



Korpuslinguistik

Annis₂-Korpussuchtool

Suchen in Falko

Marc Reznicek, Anke Lüdeling, Amir Zeldes,
Hagen Hirschmann

hirschhx@hu-berlin.de

... und vielen anderen Mitarbeitern der HU-Korpuslinguistik

Ziele

- Wie/Was kann man in ANNIS/Falko suchen?
 - Wortformen?
 - Linguistische Muster?
 - Token-Annotationen (Lemmata/Wortarten)?
 - Syntaktische Annotationen?
 - Zielhypothesen/ Edit Tags?
- Wie sucht man nach mehreren/ beliebigen Annotationen gleichzeitig?

Überblick

- generell: ANNIS (kurz)
- Abfragen
 - reguläre Ausdrücke
 - Tokenabfolgen/ beliebige andere Abfolgen
 - auf mehreren Ebenen
 - Wortarten
 - Lemmata
 - Zielhypothesen/ Edit Tags
 - * neu: Dependenz-Analysen & topologische Felder
- Metadaten

Annis₂ SFB 632



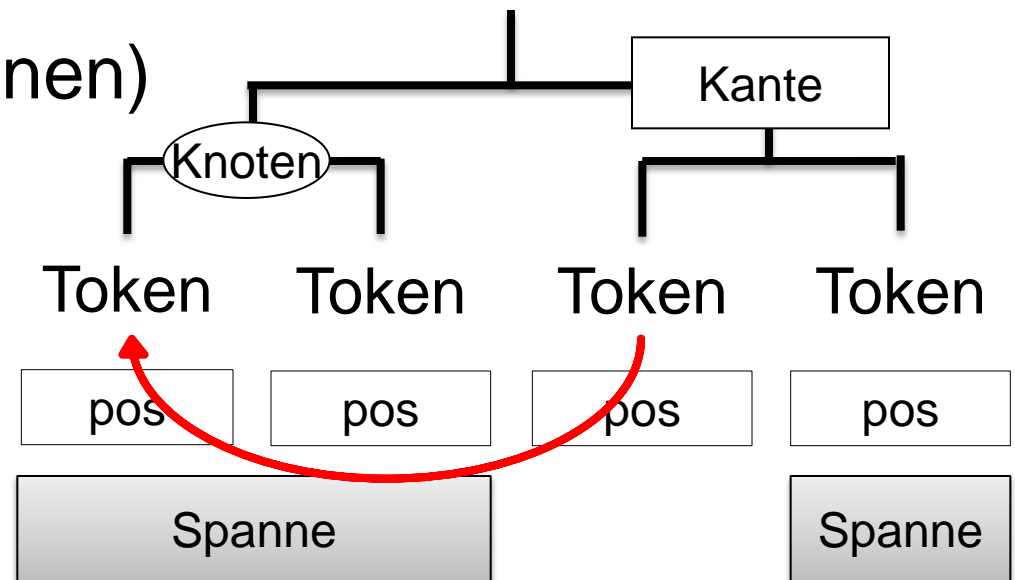
ANNotation of Information Structure

(Dipper et al. 2004; Chiarcos et al. 2008; Zeldes et al. 2009)

<http://www.sfb632.uni-potsdam.de/d1/annis/>

■ Suchmaschine für tief annotierte, multimodale Korpora

- Token (-annotationen)
- Spannen
- Bäume
- Pointer



Das Web-Interface: Tutorial

ANNIS² ▾ Tutorial logged in as "demo"

Search Form

AnnisQL:

Query Builder:

Result:

More Corpora

| <input type="checkbox"/> | Name ▲ | Texts | Tokens | |
|--------------------------|--------------------|-------|--------|--|
| <input type="checkbox"/> | FalkoEssayL1V2_0 | 95 | 70608 | |
| <input type="checkbox"/> | FalkoEssayL2V2_0 | 248 | 131599 | |
| <input type="checkbox"/> | FalkoSummaryL1V1_2 | 57 | 21211 | |

Context Left: ▾

Context Right: ▾

Results per page: ▾

1) Tutorial öffnen

Das Web-Interface: Tutorial

The screenshot displays the ANNIS2 web interface. On the left is the search form, and on the right is a help window titled 'Using the ANNIS2 Interface'. A red arrow points to a dropdown menu in the help window that currently shows 'Using the ANNIS2 interface'. A red box with the text '2) Thema wählen' is overlaid on the help window, indicating the next step in the tutorial.

ANNIS² Tutorial

Search Form

AnnisQL: "gleich"

Query Builder: Show >>

Result: Valid Query

More Corpora

| <input type="checkbox"/> | Name ▲ | Texts |
|--------------------------|--------------------|-------|
| <input type="checkbox"/> | FalkoEssayL1V2_0 | 95 |
| <input type="checkbox"/> | FalkoEssayL2V2_0 | 248 |
| <input type="checkbox"/> | FalkoSummaryL1V1_2 | 57 |

Search Export

Context Left: 5

Context Right: 5

Results per page: 10

Show Result

Get help about: Using the ANNIS2 interface

Using the ANNIS2 Interface

The ANNIS2 interface is comprised of several windows, the most important of which are the search form and the results window.

The Search Form

Search Form

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

Query Builder: Show >>

Result: Valid Query

More Corpora

| <input checked="" type="checkbox"/> | Name ▲ | Texts | Tokens | |
|-------------------------------------|--------------------------|-------|--------|---|
| <input checked="" type="checkbox"/> | GermanDiachronicTreeb... | 35 | 9540 | i |
| <input type="checkbox"/> | WikiNews | 1 | 130 | i |
| <input type="checkbox"/> | arabic.test | 1 | 11 | i |

2) Thema wählen

Das Web-Interface: Abfrage

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

ANNIS? Tutorial logged in as "demo"

Search Form

AnnisQL: "gleich"

Query Builder: Show >>

Result: Valid Query

More Corpora

| <input type="checkbox"/> | Name | Texts | Tokens | |
|-------------------------------------|--------------------|-------|--------|--|
| <input type="checkbox"/> | FalkoEssayL1V2_0 | 95 | 70608 | |
| <input checked="" type="checkbox"/> | FalkoEssayL2V2_0 | 248 | 131599 | |
| <input type="checkbox"/> | FalkoSummaryL1V1_2 | 57 | 21211 | |
| <input type="checkbox"/> | FalkoSummaryL2V1_2 | 107 | 40865 | |
| <input type="checkbox"/> | FalkoSummaryLV1_0 | 12 | 11114 | |

Search Export

Context Left: 5

Context Right: 5

Results per page: 10

Show Result

Anfragefenster

Abfrageprüfung und Anzahl der Treffer

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

Linker / rechter Kontext Anzahl der Treffer

Abfrage starten

Das Web-Interface: Suchfenster

Search Form

AnnisQL: word = "gleich"

1) Suchanfragefenster

Query Builder: Hide <<

Result: 40

2) Trefferanzahl

More Corpora

The image shows a screenshot of a search interface titled "Search Form". It is divided into three main sections: "AnnisQL:", "Query Builder:", and "Result:". The "AnnisQL:" section contains a text input field with the query "word = \"gleich\"", where the word "gleich" is underlined in red. A red arrow points from a callout box labeled "1) Suchanfragefenster" to this underlined text. The "Query Builder:" section contains a button labeled "Hide <<". The "Result:" section displays the number "40". A red arrow points from a callout box labeled "2) Trefferanzahl" to the number "40". At the bottom left, there is a button labeled "More Corpora".

Das Web-Interface: Such-Einstellungen

The image shows a screenshot of a web interface for search settings. The interface is light blue and contains several elements:

- At the top left, there are two buttons: "Search" and "Export". The "Search" button is circled in red.
- Below the buttons, there are three rows of settings:
 - "Context Left:" with a value of "5". A red arrow points from a box labeled "linker Kontext" to this value.
 - "Context Right:" with a value of "5". A red arrow points from a box labeled "rechter Kontext" to this value.
 - "Results per page:" with a value of "10". A red arrow points from a box labeled "Resultate pro Seite" to this value.
- At the bottom center, there is a button labeled "ShowResult". A red arrow points from a box labeled "Suche starten" to this button.

Das Web-Interface: Treffer

Search Result - word="gleich" (5, 5)

Page 1 of 4 Token Annotations Show Citation URL Displaying Results 1 - 10 of 39

falko (grid)

Select Displayed Annotation Levels ▾

| | | | | | | | | | | | |
|-------|-------------|-------|--------|-----|--------|--------|-----|-------|------|------|------|
| lemma | Berufsleben | sein | Frau | und | Mann | gleich | . | sein | es | aber | gut |
| word | Berufsleben | sind | Frauen | und | Männer | gleich | . | Ist | es | aber | gut |
| pos | NN | VAFIN | NN | KON | NN | ADJD | \$. | VAFIN | PPER | ADV | ADJD |
| tok | Berufsleben | sind | Frauen | und | Männer | gleich | . | Ist | es | aber | gut |

ZH1 (grid)

Select Displayed Annotation Levels ▾

| | | | | | | | | | | | |
|----------|-------------|-------|--------|-----|--------|--------|---|-------|------|------|------|
| ZH1lemma | Berufsleben | sein | Frau | und | Mann | gleich | . | sein | es | aber | gut |
| ZH1pos | NN | VAFIN | NN | KON | NN | PTKVZ | . | VAFIN | PPER | ADV | ADJD |
| ZH1 | Berufsleben | sind | Frauen | und | Männer | gleich | . | Ist | es | aber | gut |
| tok | Berufsleben | sind | Frauen | und | Männer | gleich | . | Ist | es | aber | gut |

ZH2 (grid)
ZHverb (grid)
text (grid)
Volltext

" Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt " Früher hatten die Frauen viel weniger Freiheit als heute . Es gab eine riesige Menge von gesellschaftlichen Verurteilen , wie sich die Frauen benehmen und sogar leben sollten . Nehmen wir zum Beispiel den bekannten Regel von " 3K " . Es ist verständlich , was von Frauen damals erwartet wurde und es ist auch klar , dass solch ein Leben nicht alle befriedigen konnte . Die Frauen hatten den Wunsch , am gesellschaftlichen Leben teilzunehmen , gleich wie Männer zu arbeiten und zu verdienen , aber nicht zu Hause mit den Kinder zu bleiben . Sie haben dafür gekämpft und den Kampf gewonnen . Heute gibt es keine Gesetze mehr , die den Frauen eine bestimmte gesellschaftliche Rolle vorschreiben oder ihre Möglichkeiten und Rechte begrenzen . Frauen und Männer haben gleiche Chancen im Politik , Beruf oder in der Gesellschaft . Natürlich ist es nicht überall so . Bis jetzt existieren nicht nur in den sogenannten Ländern der 3. Welt , sondern auch in den hochentwickelten Staaten verschiedene Unternehmen , in denen die Frauen weniger verdienen nur deswegen , weil sie Frauen sind , obwohl sie die gleiche Arbeit machen . Schon heute , wenn die Frauen das Gewünschte erreicht haben , taucht sofort ein neues Problem auf : der Feminismus scheint mehr ein Übel , als eine langerwartete Freiheit zu sein . Das lässt sich in allen Bereichen des Lebens merken . Im Berufsleben sind Frauen und Männer gleich . Ist es aber gut ? Dieser Regel von " 3K " hat aus dem Frauenleben nicht verschwunden . Sie haben viel Hausarbeit zu tun . Ganz oft achtet man darauf erst dann , wenn sie nicht gemacht wird . Wenn es in der Familie kleine Kinder gibt

Partitur Lernertext

Partitur Zielhypothese I

Weitere Partituransichten

Prinzip I: Variablen-Wert-Paare

■ word = "das"

Variable1
("Wortform")

Wert

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

Prinzip I: Variablen-Wert-Paare

■ word = "das"

...findet *das* (und nichts anderes)

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"

Variable3
("Lemma")

Wert

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

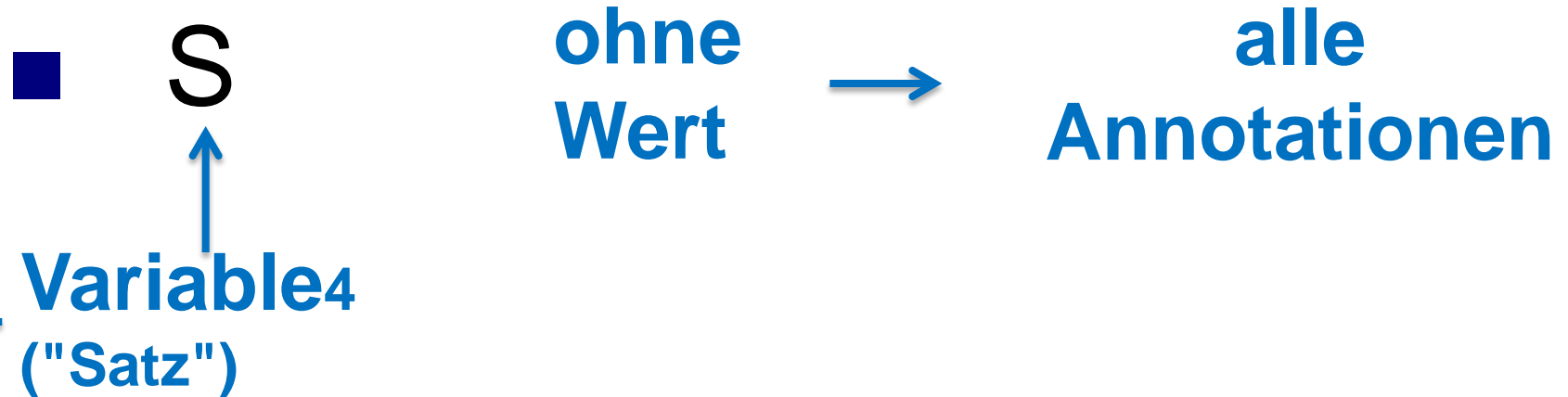
Prinzip I: Variablen-Wert-Paare

■ lemma = "d"

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

Alle Annotationen einer Ebene

Als positiv und negativ, kan die Globalisierung sein.



| | | | | |
|----------|-----------|---------|-----|---------|
| word | Als | positiv | und | negativ |
| pos | KOKOM | ADJD | KON | ADJD |
| lemma | als | positiv | und | negativ |
| S | s3 | | | |

Suche nach Strings

- Suchen Sie nach allen Vorkommen der Wortform "meinen" in FalkoEssayL2V2.3

```
word = "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)

- Annis₂ erlaubt Mustersuchen auf allen Annotationsebenen
- Mustersuchen werden statt in " " in // eingefügt
- Z. B. kann man damit nach allen Wörtern suchen, die *...mein...* enthalten.

```
word = /.*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?

word = /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?

```
word = /Frau.?!/
```

```
word = /Frau.* /
```


```
word = /Frau.+ /
```

Aufgabe

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

Aufgabe

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



word = /lang.*/

Treffer z.B.:

lange

langsam

langweilig

Aufgabe

- Versuchen Sie alle Wörter zu finden, deren Stammform auf *-lang* endet.

→ lemma = `/.*lang/`

Treffer z.B.:

bislang

lebenslang

jahrelang

Gruppieren mit **()**

- mit **()** kann man Ausdrücke als zusammengehörige Gruppen behandeln

```
word = /(ja)+/
```

findet

ja

jaja

jajaja

...

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

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

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

- Nach verschiedenen Formen:

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

- Oder Zeichenketten:

```
word=/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 |

Lösungen

```
word =/(meine|meinst|meint|meinen)/
```

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

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

→ Häufig gibt es alternative Suchanfragen für dieselben Treffermengen.

Lösungen

word =/(meine|meinst|meint|meinen)/

▶ **Problem:**
- Infinitiv
- Formen des Possessivpronomens

word =/(tlen)/

word =/mein/

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)

| ADJ ektiv | N omen | P ronomen | V erb | P ar Ti kel | KO njunktion |
|------------------|---------------|------------------|--------------|---------------------------|---------------------|
| ADJA | NN | PDS | VFIN | PTKZU | KOUI |
| ADJD | NE | PDAT | VVIMP | PTKNEG | KOUS |
| | | PIS | VVIN | 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 | VAFIN | VMFIN |
| Imperativ | VV IMP | VAIMP | |
| infinit | VV INF | VAINF | VMINF |
| Infinitiv mit <i>zu</i> | VV IZU | | |
| Partizip 2 | VV PP | VAPP | VMPP |

Aufgabe

- Suchen Sie nach Possesivpronomen

Achtung, es gibt zwei Arten:

substituierendes / attributives

pos =/PPOS(S|AT)/

Prinzip II: Relationen

- Einzelne **Variable-Wert**-Paare werden durch "&" verbunden.
- Zwischen den Paaren muss **IMMER eine Beziehung hergestellt** werden
- Auf die VW-Paare bezieht man sich mit # der Reihe nach.

Variable₁ = Wert₁ & ← Ausdruck 1: #1
Variable₂ = Wert₂ & ← Ausdruck 2: #2
#1 "Beziehung" #2

Prinzip II: Relationstypen

Variable₁ = Wert₁ & ← Ausdruck 1: #1
 Variable₂ = Wert₂ & ← Ausdruck 2: #2
 #1 "Beziehung" #2

| Operator | Description | Illustration | Notes |
|----------|---------------------|--------------|---|
| . | direct precedence | A B | For non-terminal nodes, precedence is determined by the right most and left most terminal children |
| .* | indirect precedence | A x y z B | For specific sizes of precedence spans, .n,m can be used, e.g. .3,4 - between 3 and 4 token distance |
| _=_ | identical coverage | A B | Applies when two annotation cover the exact same span of tokens |
| _i_ | inclusion | AAA B | Applies when one annotation covers a span identical to or larger than another |
| > | direct dominance | A B | A specific edge type may be specified, e.g.: >secedge to find secondary edges. Edges labels are specified in brackets, e.g. >[func="OA"] for an edge with the |

Multiple Angaben zum gleichen Token

==

- Der Operator **==** bezieht verschiedene Angaben auf das gleiche Token

word = "der" &

pos= "PDS" &

#1 == #2



Ausdruck 1: #1

Ausdruck 2: #2

Relation:
Abdeckung

Aufgabe

- Finden Sie nun Vorkommen von `word =/mein(e|st|t|en)/`, die ausschließlich finite Vollverben sind.

`word =/mein(e|st|t|en)/ &`

`pos= "VVFIN" &`

`#1 _ = _ #2`

Ausdruck 1: #1

Ausdruck 2: #2

Relation: Abdeckung

Negation !=

- **!** bedeutet Negation

- Der Operator wird vor dem "="-Zeichen eingefügt.

- Finden Sie in alle Vorkommen von**
word =/mein(e|st|t|en)/ , die nicht das Lemma
"mein" haben.

```
word =/mein(e|st|t|en)/ &  
lemma!="mein" &  
#1 _=_ #2
```

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

word = "zu"

&

pos = "NN"

&

#1.#2

Ich

bin

zu

Hause

PPER

VAFIN

APPR

NN

Satz

Subj

Adv

Tokenfolgen - Aufgabe

- Suchen Sie nach zwei aufeinanderfolgenden Adjektiven.
- Achtung: Es gibt zwei Typen von Adjektiven
ADJA & ADJD

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

Zielhypothesen

Unterschiede zwischen Zielhypothese und Originaltext sind durch "edit tags" auf der Diff-Ebene markiert.

| | | | | | | | | | |
|----------|------|------|----------|-------|---------|-----|--------------|-----------|-----|
| ZH1lemma | weil | sie | | ein | Aspekt | d | Gesellschaft | entdecken | , |
| ZH1Diff | | | MOVS | CHA | CHA | | | MOVT | |
| ZH1pos | KOUS | PPER | | ART | NN | ART | NN | VVPP | ,\$ |
| ZH1 | weil | sie | | einen | Aspekt | der | Gesellschaft | entdeckt | , |
| tok | weil | sie | entdeckt | eine | Aspekte | der | Gesellschaft | | , |

| | | | | |
|----------|-------|-----|---------|--------|
| ZH1lemma | wie | d | ander | Frau |
| ZH1Diff | | | CHA | |
| ZH1pos | KOKOM | ART | ADJA | NN |
| ZH1 | wie | die | anderen | Frauen |
| tok | wie | die | andere | Frauen |

Edit Tags

| ZHDiff | Operation in Zielhypothese |
|---------------|--------------------------------------|
| INS | Token eingefügt |
| DEL | Token gelöscht |
| CHA | Token geändert |
| MERGE | mehrere Token verbunden |
| SPLIT | Token in mehrere aufgespalten |
| MOVS | Token von hier bewegt |
| MOVT | Token hierhin bewegt |

Aufgabe

- Finden Sie alle Reflexivpronomen, die in den Lernertexten fehlen (erst einmal theoretisch).
- Jetzt praktisch:
Die Ebene für die ZH1-Wortart heißt ZH1pos. Die Ebene für die Edit Tags heißt ZH1Diff.

```
ZH1pos="PRF" &  
ZH1Diff="INS" &  
#1_=_#2
```

Aufgabe

- Wie oft benutzen Lerner Dativ "dem" statt Nominativ "der" ?

Lösung:

```
word="dem"&  
ZH1="den"&  
#1_=_#2
```

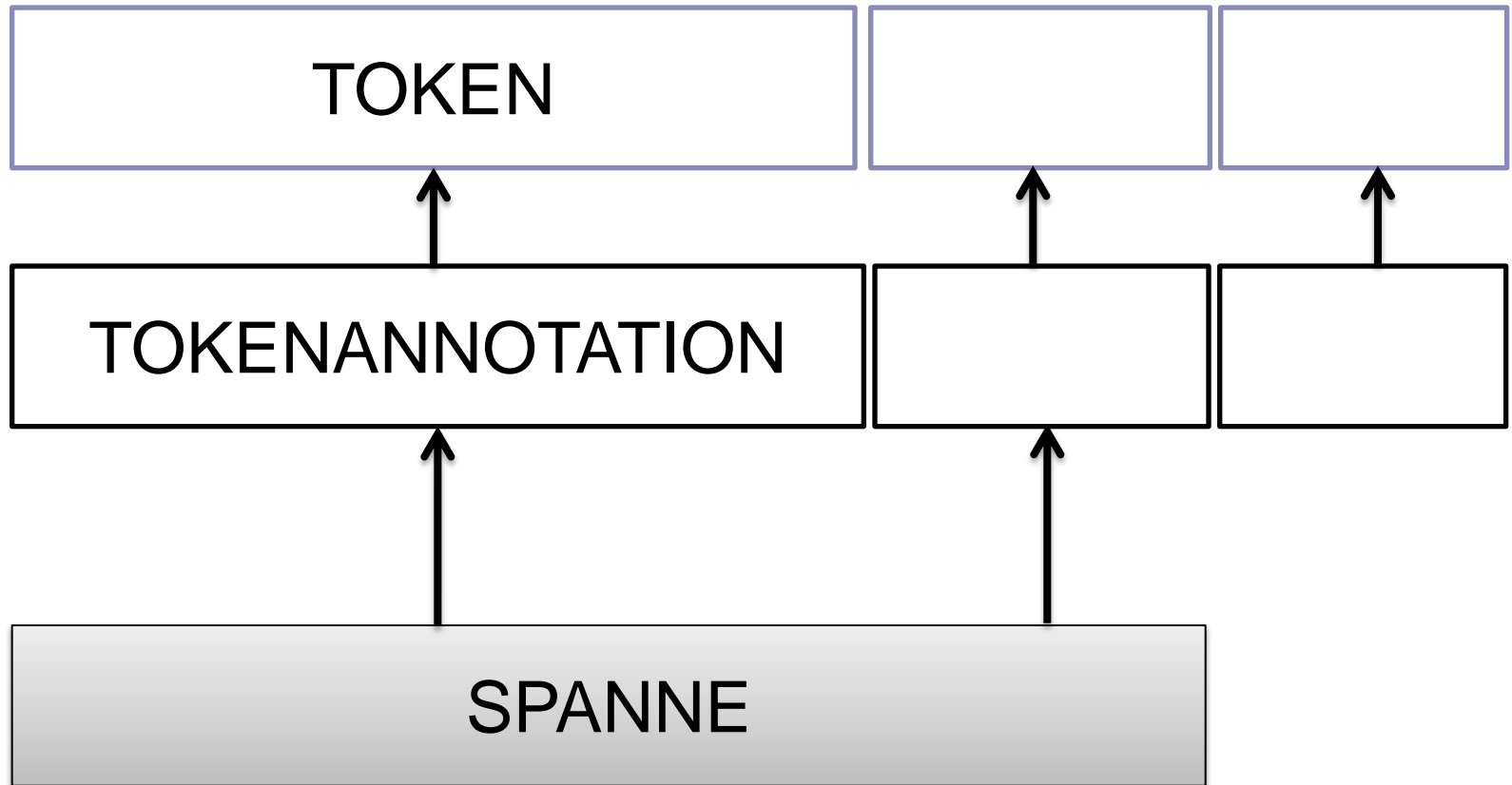

Aufgabe

- Wie oft benutzen Lerner Dativ "dem" statt Nominativ "der" ?
- Wie oft passiert das nach einer Präposition?

Lösung:

```
word="dem"& ZH1="den"& #1_=_#2 &  
ZH1pos = "APPR" & #3.#2
```

Suche auf mehreren Ebenen



Suche auf mehreren Ebenen

| | | | | |
|------------------|-------|-------|------|-------|
| word | Er | ist | zu | Hause |
| pos | PPER | VAFIN | APPR | NN |
| Root | SIMPX | | | |
| TopoField | VF | LK | MF | |

Suche auf mehreren Ebenen

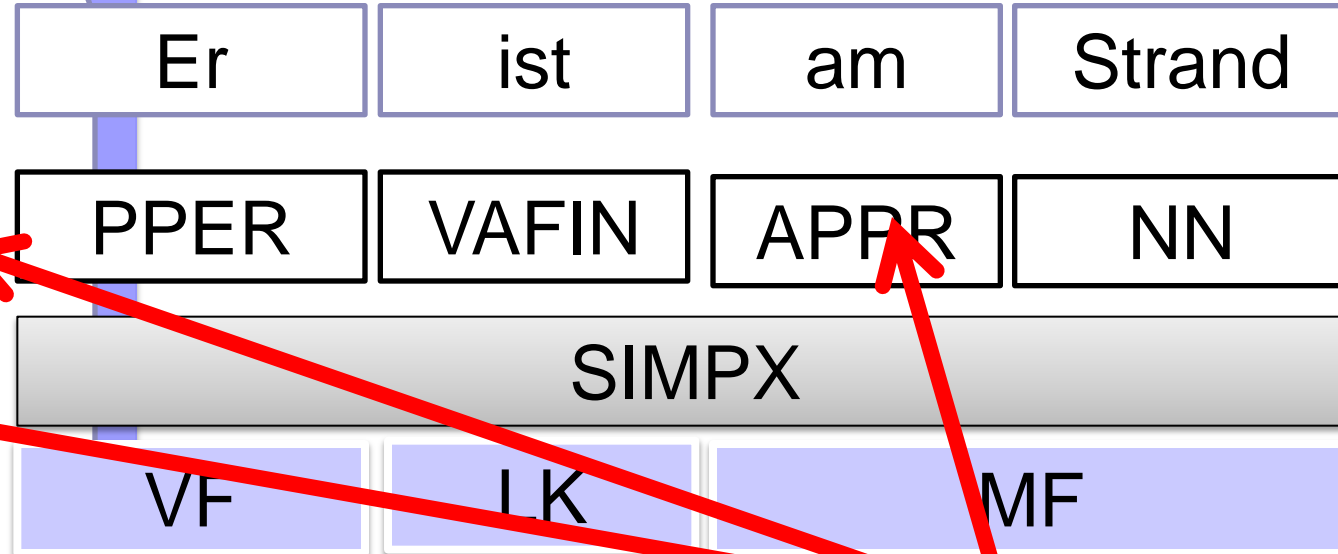
word = "zu"
&
pos = "APPR"
&
#1_ = _#2

| | | | |
|-------|-------|------|-------|
| Er | ist | zu | Hause |
| PPER | VAFIN | APPR | NN |
| SIMPX | | | |
| VF | LK | MF | |

Ausdruck 1: #1

Suche auf mehreren Ebenen

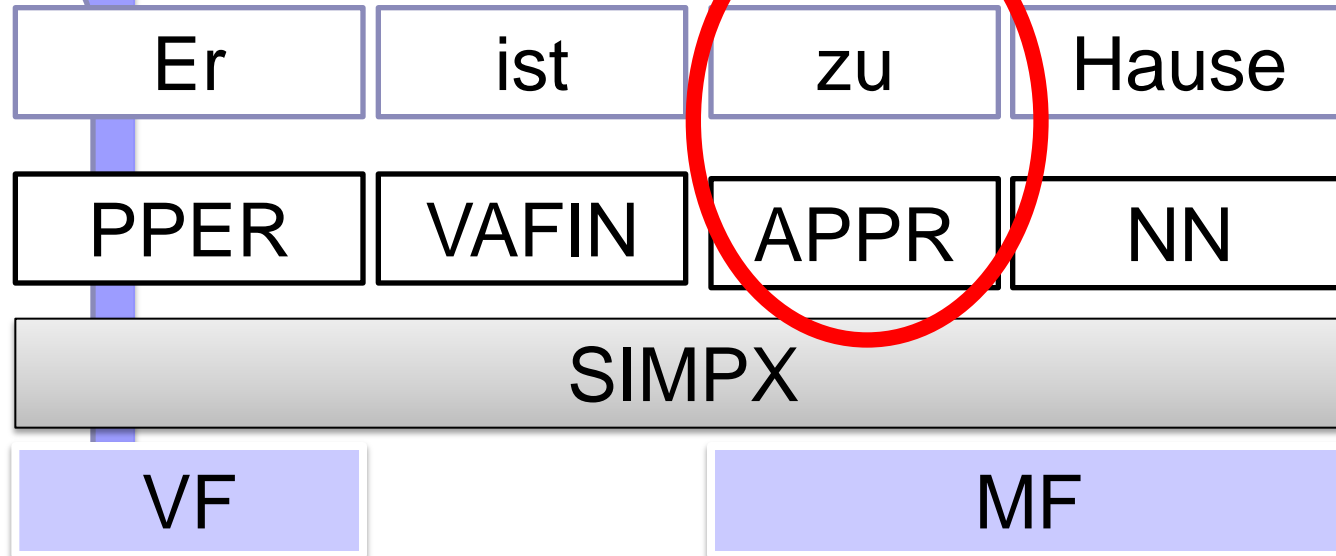
word = "zu"
&
pos = "APPR"
&
#1_=#2



Ausdruck 2: #2

Suche auf mehreren Ebenen

word = "zu"
&
pos = "APPR"
&
#1 = #2



Suche nach einem Token mit dem Wert "zu", das eine Wortartenannotation "APPR" trägt. Beide Annotationen sollen sich auf den gleichen Bereich beziehen.

Suche auf mehreren Ebenen

word = "zu"
&
pos = "APPR"
&
TopoField = "MF"
&
#1_ = _#2
&
#3_ i _#1

| | | | |
|----|-----|-------|-------|
| Er | ist | #1 zu | Hause |
|----|-----|-------|-------|

| | | | |
|------|-------|------|----|
| PPER | VAFIN | APPR | NN |
|------|-------|------|----|

| | | | |
|-------|--|--|--|
| SIMPX | | | |
|-------|--|--|--|

| |
|----|
| VF |
|----|

| |
|-------|
| #3 MF |
|-------|

Das gesuchte Token soll außerdem im Mittelfeld (MF) stehen.

Aufgabe

- Finden Sie alle Verben in der linken Satzklammer.

Lösung:

```
ZH1pos = /V.FIN/ &  
ZH1TopoFields = "LK" &  
#2_i_#1
```


Aufgabe

- Wie finde ich einen Satz, indem ein "weil" vorkommt?

Lösung:

```
ZH1S &  
lemma="weil" &  
#1_i_#2
```

Aufgabe

- Wie finde ich einen Satz, indem ein "weil" vorkommt?
- mit einer Bewegung auf ZH1

Lösung:

```
ZH1S &  
lemma="weil" &  
#1_i_#2 &  
ZH1Diff = /MOV./ &  
#1_i_#3
```

Metadaten finden

ANNIS? Tutorial

Search Form

AnnisQL: pos="NN"

Query Builder: Show >>

Result: 21842

More Corpora

| <input type="checkbox"/> | Name ▲ | Texts | Tokens | |
|-------------------------------------|-----------------------|-------|--------|-------------------|
| <input type="checkbox"/> | FalkoEssayL1V2_0 | 95 | 70608 | i |
| <input checked="" type="checkbox"/> | FalkoEssayL2V2_0 | 248 | 131599 | i |
| <input type="checkbox"/> | FalkoSummaryL1V1_2 | 57 | 21211 | i |
| <input type="checkbox"/> | FalkoSummaryL2V1_2 | 107 | 40865 | i |
| <input type="checkbox"/> | FalkoSummaryVLV1_0 | 12 | 11114 | i |
| <input type="checkbox"/> | HUKonstruktionsgramma | 22 | 19340 | i |

Search Result - pos="NN" (5, 5)

Page 1 of 2185 Token Annotations Show Citation URL

[i](#) Der **Feminismus** hat den Interessen der Frauen
Feminismus haben d Interesse d [unknown]
NN VAFIN ART NN ART NN

[falko \(grid\)](#)
[ZH1 \(grid\)](#)
[ZHverb \(grid\)](#)
[ZH2 \(grid\)](#)
[text \(grid\)](#)
[Volltext](#)

[i](#) Der Feminismus hat
d Feminismus haben d Interesse d [unknown] mehr schaden als
ART NN VAFIN ART NN ART NN ADV VVPP KOKOM

[falko \(grid\)](#)
[ZH1 \(grid\)](#)
[ZHverb \(grid\)](#)
[ZH2 \(grid\)](#)
[text \(grid\)](#)
[Volltext](#)

[i](#) Feminismus hat den Interessen der **Frauen** mehr geschadet als genützt Im
Feminismus haben d Interesse d [unknown] mehr schaden als nützen im
NN VAFIN ART NN ART NN ADV VVPP KOKOM VVPP APPRART

[falko \(grid\)](#)
[ZH1 \(grid\)](#)
[ZHverb \(grid\)](#)
[ZH2 \(grid\)](#)
[text \(grid\)](#)
[Volltext](#)

[i](#) mehr geschadet als genützt Im **Bereich** der Literatur ist viel über
mehr schaden als nützen im Bereich d Literatur sein viel über
ADV VVPP KOKOM VVPP APPRART NN ART NN VAFIN ADV APPR

[falko \(grid\)](#)
[ZH1 \(grid\)](#)
[ZHverb \(grid\)](#)

text- & lernerbezogene
Metadaten anzeigen

Metadaten finden

- Metadaten:
 - Variablen und Werte für
 - Text
 - Lerner

Meta Data for id 916

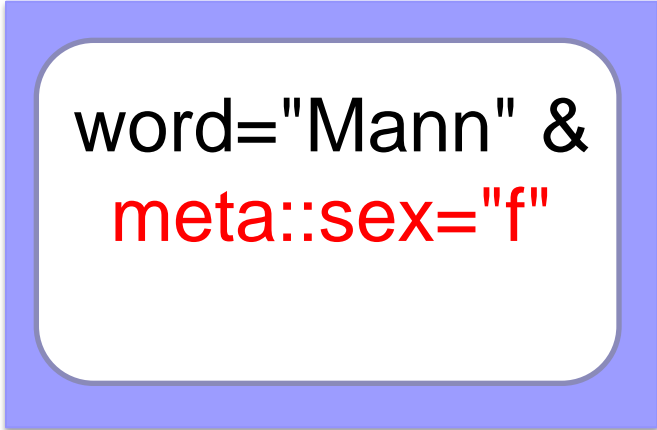
| | |
|----------------------|---|
| SPK0:l2_1_duration | 96 |
| SPK0:l2_1_langschool | N/A |
| SPK0:l2_1_school | N/A |
| SPK0:L2index | deu:96:N/A:N/A:N/A:N/A:N/A |
| SPK0:name | 9d3abd29757a213f0244c7126c50b440e7b302a1 |
| SPK0:reg | l1:eng,l1:N/A,l1:N/A,l1:N/A,l1:N/A,l1:N/A,l1:N/A,l1:N/A |
| SPK0:sex | f |
| major-subject | deutschintensiv |
| production-modality | essay |
| subcorpus | fk_2006_07 |
| topic | Feminismus |
| transcription-date | 27.07.06 |
| transcriptionName | fk007_2006_07 |
| transcripator | MMM/V |
| projectName | FALKO Essay Corpus L2 2.0 |
| projectURL | Falko project site |

Nach Metadaten filtern

- Nach Metadaten sucht man mit `meta::Variable = "Wert"`
in `FalkoEssayL2v2.3`

**Finden Sie alle Wortformen von "Mann",
die von weiblichen Lernern geschrieben
wurden.**

**(Die Variable für
Geschlecht ist
`sex="f"`).**



```
word="Mann" &  
meta::sex="f"
```

Nach Metadaten filtern

- Nach der Muttersprachen von Lernern sucht man mit

`meta::reg=/l1:LÄNDERCODE/`

- Finden Sie alle Formen des Adjektivs *deutsch* in den Texten englischer Muttersprachler (Code= `eng`)

Sprachkürzel in Falko (Auswahl)

| | | | |
|-----|---------------|-----|----------------|
| afr | afrikaans | | niederländisch |
| dan | dänisch | | |
| deu | deutsch | nor | norwegisch |
| ell | neugriechisch | pol | polnisch |
| eng | englisch | rus | russisch |
| fin | finnisch | spa | spanisch |
| fra | französisch | swe | schwedisch |
| heb | hebräisch | tur | türkisch |
| hun | ungarisch | ukr | ukrainisch |
| isl | isländisch | uzb | usbekisch |
| ita | italienisch | xho | xhosa |
| jpn | japanisch | yid | jiddisch |
| lat | lateinisch | zho | zulu |

Nach Metadaten filtern

- Nach der Muttersprachen von Lernern sucht man mit

`meta::reg=/l1:LÄNDERCODE/`

- Finden Sie alle Formen des Adjektivs *deutsch* in den Texten englischer Muttersprachler (Code= `eng`)

```
lemma="deutsch"&  
meta::reg=/l1:eng/
```


Nach Metadaten filtern

- Soll die Sprachbiographie genauer beschrieben werden, muss zwischen beiden Informationen ein **.** stehen.
- **meta::reg=/Variable1:Wert1.*Variable2:Wert2/**
- **Finden Sie alle Formen von "deutsch" in den Texten dänischer Muttersprachler mit L2 Englisch.**

```
lemma="deutsch"&  
meta::reg=/l1:dan.*l2:eng/
```

Token zählen

- Gibt man nur eine Variable ohne Wert an, bekommt man die Anzahl der gefüllten Zellen

word



Anzahl der Tokens

lemma



Anzahl der Lemmata

- So kann man auch die Anzahl der Token für eine bestimmte Lernergruppe ermitteln.

word &
meta::reg=/l1:ita/

Token zählen

- Wie viele Token von japanischen Lernern sind in Falko enthalten?

word &
meta::reg=/l1:jpn/

Ein erster Vergleich

- Vergleichen Sie die Häufigkeit der "ung"-Derivationen bei Dänen und Italienern
- Wichtig: Wieviele Tokens gibt es pro Sprachgruppe
- Wieviele Unga pro Token gibt es also?

Gesamtanzahl der Token für die Muttersprachen

word&
meta::reg=/l1:dan/

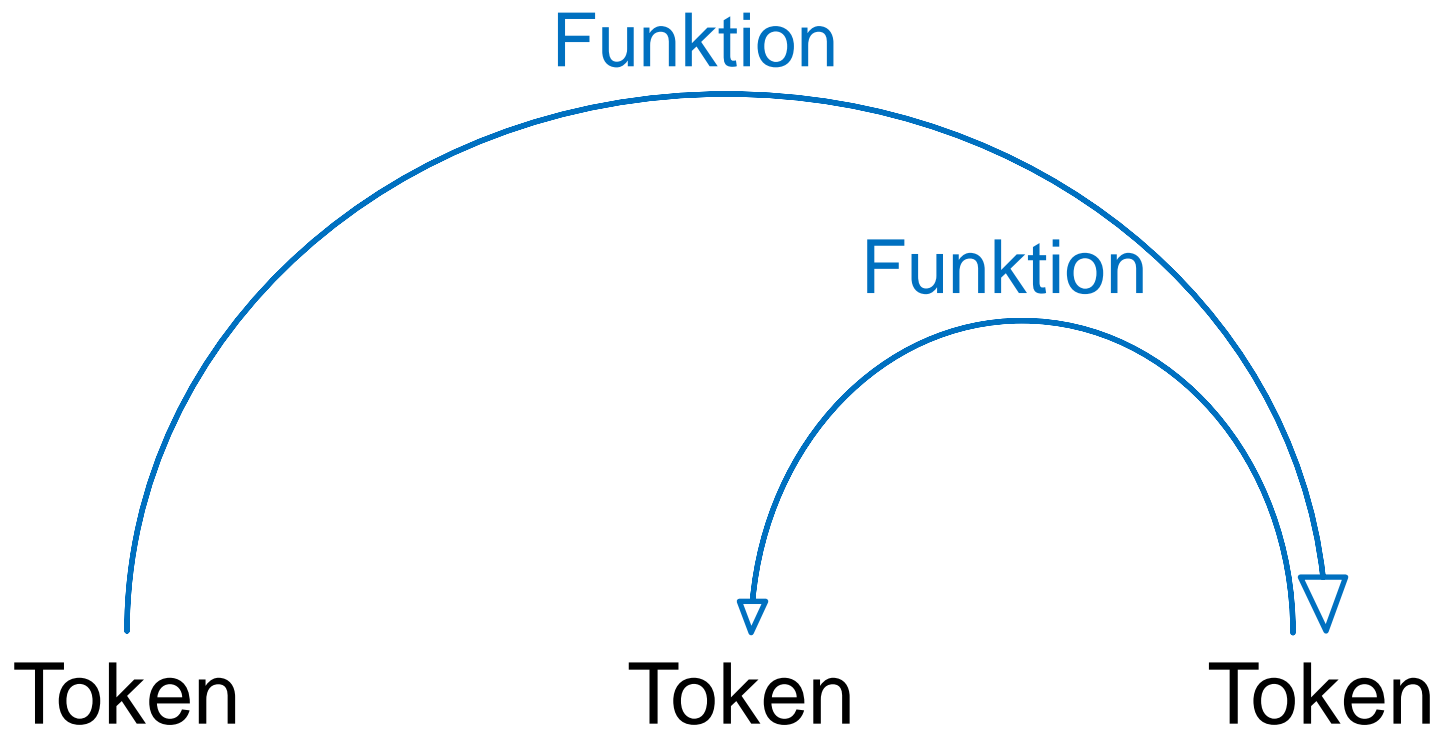
word&
meta::reg=/l1:ita/

Ambige Zeichen: Operatoren und Wort-/Satzzeichen

- Erinnerung: Der Operator " **■** " bedeutet "**ein beliebiges Zeichen**".
- Wie finde ich dann "**usw.**"?
- Der Operator "****" bedeutet "**das folgende Zeichen ist wörtlich gemeint**".

word=/usw\./

Suchen nach Abhängigkeiten



Dependenzschema

POS="NN" &
POS = "NN" &
#2 ->dep #1



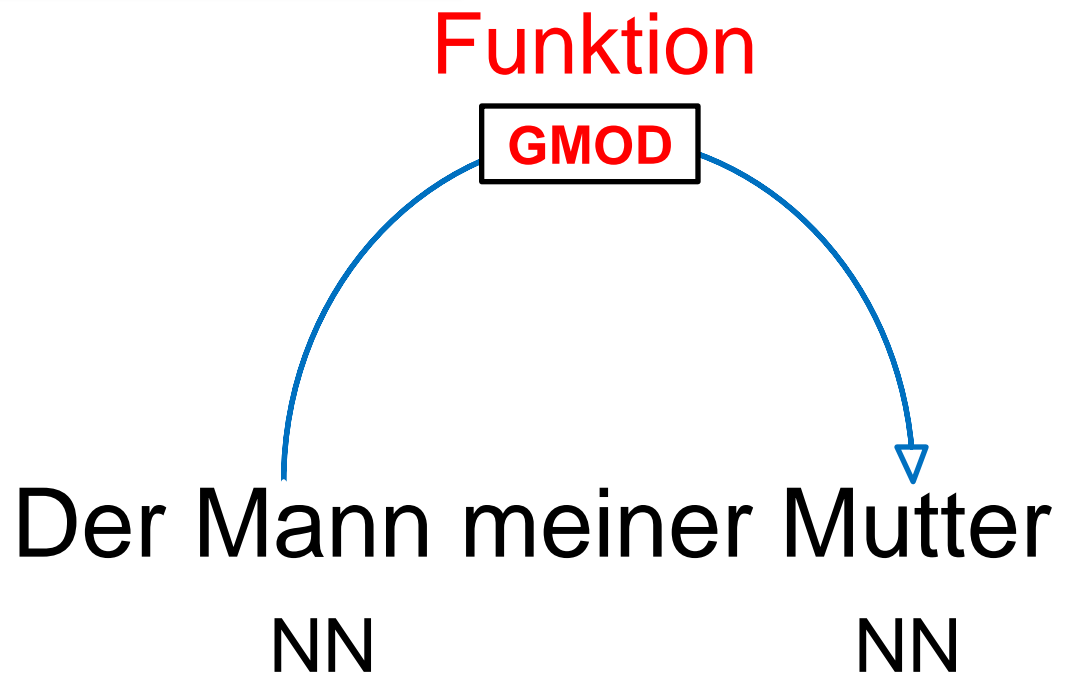
Suche nach Satzfunktionen

Einige Abhängenzrelationen

| | |
|------|--------------------|
| SUBJ | Subjekt |
| OBJA | Akkusativobjekt |
| OBJD | Dativobjekt |
| OBJG | Objekt im Genitiv |
| OBJC | Objektsatz |
| ATTR | Attribut |
| DET | Determinierer |
| GMOD | Genitivmodifikator |

Dependenzschema

```
pos="NN" &  
pos = "NN" &  
#2 ->dep[func="GMOD"] #1
```



Dependenzschema

ACHTUNG: Bisher hängen die Abhängigkeiten von der ZH1 ab! Nicht von word !

```
ZH1pos="NN" & ZH1 & #1_=_#2 &  
ZH1pos = "NN" & ZH1 & #3_=_#4 &  
#4 ->dep[func="GMOD"] #2
```

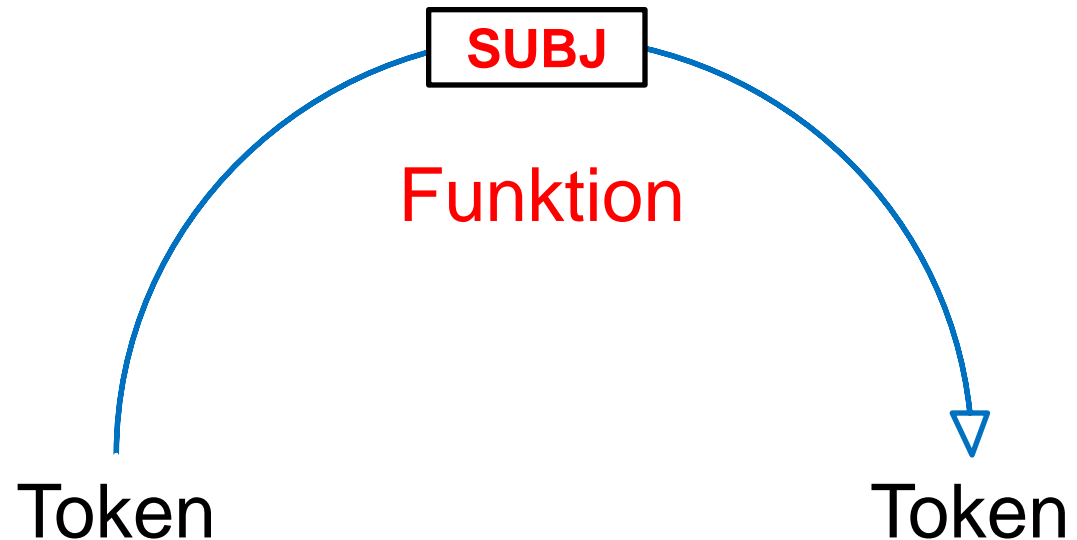
GMOD

ZH1: Der Mann meiner Mutter
tok: Der man meine Mutter

Dependenzschema

ZH1 &
ZH1 &
#2->dep[function="SUBJ">#1

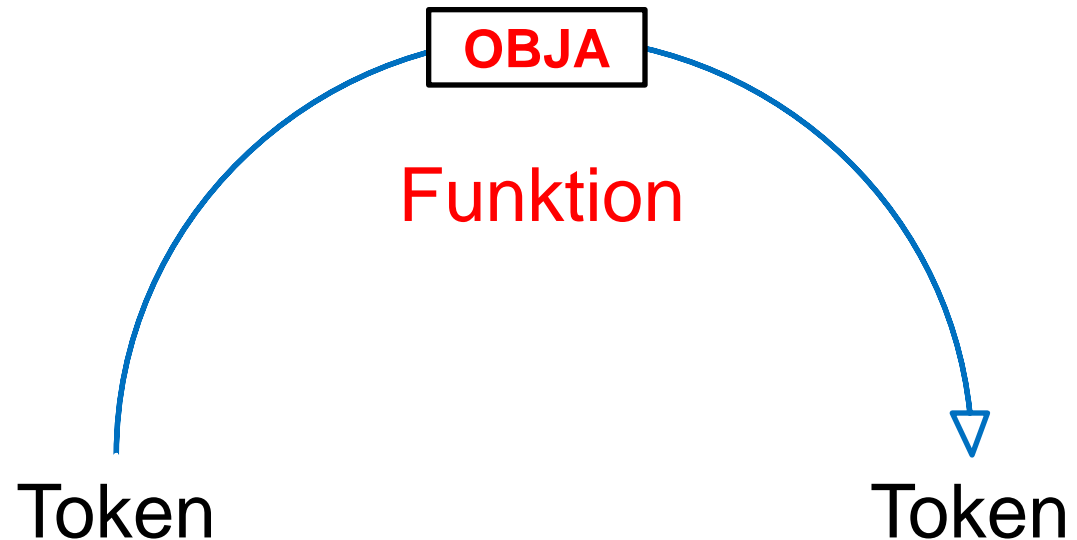
**Finden Sie alle
Subjekte**



Dependenzschema

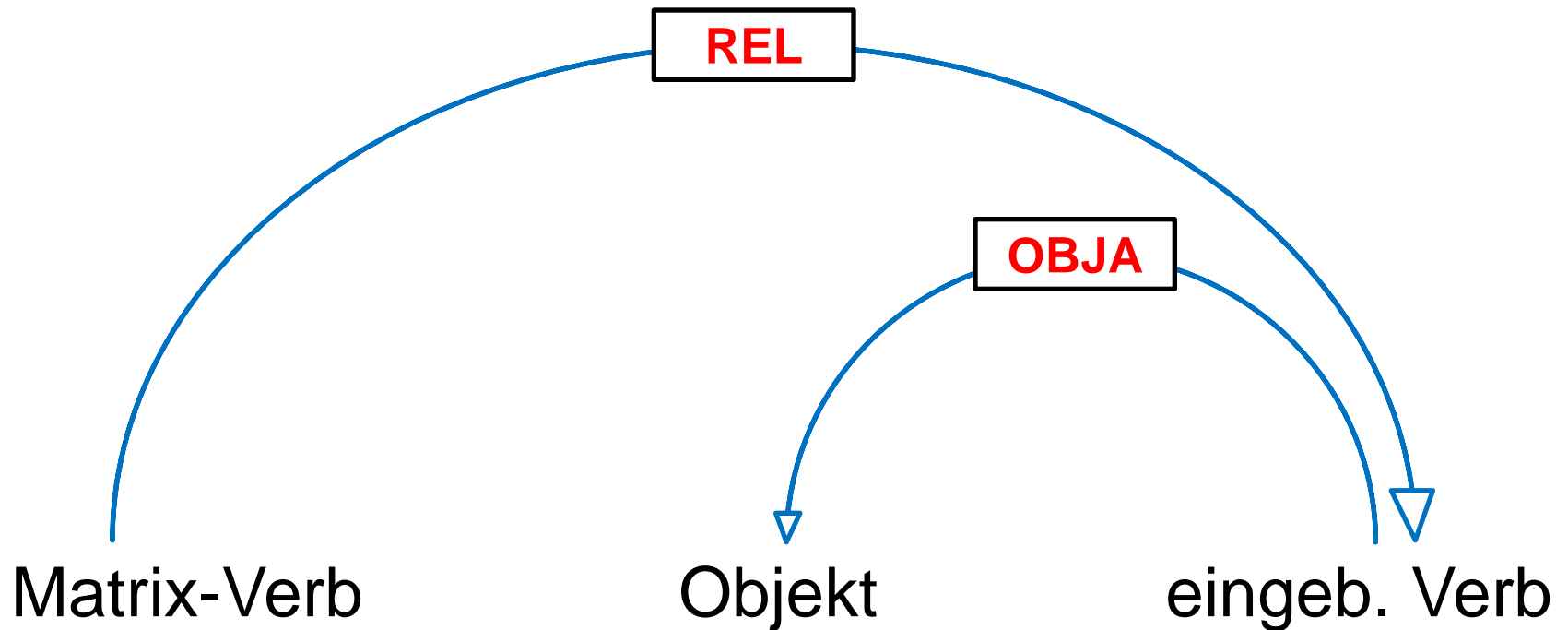
ZH1 &
ZH1 &
#2->dep[function="OBJA"]#1

**Finden Sie alle
Verben mit einem
Akkusativobjekt.**



Abhängenzschema

Finden Sie die gleichen Verben in Relativsätzen (REL).



Aufgabe

- Finden Sie alle Subjekte im Vorfeld
in Ihrem Beispieltext

Lösung:

```
ZH1 &  
ZH1 &  
TopoField = "VF" &  
#1_i_#3 &  
#2 ->dep[func="SUBJ" ]#1
```

Zusammenfassung - Operatoren

| | |
|--------|--|
| . | 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 zu Tokenrelationen:

#1.#2

#1 wird direkt gefolgt von #2.

#1.*#2

#1 wird indirekt gefolgt von #2.

#1_=#2

#1 und #2 beziehen sich auf die gleichen Token.

#1_i_#2

#1 ist in #2 enthalten.

#1 ->dep[func="SUBJ"]#2 #2 ist Subjekt von #1.

Zusammenfassung

- Mit ANNIS kann man:
 - in unterschiedlichen Korpora (auch gleichzeitig) suchen
 - die Ergebnisse quantifizieren
 - die Ergebnisse exportieren
 - sehr heterogene Annotationen gleichzeitig durchsuchen
- Man kann auch nach Metadaten filtern (Tutorial)
- CIA und EA möglich, aber (noch) keine expliziten Fehlertags (Nachteil?)

Vorsicht

- Ein Korpus entspricht nicht der ganzen Sprache
- Unterschiedliche Korpora zeigen unterschiedliche Ergebnisse (das ist auch interessant!)
- Manchmal sind Korpora fehlerhaft
- Trotzdem können Korpora Hypothesen gut unterstützen oder widerlegen

Diskussion...

- Kasusfehler
 - in verschiedenen Wortarten
 - in Kombination mit Präpositionen
- Verbstellungsfehler
- Vorkommen (Distribution) von Passivkonstruktionen
- Transferfehler
- ... und alles Weitere, woran wir interessiert sind

Herzlichen Dank!

Literatur

- **Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik (2008):** Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache* 45 (2), S. 67–73.
- **Reznicek, Marc; Walter, Maik; Schmidt, Karin; Lüdeling, Anke; Hirschmann, Hagen; Krummes, Cedric; Andreas, Thorsten (2010):** Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 1.0. Berlin: Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin. URL: <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko> [Stand: 12. Oktober 2010].
- **Zeldes, Amir; Ritz, Julia; Lüdeling, Anke; Chiarcos, Christian (2009):** ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009*.

Texte suchen

- Jeder Text beginnt mit der Annotation `TXTstructure = "start "` und endet mit `"end"`.
- Suchen Sie alle Texte von italienischen Muttersprachlern.

```
TXTstructure="start" &  
meta::reg=/l1:ita/
```

Texte suchen

- Wie viele Texte wurden von französischen Muttersprachlern geschrieben?

Sprachkürzel in Falko (Auswahl)

| | | | |
|-----|---------------|-----|----------------|
| afr | afrikaans | | niederländisch |
| dan | dänisch | | |
| deu | deutsch | nor | norwegisch |
| ell | neugriechisch | pol | polnisch |
| eng | englisch | rus | russisch |
| fin | finnisch | spa | spanisch |
| fra | französisch | swe | schwedisch |
| heb | hebräisch | tur | türkisch |
| hun | ungarisch | ukr | ukrainisch |
| isl | isländisch | uzb | usbekisch |
| ita | italienisch | xho | xhosa |
| jpn | japanisch | yid | jiddisch |
| lat | lateinisch | zho | zulu |

Texte suchen

- Wie viele Texte wurden von französischen Muttersprachlern geschrieben?

```
TXTstructure="start" &  
meta::reg=/l1:fra/
```