ANNIS User Guide

Amir Zeldes <annis-admin@ling.uni-potsdam.de>

Version 2.2.1

18 April 2012

Table of Contents

1. Introduction

ANNIS is an open source, browser-based search and visualization architecture for multi-layer corpora. It can be used to search for complex graph structures of annotated nodes and edges forming a variety of linguistic structures, such as constituent or dependency syntax trees, coreference and parallel alignment edges, span annotations and associated multi-modal data (audio/video). This guide provides an overview of the current ANNIS system, first steps for installing either a local instance or an ANNIS server with a demo corpus, as well as tutorials for converting data for ANNIS and running queries with AQL (ANNIS Query Language).

2. New Features in Version 2.2.1

Features:

- Subcorpus and document path for each hit can be shown in search result list
- Metadata coming from documents, subcorpora and top level corpora is sorted separately in the metadata display
- New grid-tree visualizer for a span visualization of hierarchical structures (e.g. topological fields in German; see Section 5.1)
- More intuitive search button position in GUI and other small GUI improvements
- Updated tutorial
- Faster import when large corpora are already in the system

Bugfixes:

- Support for JDK 7
- Fixed import bug which could disrupt queries of the sort >secedge m,n (see https://bugs.launchpad.net/bugs/870108)
- Fixed bugs in some forms of query disjunction with ‘|’
- Fixed resizing of visualizers in Chrome browsers
- Fixed escaping of characters in WEKA export

(For change logs of previous version see their respective distributions or user guides)

3. Installing ANNIS

3.1. Installing a Local Version (ANNIS Kickstarter)

Local users who do not wish to make their corpora available online can install ANNIS Kickstarter under most versions of Linux, Windows and Mac OS. To install Kickstarter follow these steps:

1. Download and install PostgreSQL 8.4 for your operating system from http://www.postgresql.org/download/ and **make a note of the administrator password** you set during the installation. After installation, Postgres may automatically launch the Postgres Stack Builder to download additional components – you can safely skip this step and cancel the Stack Builder if you wish. You may need to restart your OS if the Postgres installer tells you to.

2. Download and unzip Annis-Kickstarter-2.2.1-distribution.zip [http://launchpad.net/annis/2.2/2.2.1/+download/Annis-Kickstarter-2.2.1-distribution.zip] from the ANNIS website.

3. Start AnnisKickstarter.bat if you’re using Windows or run the bash script AnnisKickstarter.sh otherwise (this may take a few seconds the first time you run Kickstarter). At this point your Firewall may try to block Kickstarter and offer you to unblock it – do so and Kickstarter should start up.

   ### Note

   For most users it is a good idea to give Java more memory (if this is not already the default). You can do this by editing the script AnnisKickstarter and typing the following after the call to start java (before -splash:splashscreen.gif):

   ```
   -Xss1024k -Xmx1024m
   ```

   (To accelerate searches it is also possible to give the Postgres database more memory, see the link in the next section below).

4. Once the program has started, if this is the first time you run Kickstarter, press “Init Database” and supply your PostGres administrator password from step 1.

5. Download and unzip the pcc2 demo corpus [http://korpling.german.hu-berlin.de/~annis/downloads/sample_corpora/pcc2_relAnnis.zip] from the ANNIS website.

6. Press “Import Corpus” and navigate to the directory containing the directory `pcc2_v2_relAnnis/`. Select this directory (but do not go into it) and press OK.

7. Once import is complete, press “Launch Annis frontend” and login with the username and password “test” to test the corpus (try selecting the pcc2 corpus, typing `pos="NN"` in the AnnisQL box and clicking “Show Result”. See the section “Running Queries in ANNIS” in this guide for some more example queries, or press the Tutorial button at the top left of the interface).

3.2. Building and Installing an ANNIS Server

The ANNIS server version can be installed on UNIX based server, or else under Windows using Cygwin [http://www.cygwin.com/], the freely available UNIX emulator. To install the ANNIS server:

1. Install a PostgreSQL server for your operating system from http://www.postgresql.org/download/

2. Install a web server such as Tomcat [http://tomcat.apache.org/] or Jetty [http://www.mortbay.org/jetty/]

3. Make sure you have JDK 6 [http://java.sun.com/javase/downloads/index.jsp] and Maven 2 [http://maven.apache.org/] (or install them if you don’t)

4. If you’re using Cygwin and Windows you will also need to install the “patch” program via the Cygwin package manager

5. Download and unzip Annis-2.2.1.zip [http://launchpad.net/annis/2.2/2.2.1/+download/Annis-2.2.1.zip], then run the following commands (replacing the appropriate directories):

   ```
   cd <unzipped source>/Annis-Service
   mvn -DskipTests=true install
   mvn -DskipTests=true assembly:assembly tar xzvf
   target/annis-service-<version>-distribution.tar.gz -C <installation directory
   ```

6. Next initialize your ANNIS database (only the first time you use the system):

7. Set the environment variables (each time when starting up)

   ```
   export ANNIS_HOME=<installation directory>
   export PATH=$PATH:$ANNIS_HOME/bin
   ```

8. Now you can import some corpora:

   ```
   annis-admin.sh import path/to/corpus1 path/to/corpus2 ...
   ```

   ### Important

   The above import-command calls other PostgreSQL database commands. If you abort the import script with Ctrl+C, these SQL processes will not be automatically terminated; instead they might keep hanging and prevent access to the database. The same might happen if you close your shell before the import script terminates, so you will want to prefix it with the "nohup"-command.

9. Now you can start the ANNIS service:

   ```
   annis-service.sh start
   ```

10. To get the Annis front-end running, first compile it:

    ```
    cd <unzipped source>
    mvn -DskipTests=true install
    ```

    If no error occurs the war-file will be available under 4 `<unzipped source>/Annis-web/target/Annis-web.war`.

11. And configure your web server as described here: https://korpling.german.hu-berlin.de/p/projects/annis/wiki/Tomcat [https://korpling.german.hu-berlin.de/p/projects/annis/wiki/Tomcat]

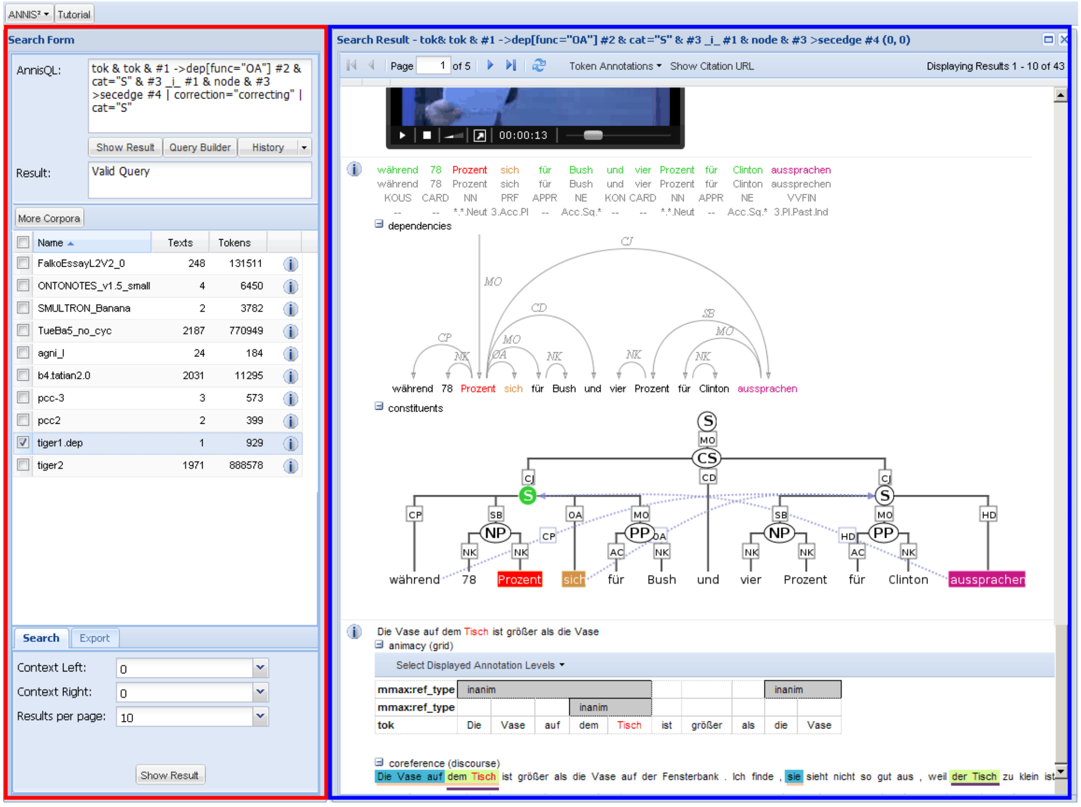
The latest instructions for compiling and installing the ANNIS Server can also be found at: https://korpling.german.hu-berlin.de/p/projects/annis/wiki/Documentation [https://korpling.german.hu-berlin.de/p/projects/annis/wiki/Documentation]

We also **strongly recommend** reconfiguring the Postgres server’s default settings as described here: https://korpling.german.hu-berlin.de/p/projects/annis/wiki/PostgreSQL

4. Running Queries in ANNIS

4.1. The ANNIS Interface

Figure 1. ANNIS interface



AnnisQL: tok & tok & #1 ->dep[func="OA"] #2 & cat="S" & #3 _i_ #1 & node & #3 >secedge #4 | correction="correcting" | cat="S"

Result: Valid Query

☐	Name	Texts	Tokens
☐	FalkoEssayL2V2_0	248	131511
☐	ONTONOTES_v1.5_small	4	6450
☐	SMULTRON_Banana	2	3782
☐	TueBa5_no_cyc	2187	770949
☐	agni_I	24	184
☐	b4.tatian2.0	2031	11295
☐	pcc-3	3	573
☐	pcc2	2	399
☑	tiger1.dep	1	929
☐	tiger2	1971	888578

Context Left: 0

Context Right: 0

Results per page: 10

Search Result - tok& tok & #1 ->dep[func="OA"] #2 & cat="S" & #3 _i_ #1 & node & #3 >secedge #4 (0, 0)

Page 1 of 5

Displaying Results 1 - 10 of 43

während	78	Prozent	sich	für	Bush	und	vier	Prozent	für	Clinton	aussprachen
während	78	Prozent	sich	für	Bush	und	vier	Prozent	für	Clinton	aussprechen
KOUS	CARD	NN	PRF	APPR	NE	KON	CARD	NN	APPR	NE	VVFIN
--	--	*.*.Neut	3.Acc.Pl	--	Acc.Sg.*	--	--	*.*.Neut	--	Acc.Sg.*	3.Pl.Past.Ind

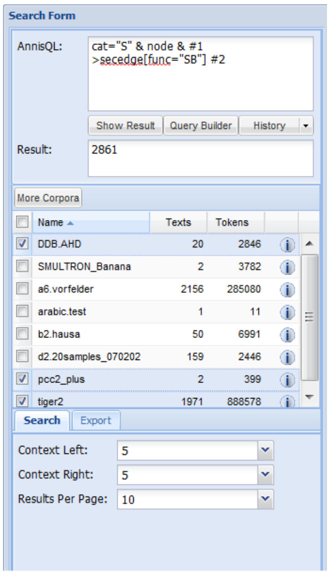
Die Vase auf dem Tisch ist größer als die Vase

mmax:ref_type	inanim								inanim	
mmax:ref_type				inanim						
tok	Die	Vase	auf	dem	Tisch	ist	größer	als	die	Vase

Die Vase auf dem Tisch ist größer als die Vase auf der Fensterbank . Ich finde , sie sieht nicht so gut aus , weil der Tisch zu klein ist

The ANNIS interface (see Figure 1, “ANNIS interface”) is comprised of several windows, the most important of which are the search form (in the red box above left) and the results window (in the blue box above).

4.1.1. The Search Form



AnnisQL: cat="S" & node & #1 >secedge[func="SB"] #2

Result: 2861

☐	Name	Texts	Tokens
☑	DDB.AHD	20	2846
☐	SMULTRON_Banana	2	3782
☐	a6.vorfelder	2156	285080
☐	arabic.test	1	11
☐	b2.hausa	50	6991
☐	d2.20samples_070202	159	2446
☑	pcc2_plus	2	399
☑	tiger2	1971	888578

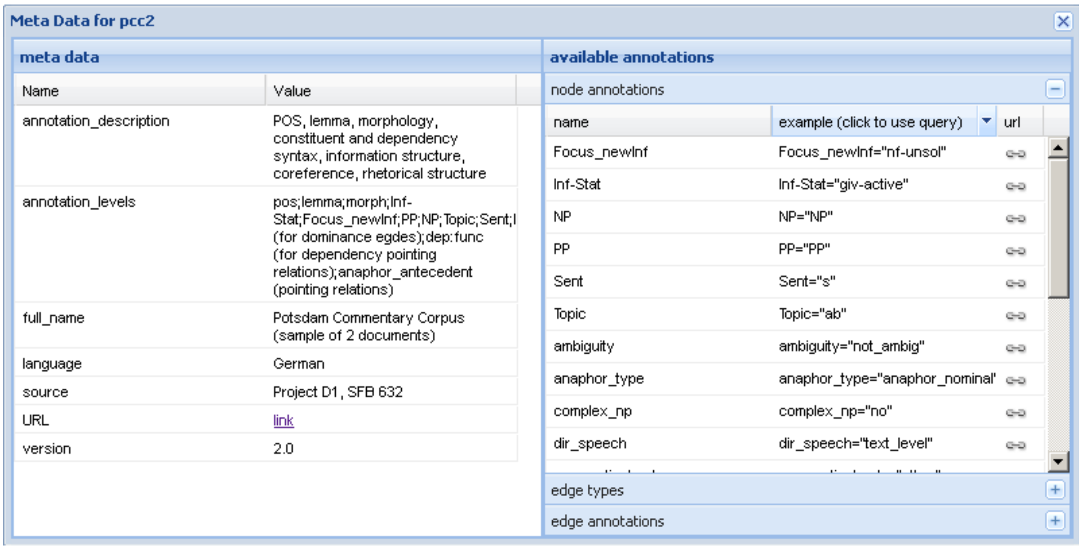
Context Left: 5

Context Right: 5

Results Per Page: 10

The Search Form on the left of the interface window is available immediately after login. In the middle, the list of currently available corpora is shown. Using the checkboxes on the left of each corpus, it is possible to select which corpora should be searched in (hold down 'shift' to select multiple corpora simultaneously). If you cannot see a corpus that should be available to you, or else if the corpora list is too cluttered, you may click on "more corpora" to open the corpora window. You may then drag and drop the desired or unwanted corpora between the list and the window.

Pressing the ⓘ button next to a corpus in the list will open the corpus explorer window (see picture below), which shows metadata for the entire corpus on the left and a list of available annotations and example queries on the right. 6 Clicking on a query will copy it to the "AnnisQL" field at the top of the form. Pressing the link icon will give you a citation link that can be used to access the query from any browser. If the corpus contains hierarchical structures, such as dominance edges or pointing relations, there will be separate segments on the right hand side of the corpus explorer to show the available edge names and annotations together with example queries. Clicking on a query will copy it to the "AnnisQL" field at the top of the form. Pressing the link icon will give you a citation link that can be used to access the query from any browser. If the corpus contains hierarchical structures, such as dominance edges or pointing relations, there will be separate segments on the right hand side of the corpus explorer to show the available edge names and annotations together with example queries.



Meta Data for pcc2

meta data

Name	Value
annotation_description	POS, lemma, morphology, constituent and dependency syntax, information structure, coreference, rhetorical structure
annotation_levels	pos;lemma;morph;Inf-Stat;Focus_newInf;PP;NP;Topic;Sent;I (for dominance egdes);dep:func (for dependency pointing relations);anaphor_antecedent (pointing relations)
full_name	Potsdam Commentary Corpus (sample of 2 documents)
language	German
source	Project D1 , SFB 632
URL	link
version	2.0

available annotations

node annotations

name	example (click to use query)	url
Focus_newInf	Focus_newInf="nf-unsol"	🔗
Inf-Stat	Inf-Stat="giv-active"	🔗
NP	NP="NP"	🔗
PP	PP="PP"	🔗
Sent	Sent="s"	🔗
Topic	Topic="ab"	🔗
ambiguity	ambiguity="not_ambig"	🔗
anaphor_type	anaphor_type="anaphor_nominal'	🔗
complex_np	complex_np="no"	🔗
dir_speech	dir_speech="text_level"	🔗

edge types

edge annotations

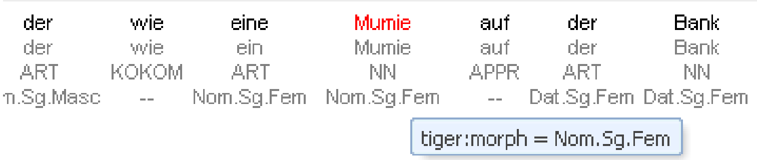
The "AnnisQL" field at the top of the form is used for inputting queries manually (see the tutorials on the ANNIS Query Language). As soon as one or several corpora are selected and a query is entered or modified, the query will be validated automatically and possible errors in the query syntax will be commented on in the "Result" box below. When modifying a query, a delay of two seconds ia activated before the query is re-sent to the server for validation.

Once a valid query has been entered, pressing the "Show Result" button (or using the shortcut ctrl +Enter) will retrieve the number of matching positions in the selected corpora in the Result box and open the Result Window to display the first set of matches. Queries from the current session are saved in the query history and can be accessed using the button underneath the result field.

The context surrounding the matching expressions in the result list is determined by the "context left" and "context right" options at the bottom of the search form, and can be set to up to 10 tokens on each side, though some corpora allow longer spans, such as entire texts, to be viewed using special discourse visualizations.

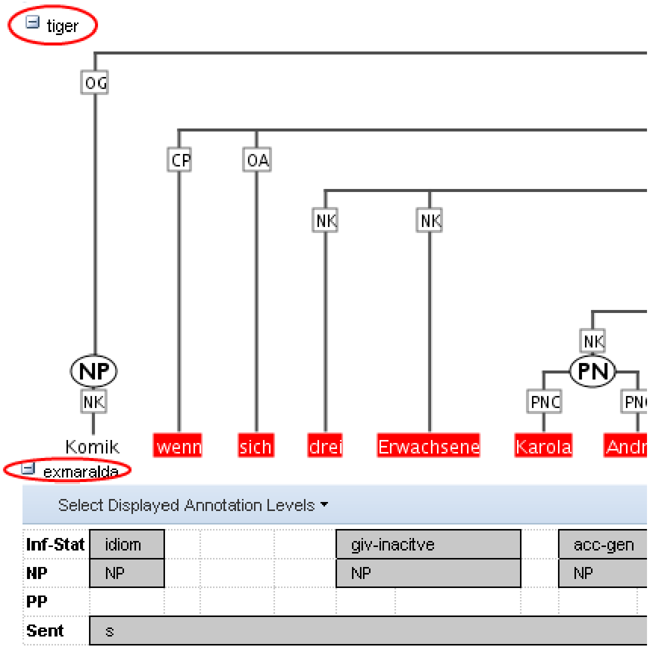
4.1.2. The Result Window

The result window shows search results in pages of 10 hits each by default (this can be changed in the Search Form). The toolbar at the top of the window allows you to navigate between these pages. The "Token Annotations" button on the toolbar allows you to toggle 7 the token based annotations, such as lemmas and parts-of-speech, on or off for you convenience. The "Citation URL" button provides a hyperlink which you can e-mail or cite, allowing others to reproduce your query.



der	wie	eine	Mumie	auf	der	Bank
der	wie	ein	Mumie	auf	der	Bank
ART	KOKOM	ART	NN	APPR	ART	NN
n.Sg.Masc	--	Nom.Sg.Fem	Nom.Sg.Fem	--	Dat.Sg.Fem	Dat.Sg.Fem

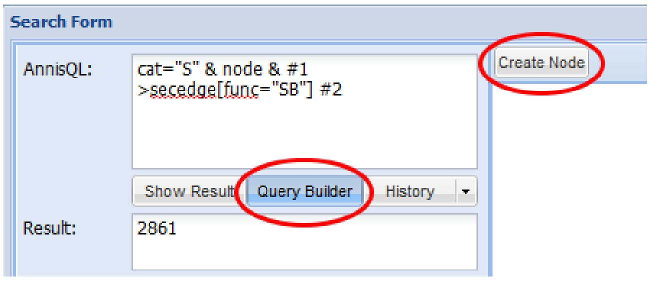
The result list itself initially shows a KWIC (key word in context) concordance of matching positions in the selected corpora, with the matching region marked red and the context in black on either side. Token annotations are displayed in gray under each token, and hovering over them with the mouse will show the annotation name and namespace. More complex annotation levels can be expanded, if available, by clicking on the plus icon next to the level's name, e.g. tiger and exmaralda for the annotations in the tree and grid views in the picture below (circled in red).



Inf-Stat	idiom		giv-inacitve	acc-gen
NP	NP		NP	NP
PP				
Sent	s			

4.2. Using the ANNIS Query Builder

To open the graphical query builder, click on the Query Builder button on the Search Form (clicking the button again will close the Query Builder). On the left-hand side of the toolbar at the top of the query builder canvas, you will see the Create Node button. Use this button to define nodes to be searched for (tokens, non-terminal nodes or annotations). Creating nodes and modifying them on the canvas will immediately update the AnnisQL field in the Search Form with your query, though updating the query on the Search Form will not create a new graph in the Query Builder.

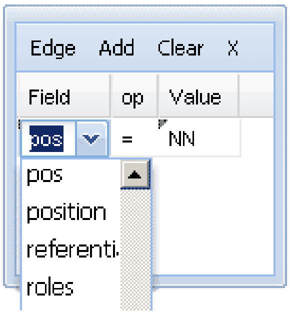


Search Form

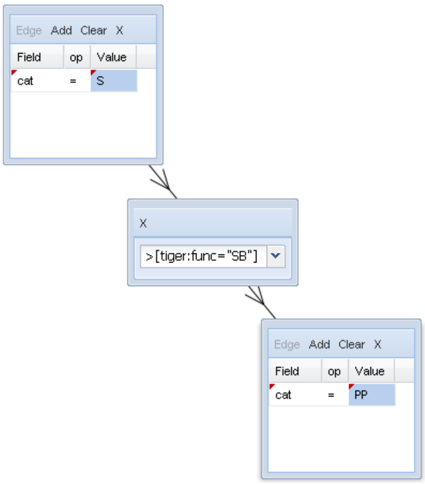
AnnisQL: cat="S" & node & #1
>secedge[func="SB"] #2

Result: 2861

In each node you create you may click on "Add" to specify an annotation value. The annotation name can be typed in or selected from a drop down list. The "Op[erator]" field in the middle allows you to choose between an exact match (the '=' symbol) or wildcard search using Regular Expressions (the '~' symbol). The annotation value is given on the right, and should NOT be surrounded by quotations (see the example below). It is also possible to specify multiple annotations applying to the same position by clicking on "Add" multiple times. Clicking on "Clear" will delete the values in the node. To search for word forms, simply leave the field name on the left empty and type directly on the right under "Value". A node with no data entered will match any node, that is an underspecified token or non-terminal node or annotation.



To specify the relationship between nodes, first click on the "Edge" button at the top left of one node, and then click the "Dock" button which becomes available on the other nodes. An edge will connect the nodes with an extra box from which operators may be selected (see below). For operators allowing additional labels (e.g. the dominance operator > allows edge labels to be specified), you may type directly into the edge's operator box, as in the example with a "func" label in the image below. Note that the node clicked on first (where the "Edge" button was clicked) will be the first node in the resulting quey, i.e. if this is the first node it will dominate the second node (#1 > #2) and not the other way around, as also represented by the arrows along the edge.



4.3. Searching for Word Forms

To search for word forms in ANNIS2, simply select a corpus (in this example the small pcc2 demo corpus) and enter a search string between double quotation marks, e.g.:

```
"statisch"
```

Note that the search is case sensitive, so it will not find cases of capitalized 'Statisch', for example at the beginning of a sentence. In order to find both options, you can either look for one form OR the other using the pipe sign (|):

```
"statisch" | "Statisch"
```

or else you can use regular expressions, which must be surrounded by slashes (/) instead of quotation marks:

```
/[Ss]tatisch/
```

To look for a sequence of multiple word forms, enter your search terms separated by & and then specify that the relation between the elements is one of precedence, as signified by the period (.) operator:

```
"so" & "statisch" & #1 . #2
```

The expression #1 . #2 signifies that the first element ("so") precedes the second element ("statisch"). For indirect precedence (where other tokens may stand between the search terms), use the .* operator:

```
/[Ss]o/ & "statisch" & "wie" & #1 . #2 & #2 .* #3
```

The above query finds sequences beginning with either "So" or "so", followed directly by "statisch", which must be followed either directly or indirectly (.*) by "wie". A range of allowed distances can also be specified numerically as follows:

```
/[Ss]tatisch/ & "wie" & #1 .1,5 #2
```

Meaning the two words may appear at a distance of 1 to 5 tokens. The operator .* allows a distance of up to 50 tokens by default, so searching with .1,50 is the same as using .* instead. Greater distances (e.g. .1,100 for 'within 100 tokens') should always be specified explicitly. Finally, we can add metadata restrictions to the query, which filter out documents not matching our definitions. Metadata attributes must be preceded by the prefix meta:: and may not be bound (i.e. they are not referred to as #1 etc. and the numbering of other elements ignores their existence):

```
/[Ss]tatisch/ & "wie" & #1 .1,5 #2 & meta::Genre="Sport"
```

To view metadata for a search result or for a corpus, press the "i" icon next to it in the result window or in the search form respectively.

4.4. Searching for Annotations

Annotations may be searched for using an annotation name and value. The names of the annotations vary from corpus to corpus, though many corpora contain part-of-speech and lemma annotations with the names pos and lemma respectively (annotation names are case sensitive). For example, to search for all forms of the German verb sein 'to be' in a corpus with lemma annotation such as pcc2, simply select the pcc2 corpus and enter:

```
lemma="sein"
```

Negative searches are also possible using != instead of =. For negated tokens (word forms) use the reserved attribute tok. For example:

```
lemma!="sein"
```

or:

```
tok!="ist"
```

Metadata can also be negated similarly:

```
lemma="sein" & meta::Genre!="Sport"
```

To only find finite forms of this verb in pcc2, use the part-of-speech (pos) annotation concurrently, and specify that both the lemma and pos should apply to the same element:

```
lemma="sein" & pos="VAFIN" & #1 _=_ #2
```

The expression #1 _=_ #2 uses the span identity operator to specify that the first annotation and the second annotation apply to exactly the same position in the corpus. Annotations can also apply to longer spans than a single token: for example, in pcc2, the annotation Inf-Stat signifies the information structure status of a discourse referent. This annotation can also apply to phrases longer than one token. The following query finds spans containing new discourse referents, not previously mentioned in the text:

```
exmaralda:Inf-Stat="new"
```

If the corpus contains no more than one annotation type named Inf-Stat, the optional namespace (in this case exmaralda:) may be dropped; if there are multiple annotations with the same name but different namespaces, dropping the namespace will find all of those annotations. In order to view the span of tokens to which this annotation applies, enter the and click on "Show Result", then open the exmaralda annotation level to view the grid containing the span. Further operators can test the relationships between potentially overlapping annotations in spans. For example, the operator _i_ examines whether one annotation fully contains the span of another annotation (the i stands for 'includes'):

```
Topic="ab" & Inf-Stat="new" & #1 _i_ #2
```