



Korpuslinguistik

Annis₃-Korpussuchtool

Suchen in tief annotierten Korpora

Anke Lüdeling, Marc Reznicek, Amir Zeldes,
Hagen Hirschmann

... und anderen Mitarbeitern der HU-Korpuslinguistik

Ziele

- Wie/Was kann man in ANNIS suchen?
 - Wortformen?
 - Linguistische Muster?
 - Token-Annotationen (Lemmata/Wortarten)?
 - Spannenannotationen
 - Syntaktische Annotationen?
 - Konstituenten
 - Abhängigkeiten
- Wie sucht man nach mehreren/beliebigen Annotationen gleichzeitig?
- Wie filtert man nach Metadaten?

Ziele

- Wie/Was kann man in ANNIS suchen?
 - Wortformen?
 - Linguistische Muster?
 - Token-Annotationen (Lemmata/Wortarten)?
 - Spannenannotationen
 - Syntaktische Annotationen?
 - Konstituenten
 - Abhängigkeiten
- Wie sucht man nach mehreren/beliebigen Annotationen gleichzeitig?
- Wie filtert man nach Metadaten?

Annis 3

- **Annis-Portal (öffentlich)**

korpling.org/annis3/

- **Annis-Portal speziell für Lernerkorpora (Falko) (öffentlich)**

korpling.org/falko-suche/

Das Web-Interface: Abfrage

<http://korpling.german.hu-berlin.de/falko-suche/search.html>

The screenshot shows the ANNIS search interface. At the top, there are navigation links for 'About ANNIS' and 'Report Problem'. Below this is a large text input area labeled 'Please enter AQL query'. To the right of the input area is a 'Query Builder' button. Below the input area are buttons for 'Search', 'More', and 'History'. A message box below the buttons says 'Welcome to ANNIS! A tutorial is available on the right side.' Below the message box are tabs for 'Corpus List' and 'Search Options'. The 'Corpus List' tab is active, showing a table of corpora. The 'Maerchenkorpus' row is highlighted in blue. To the right of the table are buttons for 'Trefferexport' and 'Statistik'.

Name	Texts	Tokens		
kobaltL2v1.4	51	33.368		
Maerchenkorpus	211	295.880		
Mercurius	2	187.423		
NoSta-D-Anselm	2	2.710		
NoSta-D-Kafka	2	10.388		

Suche-Fenster
(hier Eingabe)

Abfrage starten

Trefferexport
und Statistik

Pro Korpus Auflistung aller
Annotationen und Metadaten

Auswahl der Korpora für die Suche
(STRG+Klick für Mehrfachauswahl)

Überführung exemplarischer Anfragen ins Suchfenster

The screenshot displays the ANNIS interface. On the left, a search window contains the query `pos="NN"`. Below it are buttons for Search, More, and History, and a note: "Valid query, click on 'Search' to start searching." A "Corpus List" table shows the selected corpus:

Name	Texts	Tokens
Maerchenkorpus	211	295.880

The main area shows "Corpus information for Maerchenkorpus (ID: 9280)". It includes a metadata table:

Name	Value
Herausgeber	Maik Walter
Kontakt	walter@textbewegung.de
Kooperation	Humboldt-Universität zu Berlin, Carolin Odebrecht
Projekt	Textbewegung: Theater & Sprache www.textbewegung.de
Projektbeschreibung	Das Märchenkorpus enthält die 201 Kinder- und Hausmärchen sowie die im 2. Band abgedruckten 10 Kinderlegenden in der von den Brüder Grimm herausgegeben Ausgabe letzter Hand. Das Korpus wurde für das Vertiefungsseminar "Dramapädagogik des Märchens: Linguistik, Didaktik und Theater" kompiliert und aufbereitet. Das Vertiefungsseminar fand im Sommersemester 2013 am Deutschen Seminar der Universität Tübingen unter Leitung von Maik Walter statt (vgl. Maik Walter (i.E.): Es VERBte (ein)mal. Linguistisches Forschungstheater im Grimm-Jahr 2013. Erscheint in Zeitschrift für Theaterpädagogik 63. 29.Jahrgang. Themenheft: Forschung, Fachdiskurse & Labore).
Titel	Märchenkorpus

On the right, the "Available annotations" section shows "Node Annotations" with a table:

Name	Example (click to use query)
lemma	lemma="<unknown>"
pos	pos="NN"

Below this are sections for "Edge Annotations", "Edge Types", and "Meta Annotations".

Prinzip I: Variablen-Wert-Paare

- Jedes Textkorpus enthält eine Ebene mit fortlaufendem, tokenisiertem Text
- Meistens: "tok"
- tok = "Frau"
 - Findet alle Vorkommen von "Frau" im Text

tok	Sofern	die	Frau	herrscht
-----	--------	-----	------	----------

Prinzip I: Variablen-Wert-Paare

- tok = "Frau"

- "Finde mir alle Vorkommen von 'Frau' auf der Ebene (Variable) 'tok' "

- tok

- "Finde mir alle Vorkommen auf der Ebene 'tok' "

- Korpus explorieren

Prinzip I: Variablen-Wert-Paare

- Jedes Korpus in ANNIS enthält Annotationen (linguistische Informationen über die Sprache im Korpus)
- Unterschiedliche Annotationen werden auf Annotationsebenen gespeichert
- Meistens:
 - pos → Wortart
 - lemma → Lemma (Grundform)

Prinzip I: Variablen-Wert-Paare

- **pos = "NN"**
 - "Finde mir alle Vorkommen von 'NN' auf Ebene (Variable) 'pos' "
('NN' steht für 'normales Nomen')
- **pos**
 - "Finde mir alle Vorkommen auf der Ebene 'pos' "
 - "Finde alle Wortarten"

Prinzip I: Variablen-Wert-Paare

■ tok = "das"

Variable1
("Wortform")

Wert

tok	Sofern	das	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen

Prinzip I: Variablen-Wert-Paare

■ **pos** = **"ART"**

Variable2
("Wortart")

Wert

word	Sofern	das	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen

Prinzip I: Variablen-Wert-Paare

■ pos = "NN"

...findet *Riesen, Frauen, Student, ...*

Prinzip I: Variablen-Wert-Paare

■ lemma = "d"

Variable
("Lemma")

Wert

word	Sofern	das	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen

Prinzip I: Variablen-Wert-Paare

■ lemma = "d"

...findet *die, dem, den, ...*

beliebig erweiterbar...

■ **satz** = "NS"

Variable4
("Satztyp")

Wert

word	Sofern	das	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen
satz	NS			

beliebig erweiterbar...

- **satz = "NS"**

...findet alle Nebensätze wie
Sofern das System herrscht

(sofern die Daten wie gezeigt annotiert sind)

Suche nach Strings

- Suchen Sie nach allen Vorkommen der Wortform "meinen" in FalkoEssayL2V2.4

```
tok = "meinen"
```

- Was wird gefunden?
- Ist das interessant?
- Was wird nicht gefunden, was interessant sein könnte?

Lemmata

- "Basisformen" von Wörtern
- **Suchen Sie nach allen Vorkommen der Formen des Verbs *meinen*:**

lemma = "meinen"

- → Problem: Lemmatisierung ist willkürlich; man muss wissen, wie lemmatisiert wurde.
- Beispiel: Lemma von *sich*

Lemmata

- "Basisformen" von Wörtern
- **Suchen Sie nach allen Vorkommen der Formen des Possessivartikels:**

lemma = "mein"

Mustersuche (reguläre Ausdrücke)

- Annis erlaubt Mustersuchen auf allen Annotationsebenen
- Mustersuchen werden statt in " " in // eingefügt
- Z. B. kann man damit nach allen Wörtern suchen, die *...mein...* enthalten.

```
tok = /.*mein.*/
```

Mustersuche: Joker .

- ein beliebiges Zeichen al. → *als*, *alt*, ...
- ■ zwei beliebige Zeichen al.. → *alle*, *alte*, *also*
- ■ ■ drei beliebige Zeichen al... → *alles*, *altes*,
alias, ...

Aufgabe

- Welche Wortformen bekommen Sie mit?

tok = /g.b./

Mustersuche: ? und * +

das[↷]s?

das vorherige Zeichen ist optional
→ ϕ , s → da, das

das[↷]s*

das vorh. Zeichen kommt 0- bis ∞ mal vor
→ ϕ , s, ss, ... → da, das, dass, dassssssssss

das[↷]s+

das vorh. Zeichen kommt 1- bis ∞ mal vor
→ s, ss, ... → das, dass, dassssssssssss

Aufgabe

- Was passiert, wenn Sie die Operatoren kombinieren?

```
tok = /Mann.*/
```

```
tok = / Mann.* /
```

```
tok = / Mann.+ /
```

Aufgabe

- **Versuchen Sie alle Wörter (Grundformen) zu finden, die auf *-lang* enden.**

Aufgabe

- Versuchen Sie alle Wörter (Grundformen) zu finden, die auf *-lang* enden.

lemma = */.+lang/*

Treffer z.B.:

bislang

lebenslang

jahrelang

Aufgabe

- Versuchen Sie alle Wörter (Grundformen) zu finden, die mit *lang-* beginnen.

lemma = /lang.+/

Treffer z.B.:

lange

langsam

langweilig

Gruppieren mit ()

- mit () kann man Ausdrücke als zusammengehörige Gruppen behandeln

```
tok = /(ja)+/
```

findet

ja

jaja

jajaja

...

Alternativen: a oder b = **(a|b)**

- Mit Klammern und | ("oder") kann man gleichzeitig nach verschiedenen Wörtern suchen:

```
tok = /(Mann|Frau|Kind)/
```

- Nach verschiedenen Formen:

```
tok = /(Mann|Mannes)/
```

- Oder Zeichenketten:

```
tok=/bes(ser|t).?/
```

Aufgabe

- Finden Sie alle Formen des Verbs *meinen* im Präsens, aber keine anderen Formen.

mein	e
mein	st
mein	t
mein	en
mein	t
mein	en

Lösung

```
tok =/mein(s?t|en?)/
```

oder

```
tok =/mein(e|st|t|en)/
```

→ Häufig gibt es alternative Suchanfragen für dieselben Treffermengen.

Lösung

t (e|st|t|en?)/

oder

tok =/mein(e|st|t|en)

► Problem:
- Infinitiv
- Formen des Possessivpronomens

Suche nach Wortart

- Es gibt unterschiedliche Wortartensysteme (→ Tagsets) für Korpora
- allgemein in der Linguistik unterschiedliche Wortartensysteme
- Die meisten deutschen Korpora benutzen das Tagset STTS
 - ADJA attributives Adjektiv
 - ADV Adverb
 - ART Artikel
 - NN normales Nomen
 - VVFIN finites Verb
 - ...

<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>

Stuttgart-Tübingen-Tagset (STTS)

ADJektiv	Nomen	Pronomen	Verb	Partikel	Konjunktion
ADJA	NN	PDS	VFIN	PTKZU	KOUI
ADJD	NE	PDAT	VVIMP	PTKNEG	KOUS
		PIS	VVINF	PTKVZ	KON
		PIAT	VVIZU	PTKANT	
		PIDAT	VVPP	PTKA	
		PPER	VAFIN		
		PPOSS	VAIMP		
		PPOSAT	VAINF		
		PRELS	VAPP		
		PRELAT	VMFIN		
		PRF	VMINF		
		PWS	VMPP		
		PWAT			
		PWAV			

Stuttgart-Tübingen-Tagset (STTS)

VERB	Vollverb	Auxiliar	Modalverb
finit	VV FIN	V AFIN	V MFIN
Imperativ	VV IMP	V AIMP	
infinit	VV INF	V AINF	V MINF
Infinitiv mit <i>zu</i>	VV IZU		
Partizip 2	VV PP	V APP	V MPP

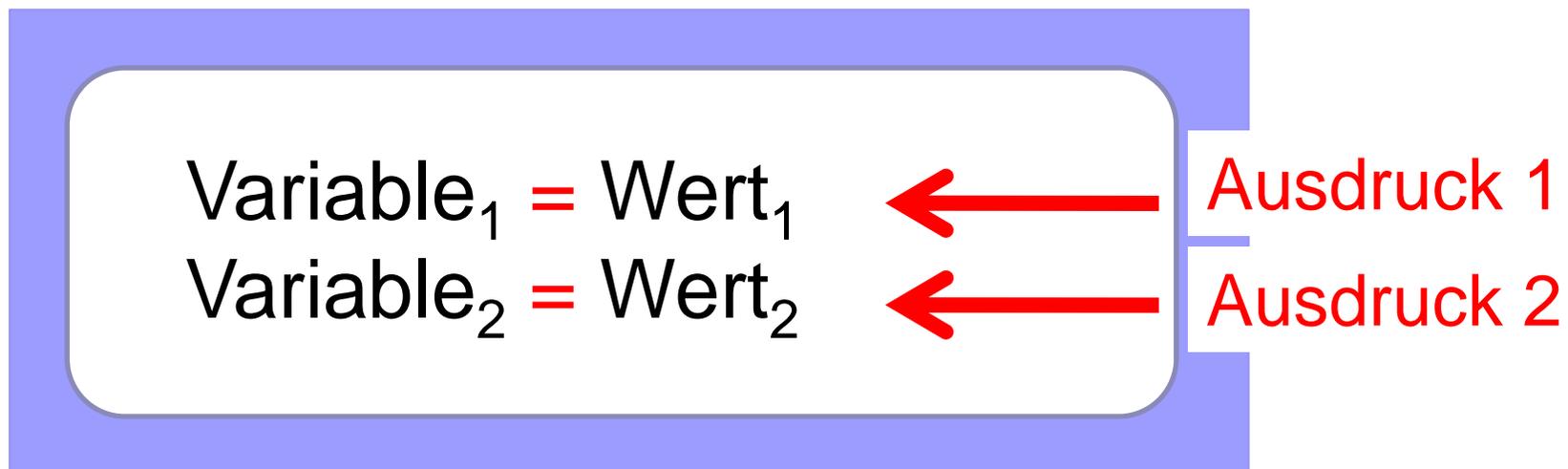
Aufgabe

- Suchen Sie in DDB.FNHD nach Possessiva

```
pos =/PPOS(S|AT)/
```

Prinzip II: Mehrere Suchbedingungen eingeben und Verknüpfen

- Wir wollen mehrere Bedingungen an die Suche stellen, z. B. nach dem Lemma "auf" suchen, und dies soll gleichzeitig die Wortart "Verbpartikel" haben, weil wir keine Präpositionen finden wollen
- Einzelne **Variable-Wert**-Paare werden durch einen passenden Befehl verbunden.



Prinzip II: Verknüpfungstypen

- Wir können bereits die zwei Suchanforderungen für "Such mir alle Vorkommen von "auf", die Verbpartikeln sind" formulieren:
`lemma = "auf"`
`pos="PTKVZ"`
- Wir benötigen nun die passende Verknüpfung dieser Suchanforderungen

Sämtliche Verknüpfungstypen in ANNIS

The screenshot shows the ANNIS interface with a search query `tok = "meinen" pos="PPOSAT"` and a search error message: `line 1:16 mismatched input 'pos' expecting {<EOF>, '&', '|'}`. A dropdown menu is open, showing options like 'ANNIS interface' and 'ANNIS Query language'. A red box highlights the text 'Suchen Sie den passenden Typ' (Search for the appropriate type) and points to the 'Complete List of Operators' option in the menu.

Operator	Description	Symbol	Applicability
.	direct		
>*	inclusion		
==	identical coverage	A B	Applies when two annotation
i	inclusion	AAA B	Applies when one annotation

Sämtliche Verknüpfungstypen in ANNIS

The screenshot shows the ANNIS interface with a search query `tok = "meinen" pos="PPOSAT"` and a corpus list. A dropdown menu is open, showing a list of operators. A red box highlights the text "Suchen Sie den passenden Typ" (Search for the appropriate type) and a red circle highlights the operator `=` in the table.

Visible: All

Name	Texts	Texts		
FalkoEssayL1v2.0				
falkoEssayL1v2.3		10.615		
FalkoEssayL2v2.0	248	132.066		
FalkoEssayL2v2.3	248	131.628		
falkoEssayL2v2.4	248	144.619		
FalkoEssayL2WHIGv2.0	195	130.187		
FalkoGeorgetownL2v1.0	92	78.151		

Operator	Description	Symbol	Applicability
.	direct		
>	directional		
>*	indirect		
>+	dominant		
=	identical coverage	A B	Applies when two annotation
i	inclusion	AAA B	Applies when one annotation

Lösung

- Suchanfrage für *auf* in der Funktion einer **Verbpartikel**

lemma = "auf"
=
pos= "PTKVZ"

Aufgabe

- Finden Sie nun im Märchenkorpus Vorkommen von lemma = "gehen", die ausschließlich Imperative sind.

lemma = "gehen"
=
pos = "VVIMP"

Negation !=

- **!** bedeutet Negation
 - Der Operator wird vor dem "="-Zeichen eingefügt.
 - **Finden Sie im Märchenkorpus alle Vorkommen von lemma ="denn" , die keine Konjunktionen sind. Welche Wortart ist das? Welches Tag wird vergeben?**

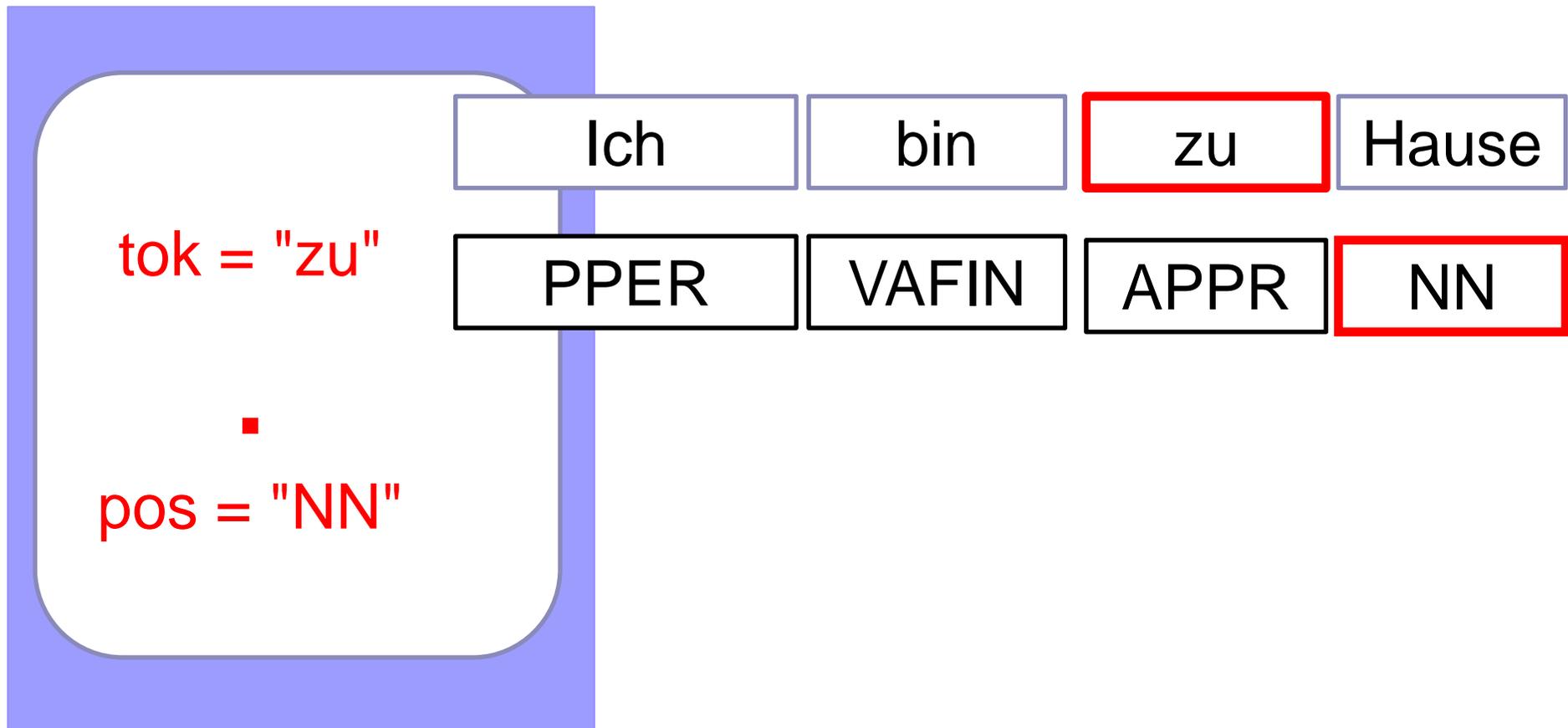
```
lemma ="denn" _=_ pos!="KON"
```

Aufgabe

- Finden Sie im Märchenkorpus alle **Abfolgen** von Präpositionen und Nomina (ohne Artikel oder Adjektiv dazwischen)
- (→Wieder zwei Bedingungen, aber jetzt ist die Verknüpfung der Bedingungen eine andere)

pos = "APPR" . pos = "NN"

Suche nach Abfolgen: z.B. Nomen folgt auf "zu"



Aufgabe

- Finden Sie in DDB.MHD alle Satzbeendungszeichen, die drei Stellen hinter einem finiten Modalverb stehen

```
pos="VMFIN" .3 pos =^$\./
```

Ambige Zeichen: Operatoren und Wort-/Satzzeichen

- Erinnerung: Der Operator " **.** " bedeutet "**ein beliebiges Zeichen**", wenn ich die Suche in **//** formuliere
- Wie finde ich aber Wörter mit Abkürzungspunkten, wenn ich in der Suchanfrage reguläre Ausdrücke verwende?
- Der Operator "****" bedeutet "**das folgende Zeichen ist wörtlich gemeint**".

Ambige Zeichen: Operatoren und Wort-/Satzzeichen

- `tok="usw."`
→ Findet alle Vorkommen von ***usw.***
- `tok=/(B|b)./`
→ Findet in `falkoEssayL2v2.4`
B. und ***b.*** (Kontext: z. ***B.***),
aber auch ***BA*** und ***BH***
- `tok=/(b|B)\./`
→ Findet nur ***B.*** und ***b.***

Aufgabe - Statistik

- Finden Sie in DDB.AHD alle Präpositionen
- Lassen Sie sich mithilfe der Funktion **More>Frequency Analysis** die Lemmata anzeigen
- Sie müssen hierzu die Suchanfrage **pos="APPR"** formulieren und in das Feld **Select annotation of node** ***lemma*** eingeben (dann unten Mitte abschicken)

Zusammenfassung: reguläre Ausdrücke

.	Ein beliebiges Zeichen
*	Beliebig viel (0 bis unendlich vom vorherigen Element)
+	Mindestens einmal (vorheriges Element)
?	Optional (vorheriges Element)
\	wörtlich (folgendes Zeichen)
!	nicht
[abc]	Menge (oder $[^abc]$ = alles <i>außer</i> die Menge)
(a b)	a oder b
a{2,3}	a 2 bis 3 mal

Zusammenfassung: Operatoren zur Verknüpfung von Suchanforderungen

X **.** **Y** \rightarrow **Y folgt direkt auf X**

X **.2** **Y** \rightarrow

Y folgt nach 2 Stellen auf X

X **.1,3** **Y** \rightarrow **Y folgt im Abstand von 1-3**

X **.******* **Y** \rightarrow **Y folgt irgendwann auf X**

X **_****=****_** **Y** \rightarrow **Y deckt sich ab mit X**

(bzw.: **_** **X** und **Y** treffen auf dasselbe Element zu)

Zusammenfassung

- Mit ANNIS kann man:
 - in unterschiedlichen Korpora (auch gleichzeitig) suchen
 - die Ergebnisse quantifizieren
 - die Ergebnisse exportieren
- Man kann auch nach Metadaten filtern (Tutorial)

Herzlichen Dank!