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
• 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
• 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
Sample queries (corresponding to corpus)

<table>
<thead>
<tr>
<th>Example Query</th>
<th>Description</th>
<th>open corpus browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma=&quot;sein&quot; &amp; meta::Genre=&quot;Sport&quot;</td>
<td>search for all words with the lemma &quot;sein&quot; in documents from the Genre &quot;Sport&quot;</td>
<td>pcc2</td>
</tr>
<tr>
<td>exmaralda:Inf-Stat=&quot;new&quot;</td>
<td>search for all Inf-Stat annotations with the value &quot;new&quot; in the &quot;exmaralda&quot; namespace</td>
<td>pcc2</td>
</tr>
<tr>
<td>&quot;statisch&quot;</td>
<td>search for the word &quot;statisch&quot;</td>
<td>pcc2</td>
</tr>
<tr>
<td>lemma!=&quot;sein&quot;</td>
<td>search for all words where the lemma is not &quot;sein&quot;</td>
<td>pcc2</td>
</tr>
<tr>
<td>tok!=&quot;ist&quot;</td>
<td>search for all words that are not &quot;list&quot;</td>
<td>pcc2</td>
</tr>
<tr>
<td>/.*lich/</td>
<td>Search for words ending with &quot;lich&quot; (regular expression)</td>
<td>pcc2</td>
</tr>
</tbody>
</table>
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Query result

Visualizations
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Corpus metadata window

Corpus metadata
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
• 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)
• Token annotation
  - Exact sequence

searching for a word form

"Jugendlichen"
"jugendlichen"
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

**Exact word forms**

- **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

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

  attribute \hspace{1cm} \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

pos="NN"
pos="ADJA"
• Token annotation
  - Exact sequence

  searching for an exact part of speech tag

  \[
  \begin{align*}
  \text{pos} &= \text{"NN"} & \text{62 hits} \\
  \text{pos} &= \text{"ADJA"} & \text{18 hits} \\
  \text{pos} & & \text{399 hits}
  \end{align*}
  \]

  searching for all values of an attribute
• 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
    . 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)
• Token annotation
  - patterns

searching for **all** nouns

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

searching for **all** adjectives

\[ \text{pos}=/\text{ADJ./} \]
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

\texttt{cat="NP"} 41 hits (pos="NN" 62 hits)

e.g. \textit{Die Jugendlichen in Zossen}
• Relations between attributes

searching for all NPs which contain a preposition

\begin{itemize}
  \item \texttt{cat="NP"} \hspace{2cm} 41 hits
  \item \texttt{pos="APPR"} \hspace{2cm} 19 hits
\end{itemize}

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
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!
• Relations between attributes

searching all NPs which are objects

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

e.g. *ein Musikcafé* --> object!

![Hierarchy diagram]

Die, Jugendlichen, in, Zossen, wollen, ein, Musikcafé
• 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)
• Question:

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

So far:

\[ \text{cat="S" & cat="NP" & "Die" & "Jugendlichen" & #1 > #2 & #2 > #3 & #2 > #4 & #3 . #4} \]
• Question:

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

So far:
\[
\text{cat} = "S" \& \text{cat} = "NP" \& "Die" \& "Jugendlichen" \& #1 > #2 \& #2 > #3 \& #2 > #4 \& #3 . #4
\]

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

[Image of the ANNIS interface with a focus on the Frequency Analysis tab and a comment indicating that the selected annotation of 'pos' is automatically created from 'pos'.]

Carolin Odebrecht & Florian Zipser

ANNIS workshop
ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Frequency analysis

[Image of a frequency analysis graph with a bar chart showing frequency distribution across different categories.]
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 "!=" for attributes

Question (inequality):

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

Die

der

Jugendlichen

ejugendliche

Nom.Pl.*

Nom.Pl.*

ART

NN

pos . pos & #1 != #2
Equality „==“ and inequality „!=“ for attributes

Question (equality):

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

\[
\text{pos . pos & #1 == #2}
\]
Equality “==” and inequality “!=“ for attributes

Question (inequality):

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

```
Die der Jugendlichen der Jugendliche
ART NN
```

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

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

Simplified:

\[\text{cat} = "S" \rightarrow \text{np#cat} = "NP" \rightarrow "\text{Die}" \rightarrow \text{jug#"Jugendlichen"} \& \#np \rightarrow \#jug\]
• Question:

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

Variables and numbers can be mixed:
\[
\text{cat} = "S" \rightarrow \text{np\#cat} = "NP" \rightarrow "Die" \cdot "Jugendlichen" \& \#\text{np} > \#4
\]

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

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

```plaintext
pos=/^(NN)|(ADJA)|(ART)/
```

(in pattern search)
• 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"
```
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

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

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

- Displays the entire text of a document
• Export data for further processing
• 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