ANNIS: Search and Visualization in Multilayer Linguistic Corpora

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!
• 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
• Search and Visualization in Multilayer Linguistic Corpora
  – Displays corpora
    • Many visualizations available
    • Corresponding to type of annotation (syntactic trees, phrase trees (RST), grids, coreferences ...)
• What ANNIS cannot do
  - Does not know how to speak natural language
    → so you have to learn AQL
• 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
• 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

- **Enter query**
- **Virtual Keyboard (e.g. arabic)**
- **Previous queries**
- **Corpus list**
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Sample queries (corresponding to corpus)
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Query result

Visualizations
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Document metadata

Document metadata window
• Basic principles of AQL (ANNIS Query Language)
  – Attributes and values
    • Searching for exact character sequences
    • Searching for patterns
  – Combinatory search
• 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
• Different types of annotations
  - Token annotation
  - Span annotation
  - Pointing relation
  - Hierarchy annotation (trees)
• 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"
• Token annotation
  - Exact sequence

  searching for a word form
  "Jugendlichen" 3 hits
  "jugendlichen" 0 hits

  → tok="jugendlichen"
• Token annotation
  – Exact sequence

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

  \[
  \text{pos} = "\text{NN}" \\
  \text{attribute} \quad \text{value}
  \]

  – Attributes can have more than one value
  – Searching for all values of an attribute
• Token annotation
  – Exact sequence

searching for an exact part of speech tag

```plaintext
pos="NN"
pos="ADJA"
```
• **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
    ```
• Span annotation
  - Exact sequence

searching for sentences

Sent="s"
• 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
• Token annotation
  – Patterns
    . matches any single character
    * zero or more of the preceding element

searching for the beginning a of word

/Jugend.*/

/jugend.*/
• Token annotation
  – Patterns
    1. matches any single character
    2. 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)
• Token annotation
  - patterns

searching for **all** nouns

\[
\text{pos}=/N./ \quad \text{includes NN & NE}
\]

searching for **all** adjectives

\[
\text{pos}=/\text{ADJ}./ \quad \text{includes ADJA & ADJD}
\]
• 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)
• Span annotation

searching for all NPs

\[ \text{cat}="\text{NP}" \]

41 hits (pos="NN" 62 hits)

e.g. *Die Jugendlichen in Zossen*
• Relations between attributes

searching for all NPs which contain a preposition

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat=&quot;NP&quot;</td>
<td>41 hits</td>
</tr>
<tr>
<td>pos=&quot;APPR&quot;</td>
<td>19 hits</td>
</tr>
</tbody>
</table>

e.g. *Die Jugendlichen in Zossen*

→ no relation between the two information!
• 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 attributes

searching for all NPs which contain a preposition

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

e.g. *Die Jugendlichen in Zossen*
• 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

![Diagram showing the hierarchy relations between tokens, node annotation, edge annotation, and span.](image)
• Relations between attributes

searching all NPs which are objects

\texttt{cat="NP"}

the syntactic function in the tree

\texttt{func="OA"}

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

searching all NPs which are objects

\texttt{node \& cat="NP" \& #1 \> [func="OA"] #2}

e.g. \textit{ein Musikcafé} --> object!

\begin{itemize}
  \item \texttt{Die Jugendlichen in Zossen wollen ein Musikcafé}
\end{itemize}
Used Relations

- Relations we used:
  
  \[ A \_i\_ B \quad \text{A includes B} \]
  
  \[ A > B \quad \text{A dominates B} \]
  
  \[ A > \left[ \text{func="OA"} \right] B \quad \text{A dominates B and B is an object} \]

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)
• Question:

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

So far:

\[
\text{cat} = "S" \& \text{cat} = "NP" \& "Die" \& "Jugendlichen" \& #1 > #2 \& #2 > #3 \& #2 > #4 \& #3 . #4
\]
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
• 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?
  - ...
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Frequency analysis
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Frequency analysis

![Graph showing frequency analysis results with download option](image-url)
Frequency analysis

Attention: A frequency analysis has to be bound to a query!
• What are the most frequent part-of-speech tags followed by a noun?

\[
\text{pos . pos="NN"}
\]

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

\[
\text{cat="S" > cat="PP" > pos}
\]
• 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

```
ersten  Zossener
erster  Zossener
Pos.Dat.Sg.Fem Pos.*.*.*
ADJA       ADJA
```

pos . pos & #1 == #2
• Equality „==“ and inequality „!=“ for attributes
• Question (inequality): two different part-of-speech tags, one directly following the other

```plaintext
pos . pos & #1 != #2
```
• Question:

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

Simplified:
\[ \text{cat}=\text{"S"} > \text{cat}=\text{"NP"} > \text{"Die" . "Jugendlichen" & #2 > #4} \]
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
• Question:

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

Simplified:

\[
\text{cat}="S" \rightarrow \text{np}\#\text{cat}="\text{NP}" \rightarrow "\text{Die}" \cdot "\text{Jugendlichen}" \& \#\text{np} > \#4
\]

Variables and numbers can be mixed:

\[
\text{cat}="S" \rightarrow \text{np}\#\text{cat}="\text{NP}" \rightarrow "\text{Die}" \cdot \text{jug}"\text{Jugendlichen}" \& \#\text{np} > \#4
\]
• Question (simple OR):

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

\[ \text{pos}=/\text{NN}|\text{ADJA}|\text{ART}/ \]  (in pattern search)
• Question (simple OR):
  
  A part-of-speech tag which is a noun, an attributive adjective or an article

  \[
  \text{pos}=/\text{NN}|\text{ADJA}|\text{ART}/ \quad \text{(in pattern search)}
  \]

• OR for expressions

  \[
  \text{pos}="\text{NN}\" \mid \text{pos}="\text{ADJA}\" \mid \text{pos}="\text{ART}\"
  \]
• Question (complex OR):
  A prepositional phrase, which is dominated by a sentence, or just a nominal phrase

\[(\text{cat}="S" \rightarrow \text{cat}="PP") \mid \text{cat}="NP"\]
• Question (nested OR):

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

\[
a\#cat="PP" \&
(b\#pos="NN" \| b\#pos="ADJA" \| b\#pos="ART") \&
#a > #b
\]
• Question (nested OR):

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

\[
a\texttt{cat} = \texttt{PP} & (b\texttt{pos} = \texttt{NN} | b\texttt{pos} = \texttt{ADJA} | b\texttt{pos} = \texttt{ART}) & \ldots
\]

Attention:
All expressions in brackets have to use the same variable
... & (b#pos="NN" | b#pos="ADJA" | b#pos="ART") & ...
• Displays the entire text of a document
ANNIS workshop

Carolin Odebrecht & Florian Zipser

ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Document browser
• Export data for further processing

CSV export
• Sometimes corpus names can get very long
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Report problem
• 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