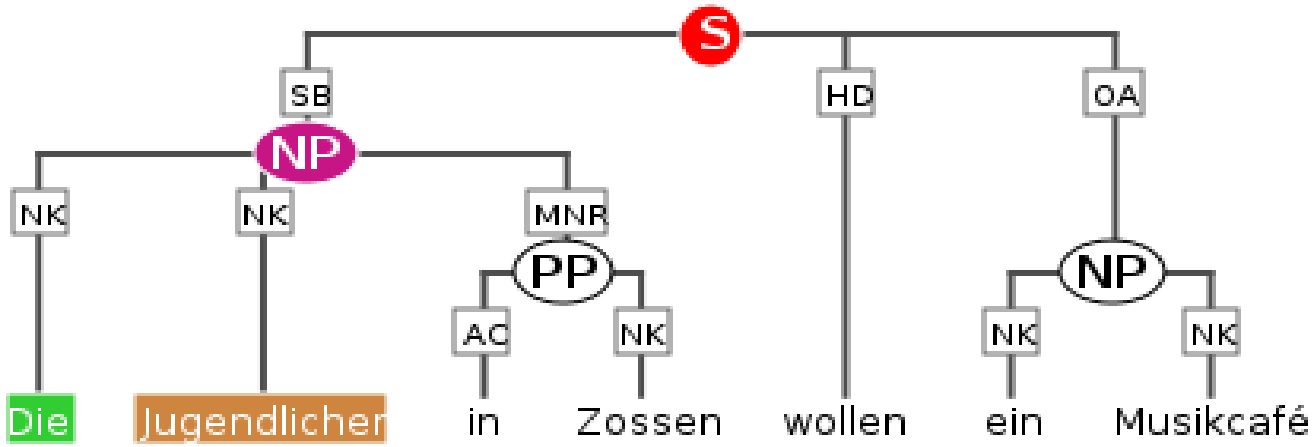


So far:
cat="S" & cat="NP" & "Die" & "Jugendlichen" & #1 > #2 & #2 > #3 & #2 > #4 & #3 . #4

Simplified syntax

- Question:

  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence



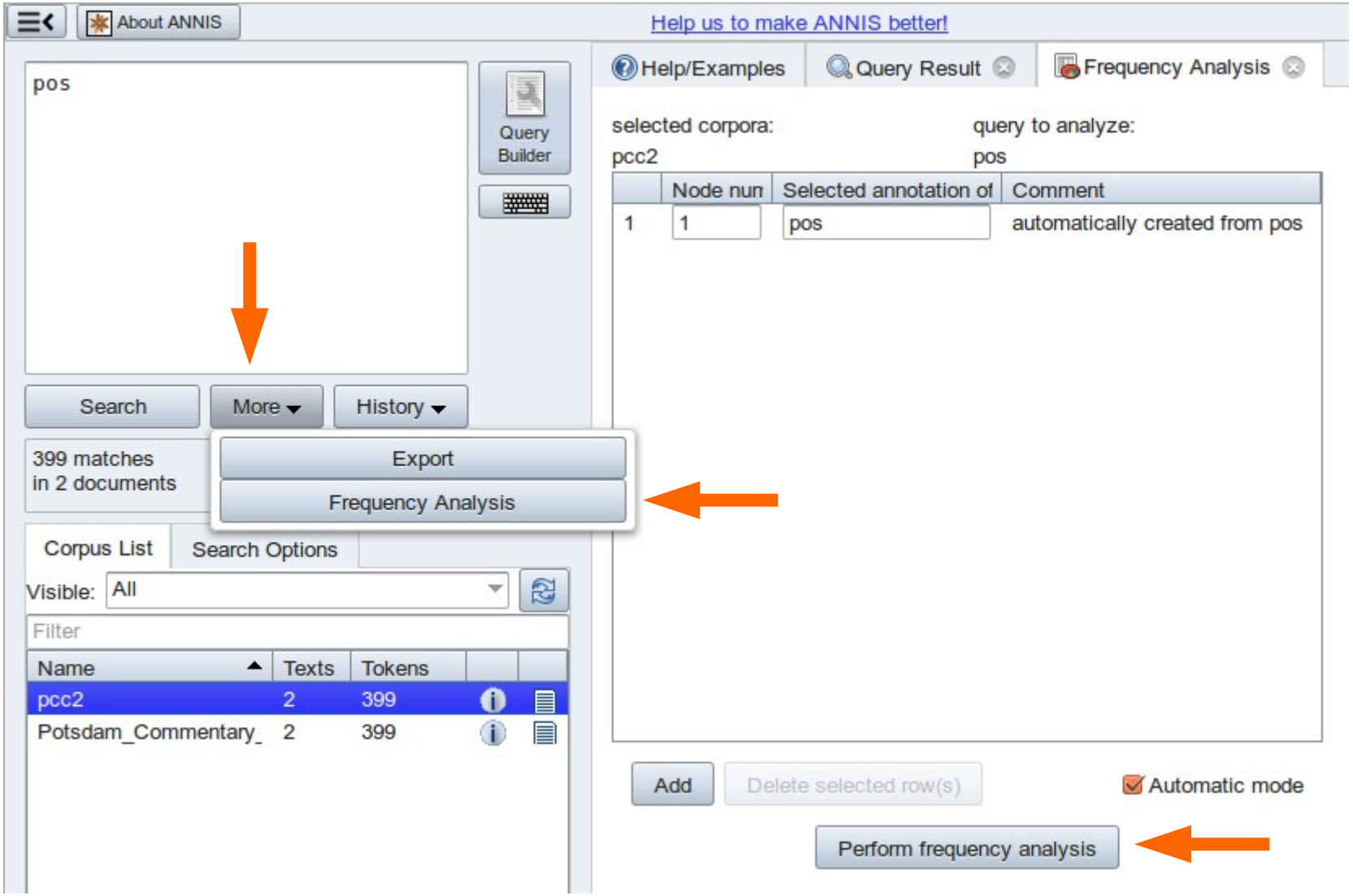
So far:
cat="S" & cat="NP" & "Die" & "Jugendlichen" & #1 > #2 & #2 > #3 & #2 > #4 & #3 . #4

Simplified:
cat="S" > cat="NP" > "Die" . "Jugendlichen" & #2 > #4

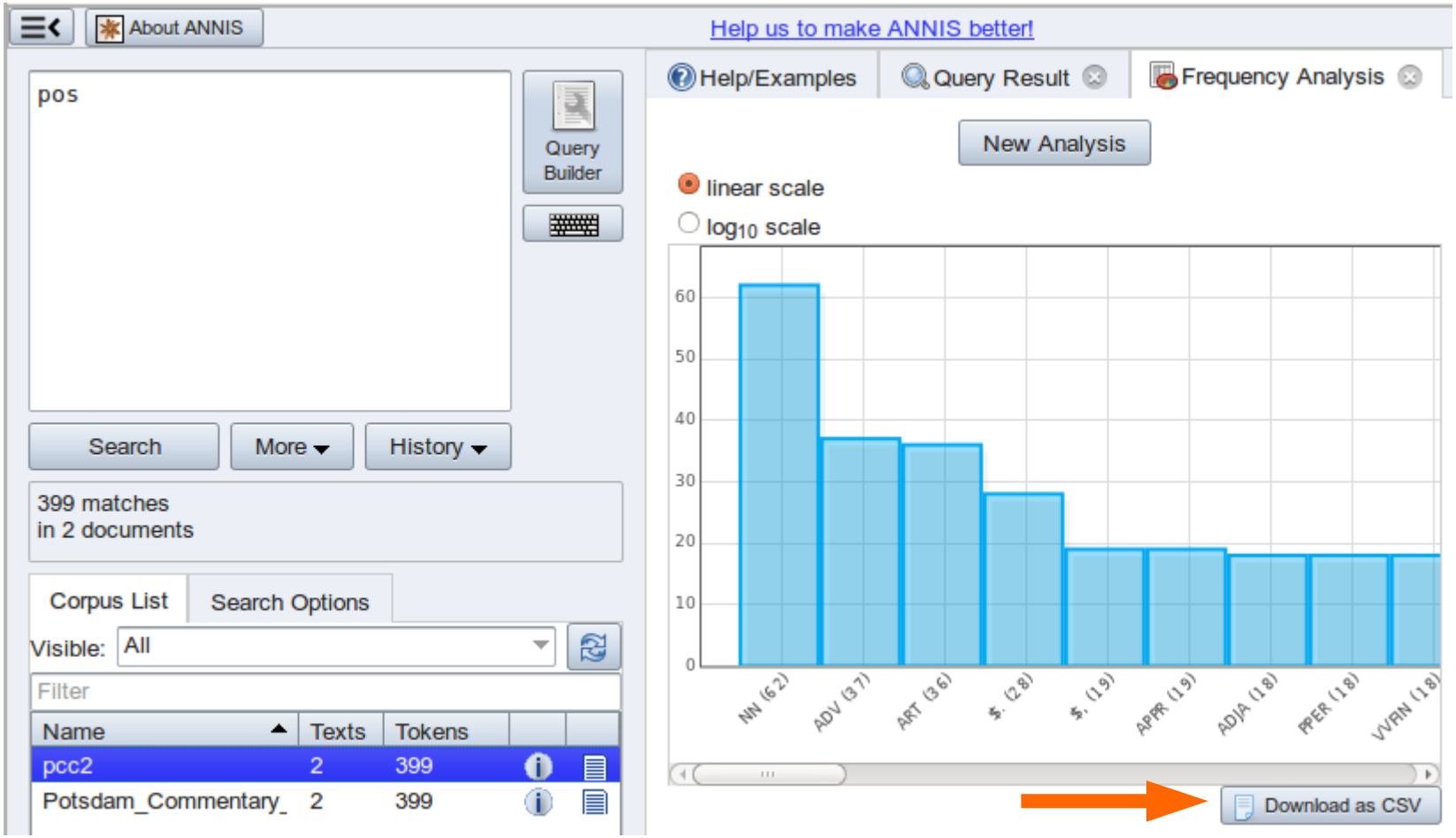
Frequency analysis

- Question:
  - How many words tagged as „NN“, „ADJA“ or „ADV“ does a corpus contain?
  - What are the most frequent part-of-speech tags followed by a noun?
  - What are the most frequent part-of-speech tags in a prepositional phrase, which is in a sentence?
  - ...

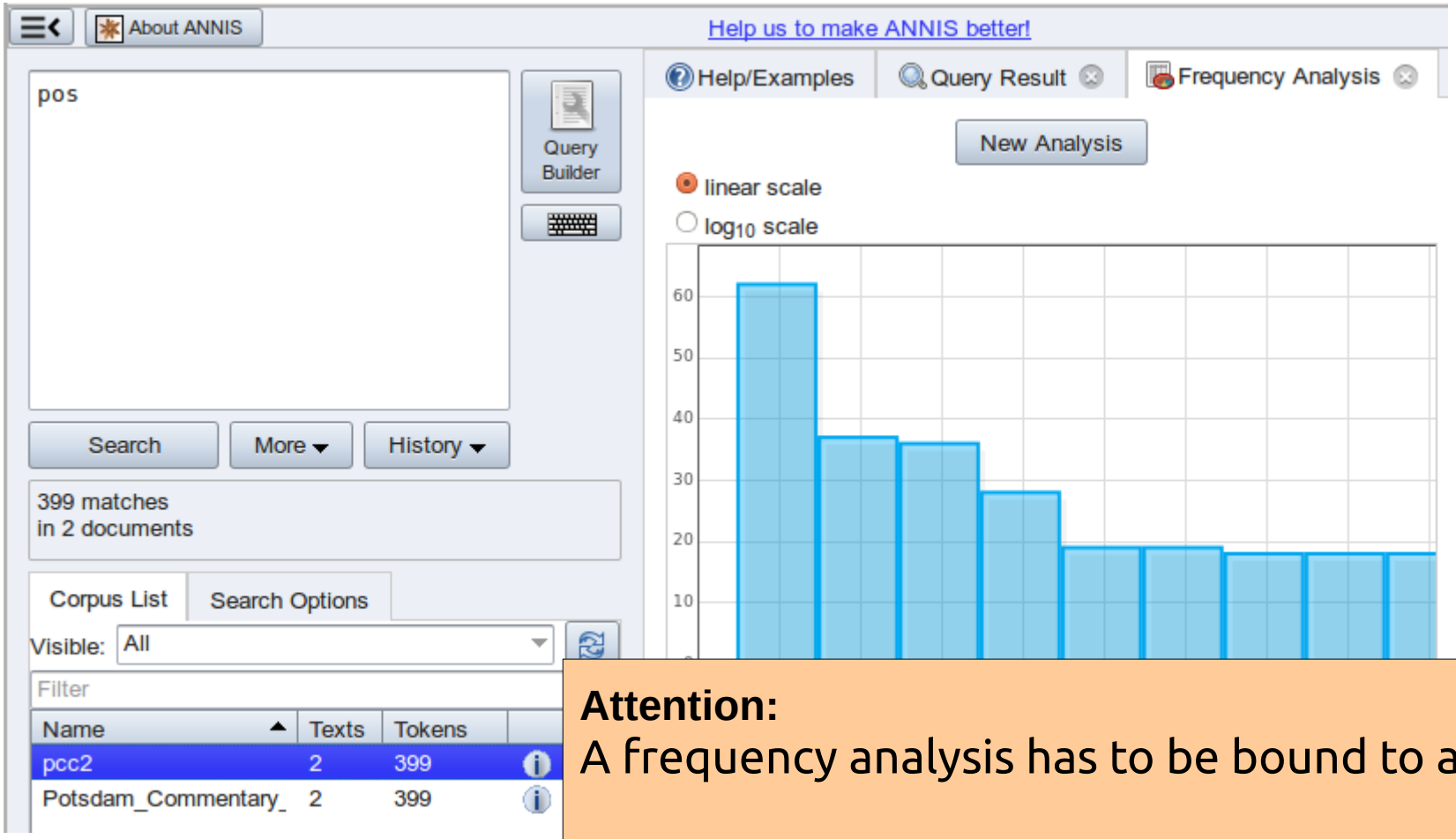
Frequency analysis



Frequency analysis



Frequency analysis

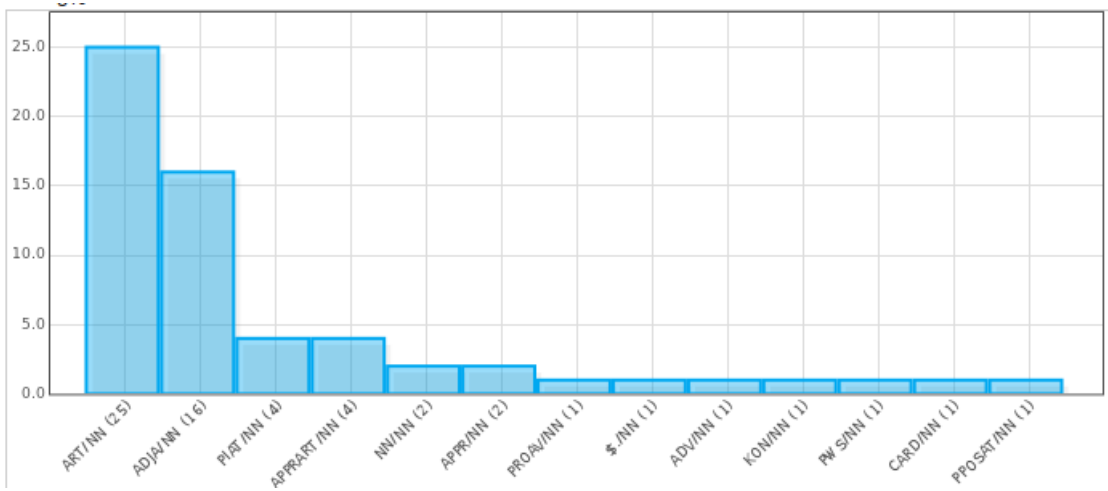


Attention:
A frequency analysis has to be bound to a query!

Frequency analysis

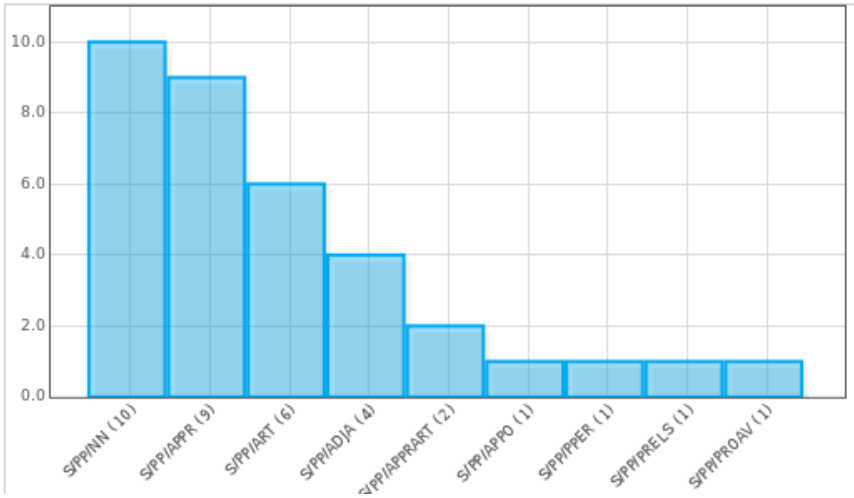
- What are the most frequent part-of-speech tags followed by a noun?

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



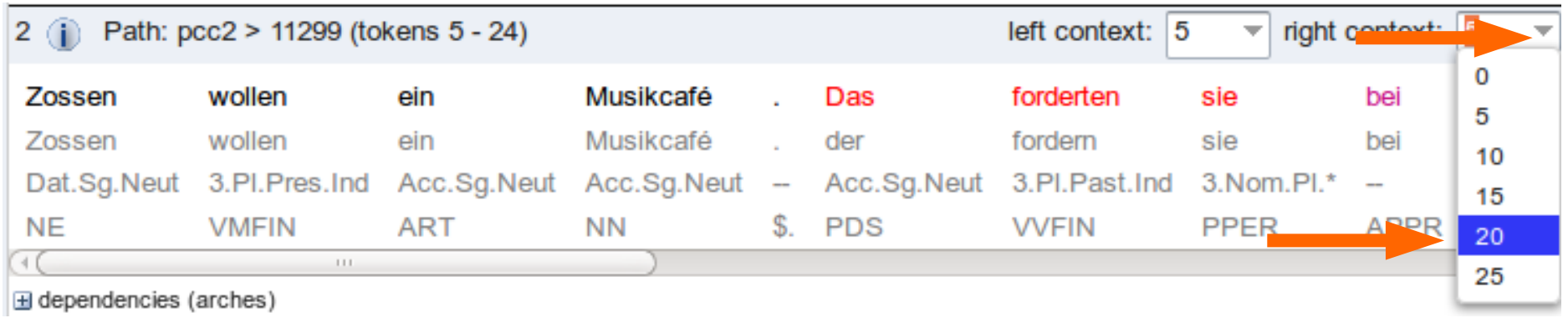
- What are the most frequent part-of-speech tags in a prepositional phrase, which is in a sentence?

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



Expand match context

- Sometimes the context is too small

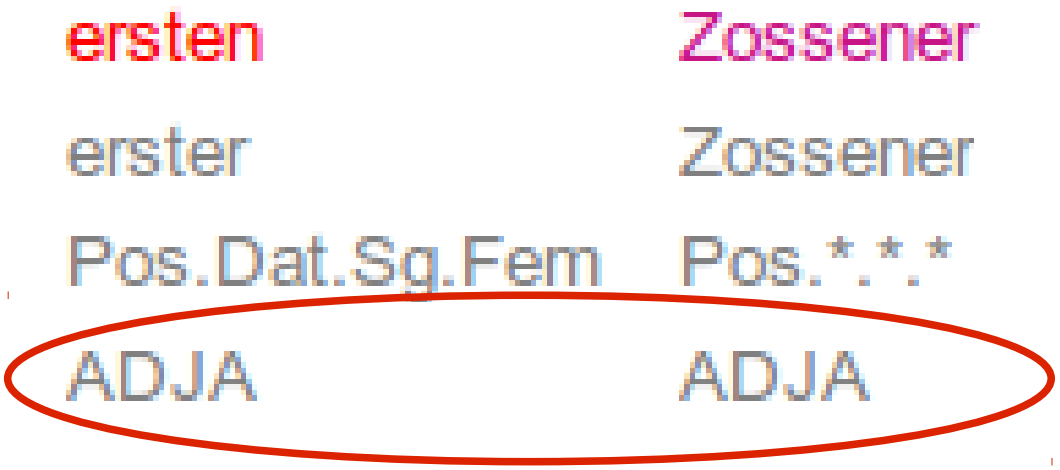


- Even more than 25 is possible, it's a free text field

Equality and Inequality

- Equality „==“ and inequality „!=“ for attributes
- Question (equality):

  two same part-of-speech tags, one directly following the other

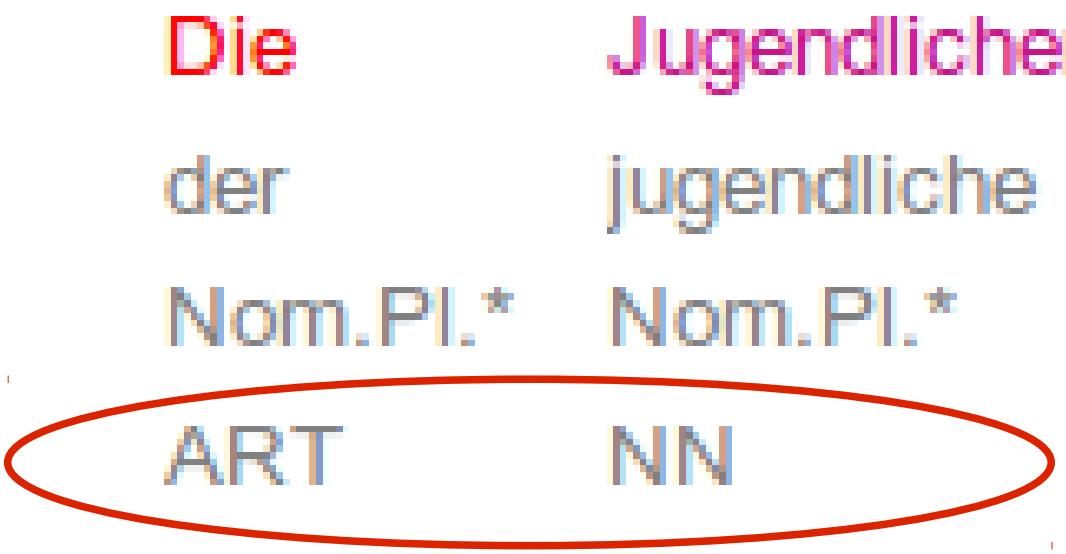



```
pos . pos & #1 == #2
```

Equality and Inequality

- Equality „==“ and inequality „!=“ for attributes
- Question (inequality):

  two different part-of-speech tags, one directly following the other

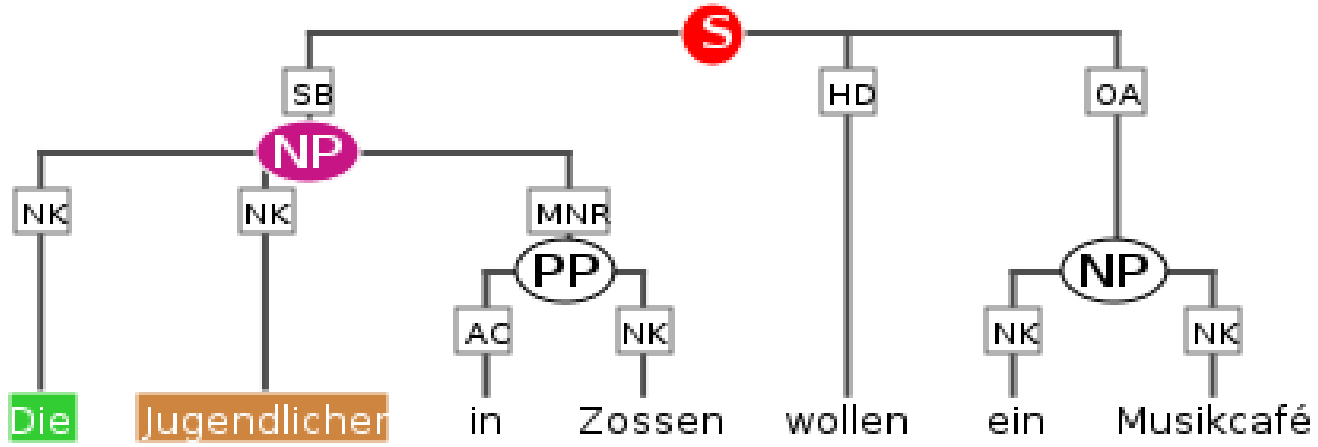



```
pos . pos & #1 != #2
```

Variables

- Question:

  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence

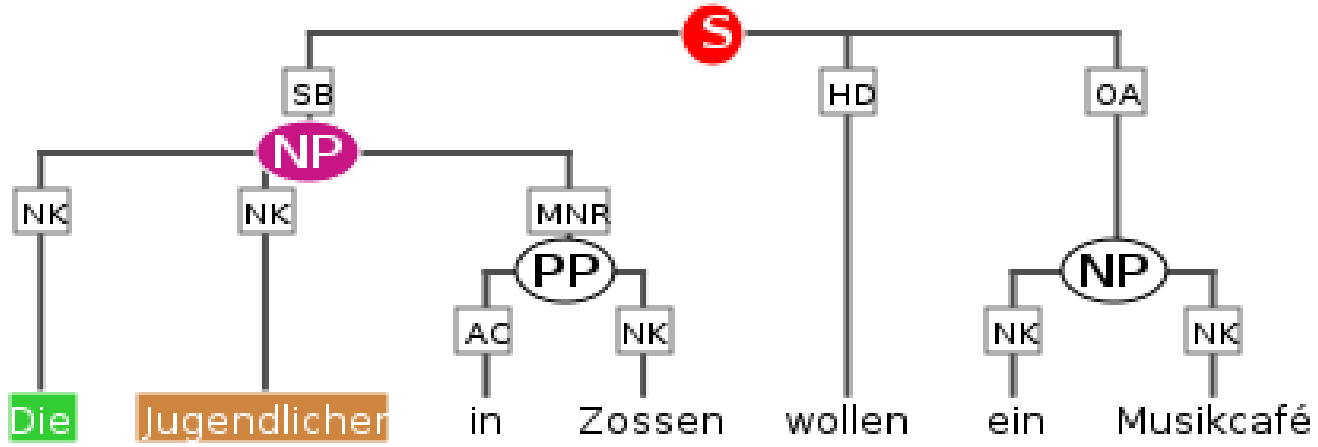


Simplified:
cat="S" > cat="NP" > "Die" . "Jugendlichen" & #2 > #4

Variables

- Question:

  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence

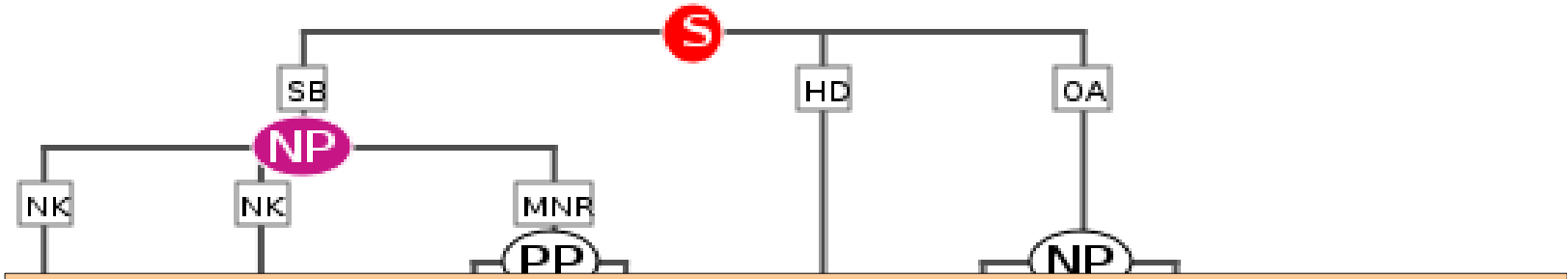


Simplified:
cat="S" > np#cat="NP" > "Die" . jug#"Jugendlichen" & #np > #jug

Variables

- Question:

  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence



Variables and numbers can be mixed:

cat="S" > np#cat="NP" > "Die" . "Jugendlichen" & #np > #4

Simplified:

cat="S" > np#cat="NP" > "Die" . jug#"Jugendlichen" & #np > #jug

Complex OR expression

- Question (simple OR):

  A part-of-speech tag which is a noun, an attributive adjective or an article

```
pos=/(NN)|(ADJA)|(ART)/          (in pattern search)
```

Complex OR expression

- Question (simple OR):

  A part-of-speech tag which is a noun, an attributive adjective or an article

```
pos=/(NN)|(ADJA)|(ART)/          (in pattern search)
```

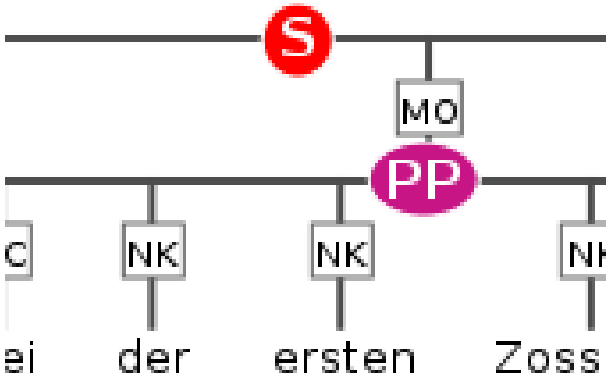
- OR for expressions

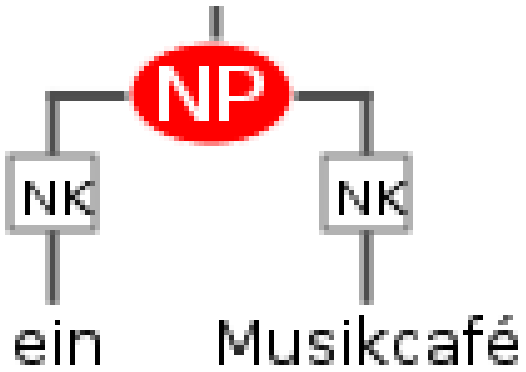
```
pos="NN" | pos="ADJA" | pos= "ART"
```

Complex OR expression

- Question (complex OR):

  A prepositional phrase, which is dominated by a sentence, or just a nominal phrase



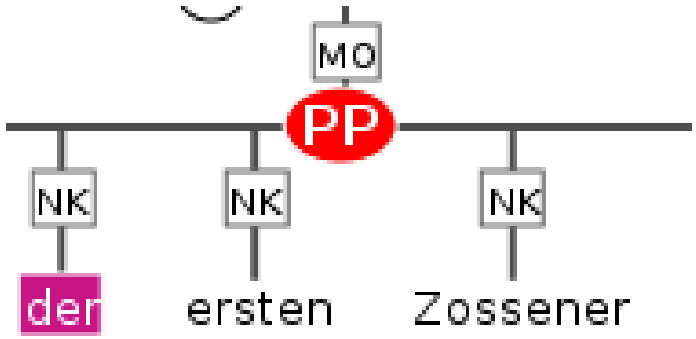



```
(cat="S" > cat="PP") | cat="NP"
```

Complex OR expression

- Question (nested OR):

  A prepositional phrase, which dominates a noun, an attributive adjective or an article

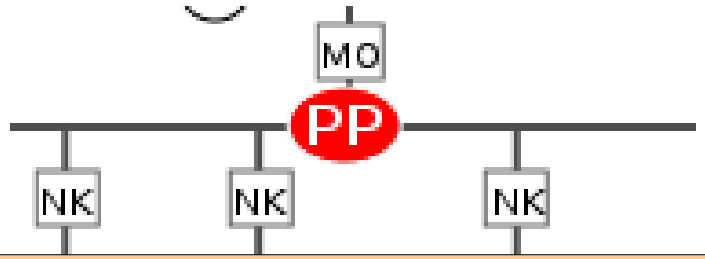



```
a#cat="PP" &
(b#pos="NN" | b#pos="ADJA" | b#pos= "ART") &
#a > #b
```

Complex OR expression

- Question (nested OR):

  A prepositional phrase, which dominates a noun, an attributive adjective or an article

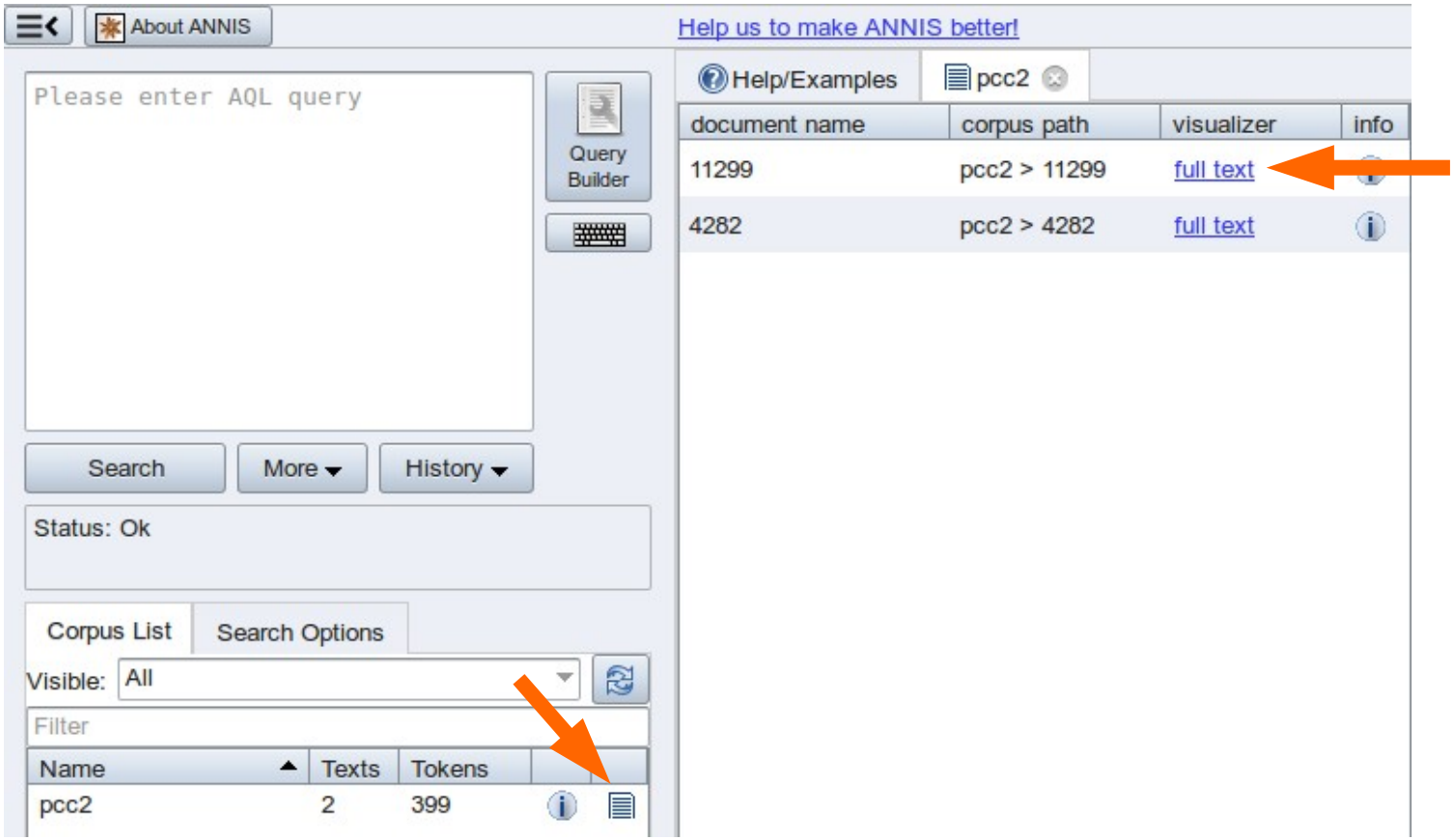


Attention:
All expressions in brackets have to use the same variable
… & (b#pos="NN" | b#pos="ADJA" | b#pos= "ART") & …

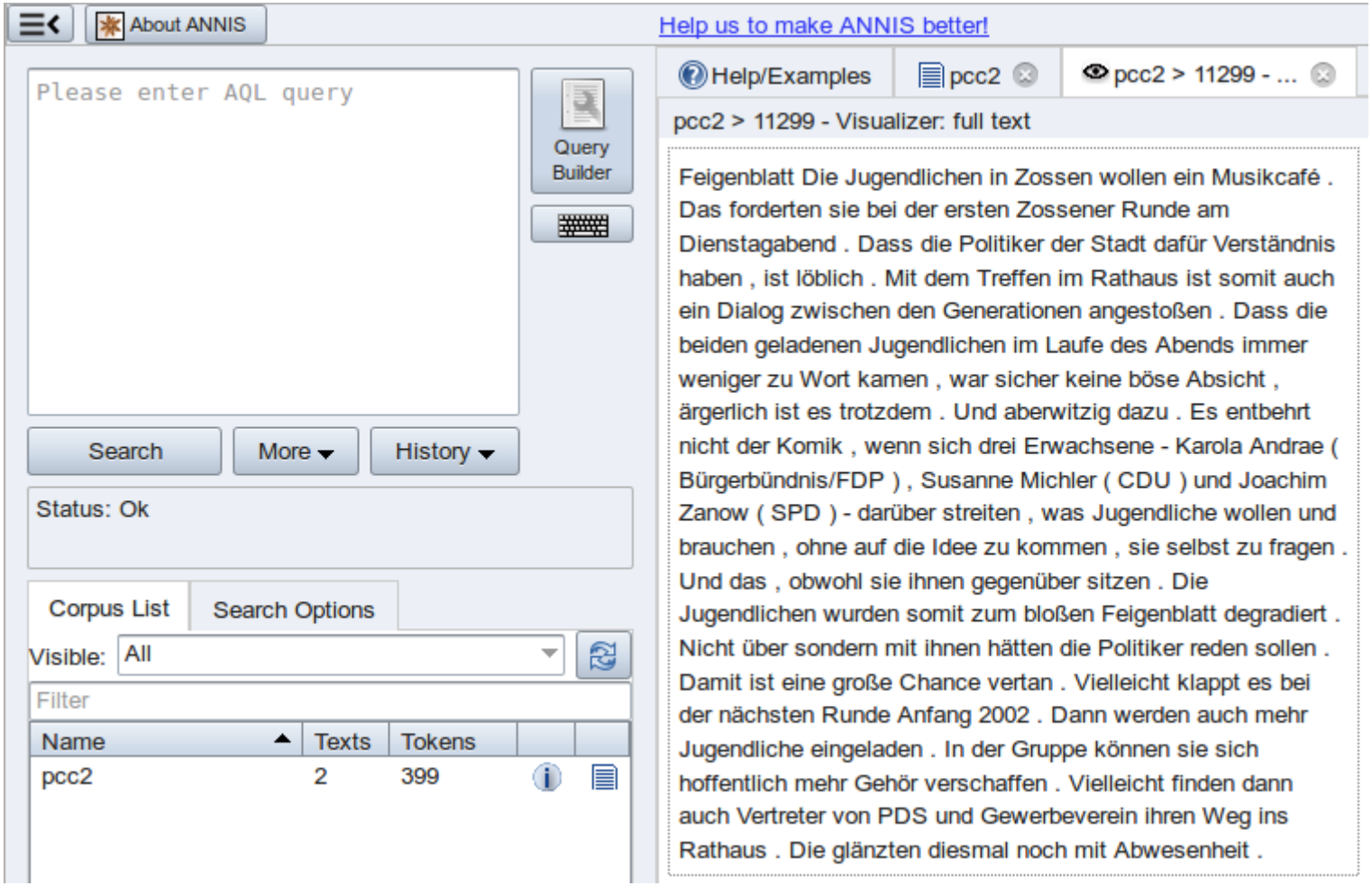
```
a#cat="P
(b#pos="NN" | b#pos="ADJA" | b#pos= "ART") &
#a > #b
```

Document browser

- Displays the entire text of a document

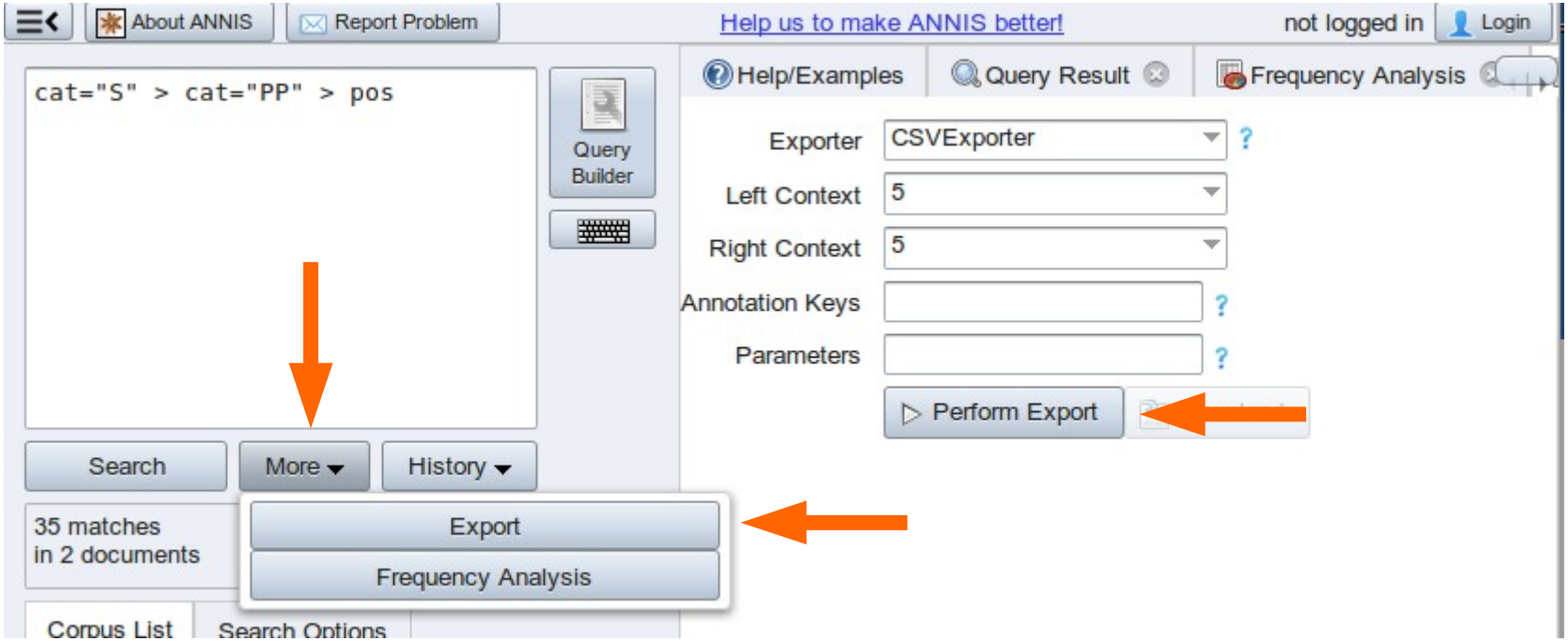


Document browser



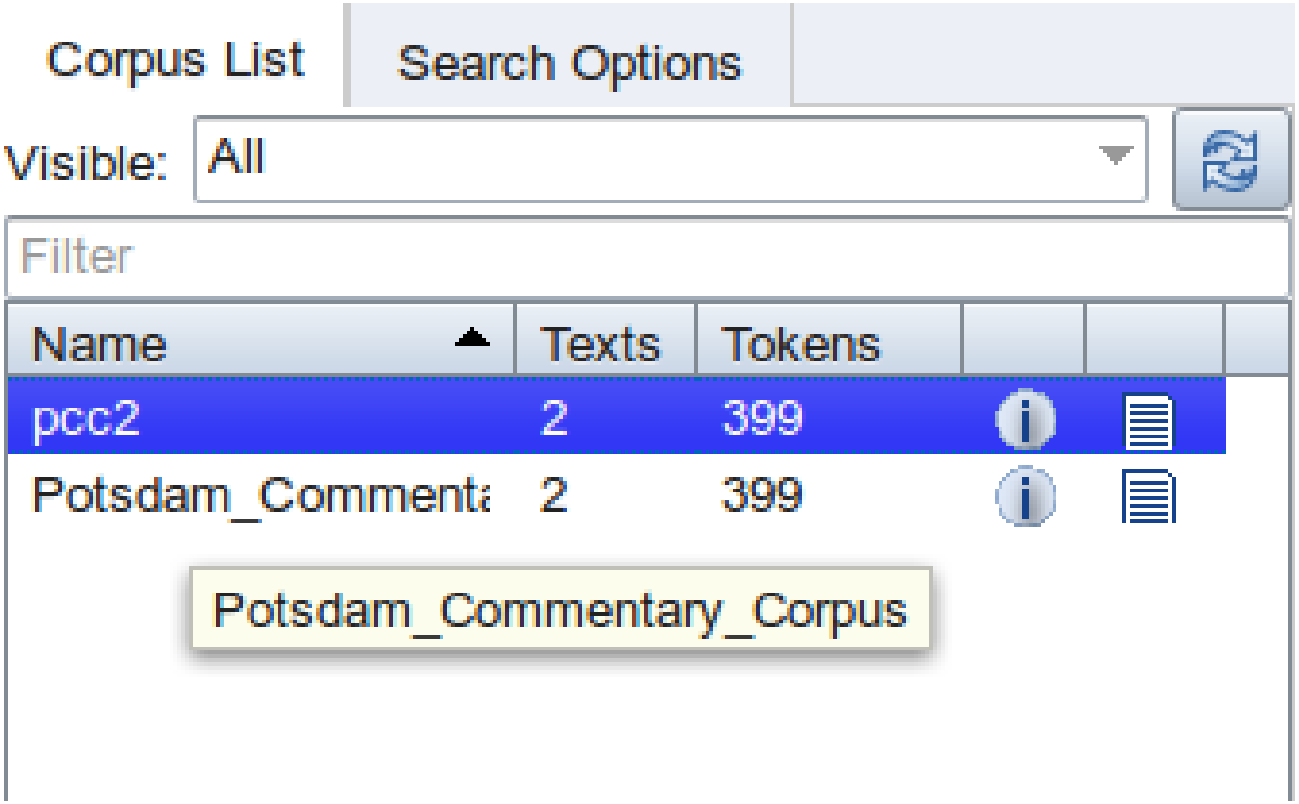
CSV export

- Export data for futher processing

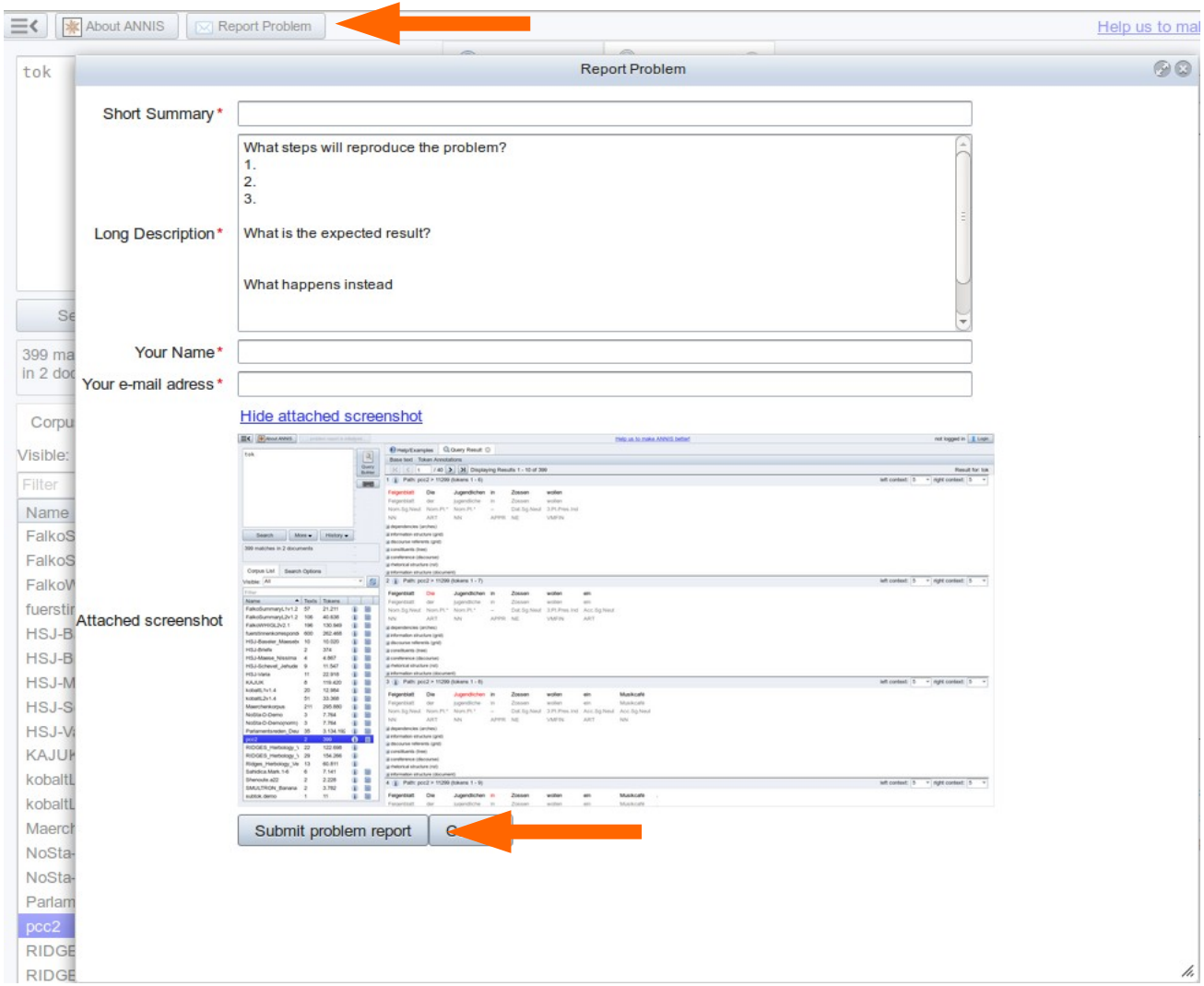


Tooltips for corpus names

- Sometimes corpus names can get very long



Report problem



Get ANNIS

- ANNIS comes in two flavors
  - A server version
  - A desktop version (ANNIS kickstarter)
  - Both are downloadable at: http://www.sfb632.uni-potsdam.de/annis/
- ANNIS is open source (Apache license 2.0) and hosted on github
  - https://github.com/korpling/ANNIS

Thanks for your attention!

Any questions?

carolin.odebrecht@hu-berlin.de,
f.zipser@gmx.de