



# Korpuslinguistik das **ANNIS**-Korpussuchtool zum Durchsuchen des Falko-Korpus

Anke Lüdeling, Thomas Krause, Marc Reznicek, Amir Zeldes,  
Hagen Hirschmann  
hirschhx@hu-berlin.de

... und weitere Mitarbeiter der Korpuslinguistik an der  
Humboldt-Universität zu Berlin

# Das Suchinterface ANNIS



- **Im Korpus-Suchinterface ANNIS wird Falko (mit allen Annotationen) durchsuchbar gemacht**
  - **ANNIS-Portal** → alle Korpora, die an der HU in ANNIS verfügbar sind:  
<https://hu.berlin/annis>
  - **ANNIS-Portal speziell für Lernerkorpora**  
→ Falko, WHiG, BeMaTaC L2, Kobalt-DaF, KanDel:  
<https://hu.berlin/annis-falko>

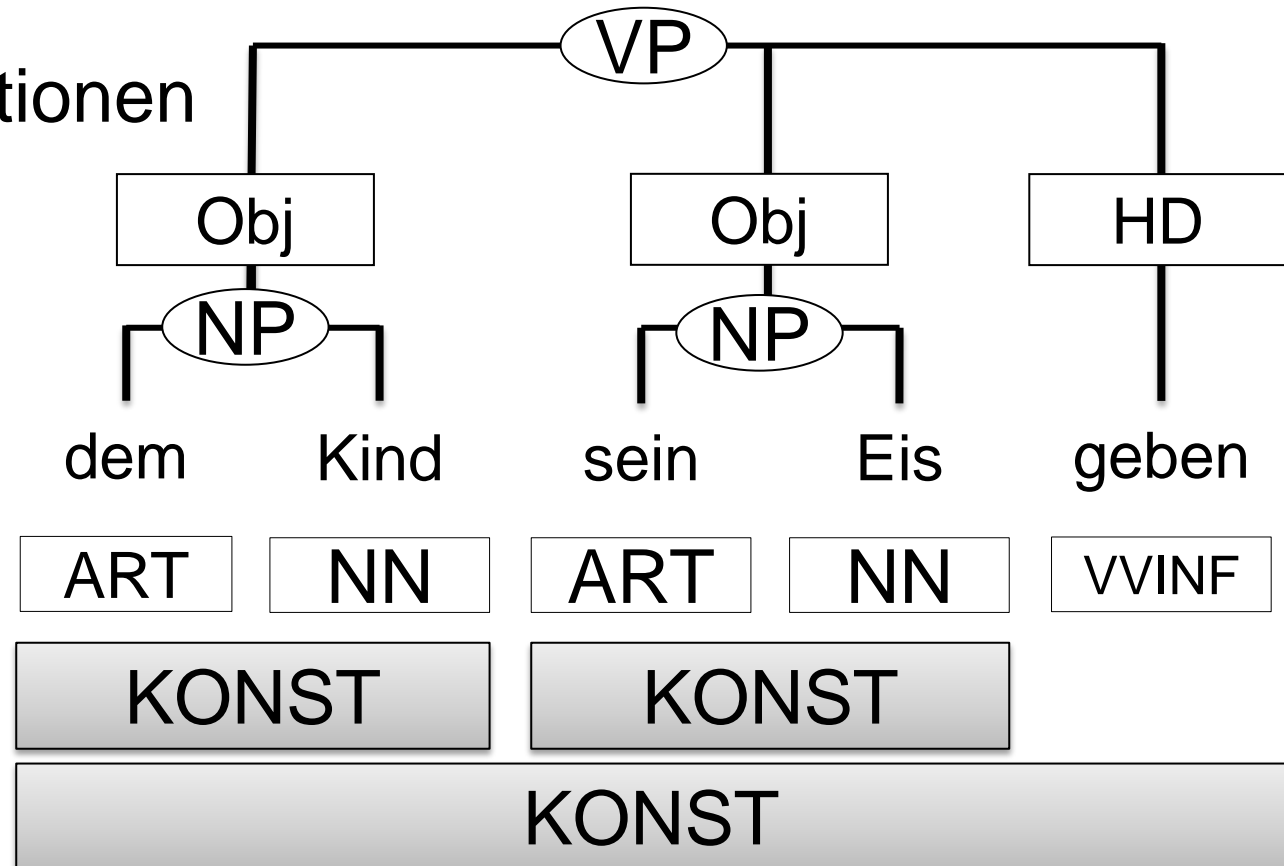
# Falko durchsuchen...

- Wie/Was kann man in ANNIS suchen?
  - Wortformen
  - Token-Annotationen:
    - Lemmata
    - Wortarten
  - dadurch: Linguistische Muster
  - Satzspannen
  - strukturelle Fehler, basierend auf Zielhypothesen
- Zusätzlich Filterung nach diversen Metadaten

# Beispiel für Annotationen in ANNIS



- Token
- Token-Annotationen
- Spannen
- Bäume



# Das Web-Interface: Tutorial



Browser address bar: <https://korpling.german.hu-berlin.de/falko-suche/>

Navigation: [About ANNIS](#) | [Report Problem](#) | [Help us make ANNIS better!](#)

Search Interface:

Please enter AQL query

Query Builder

Search [More] History

Welcome to ANNIS! A tutorial is available on the right side.

Corpus List | Search Options

Visible: Falko

Name	Texts	Tokens		
CLEG13	729	285.286	📘	📄
FalkoEssayL1v2.0	94	70.110	📘	📄
falkoEssayL1v2.3	95	70.615	📘	📄
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	📘	📄
FalkoSummaryL1v1.2	57	21.211	📘	📄
FalkoSummaryL2v1.2	106	40.638	📘	📄
FalkoWHIGL2v2.1	196	130.949	📘	📄

Help/Examples

- [Help](#)
- [Example Queries](#)

Example Query	Description
<a href="#">Q "haben"</a>	search for the word "haben"
<a href="#">Q /fAa llen/</a>	Search for the "allen" with upper or lower-case "
<a href="#">Q "der"</a>	search for the word "der"
<a href="#">Q /fEe linge/</a>	Search for the "einge" with upper or lower-case "
<a href="#">Q "sich"</a>	search for the word "sich"
<a href="#">Q /fAa uszahlt/</a>	Search for the "auszahlt" with upper or lower-ca
<a href="#">Q "Studenten"</a>	search for the word "Studenten"
<a href="#">Q /fDd ie/</a>	
<a href="#">Q "bereiten"</a>	
<a href="#">Q /fDd ie/</a>	Search for the "die" with upper or lower-case "d"
<a href="#">Q "die"</a>	search for the word "die"
<a href="#">Q /fDd ie/</a>	Search for the "die" with upper or lower-case "d"
<a href="#">Q "Universitätsabschlüsse"</a>	search for the word "Universitätsabschlüsse"
<a href="#">Q /fDd ie/</a>	Search for the "die" with upper or lower-case "d"
<a href="#">Q "bin"</a>	search for the word "bin"
<a href="#">Q /fBb in/</a>	Search for the "bin" with upper or lower-case "b"
<a href="#">Q "Meibauer"</a>	search for the word "Meibauer"
<a href="#">Q /f11 71/</a>	Search for the "171" with upper or lower-case "1"
<a href="#">Q "I"</a>	search for the word "I"
<a href="#">Q "Studenten"</a>	search for the word "Studenten"

**1) Hilfe/Tutorial öffnen**

# Das Web-Interface: Tutorial



Navigation: About ANNIS | Report Problem | [Help us make ANNIS better!](#)

Search: Please enter AQL query | Query Builder | Search | More | History

Welcome to ANNIS! A tutorial is available on the right side.

Corpus List | Search Options

Visible: Falko

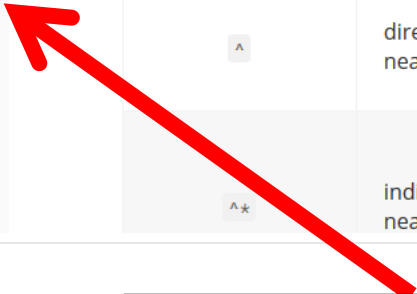
Name	Texts	Tokens		
CLEG13	729	285.286	🔍	📄
FalkoEssayL1v2.0	94	70.110	🔍	📄
falkoEssayL1v2.3	95	70.615	🔍	📄
FalkoEssayL2v2.0	248	132.066	🔍	📄
FalkoEssayL2v2.3	248	131.628	🔍	📄
falkoEssayL2v2.4	248	144.619	🔍	📄
FalkoEssayL2WHIGv2.0	195	130.187	🔍	📄
FalkoGeorgetownL2v1.0	92	78.151	🔍	📄
FalkoSummaryL1v1.2	57	21.211	🔍	📄
FalkoSummaryL2v1.2	106	40.638	🔍	📄
FalkoWHIGL2v2.1	196	130.949	🔍	📄
KanDeL_cross_cohort_v2	425	73.920	🔍	📄

Help/Examples | Help

## Complete List of Operators

AQL currently includes the following operators:

Operator	Description	Illustration	Notes
.	direct precedence	AB	For non-terminating multiple segments which consist of specified words
.*	indirect precedence	A x y z B	For specific .n,m can be 3 and 4 tokens maximum. As above, specified, e.g.
^	directly near	AB or BA	Same as precedence. In corpora where the layer or may be specified
^*	indirectly near	A x y z B or B x y z A	Like indirect precedence. The form ^ between 3 and 4 tokens default maximum



2) Thema wählen

# Das Web-Interface: Abfrage

<https://hu.berlin/annis-falko>



Suchfenster (Sucheingabe hier)

Abfrage starten

Statistik & Export

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

Name	Texts	Tokens		
FalkoEssayL2v2.3	248	131.628	i	
<b>falkoEssayL2v2.4</b>	<b>248</b>	<b>144.619</b>	<b>i</b>	<b></b>
FalkoEssayL2WHIGv2.0	195	130.187	i	
FalkoGeorgetownL2v1.0	92	78.151	i	
FalkoSummaryL1v1.2	57	21.211	i	
FalkoSummaryL2v1.2	106	40.638	i	
FalkoWHIGL2v2.1	196	130.949	i	
KanDeL_cross_cohort_v2	425	73.920	i	

# Das Web-Interface: Such-Einstellungen



The screenshot shows a search interface with the following elements:

- Search bar containing the text **"lernen"**.
- Buttons: **Search**, **More** (dropdown), and **History** (dropdown).
- Results summary: **147 matches in 64 documents**.
- Navigation tabs: **Corpus List** and **Search Options**.
- Search Options section:
  - Left Context**: dropdown menu showing **5**.
  - Right Context**: dropdown menu showing **5**.
  - Show context in**: dropdown menu showing **tokens (defa ...**.
  - Results Per Page**: dropdown menu showing **10**.
  - Order**: dropdown menu showing **Ascending**.

Red arrows point from external labels to these settings:

- Suche starten** points to the search bar.
- linker Kontext** points to the Left Context dropdown.
- rechter Kontext** points to the Right Context dropdown.
- Resultate pro Seite** points to the Results Per Page dropdown.



# Prinzip I: Variablen-Wert-Paare

■ **word** = "das"  
↑ **Variable1**  
("word" = Wortform)

↑ **Wert**  
↓

<b>word</b>	Sofern	<b>das</b>	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen

"word" = alle Wörter und Satzzeichen im Original-Lernertext

# Prinzip I: Variablen-Wert-Paare



■ word = "das"

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

# Prinzip I: Variablen-Wert-Paare

■ pos = "ART"

Variable2

("pos" = Wortart)

Wert

<b>word</b>	Sofern	das	System	herrscht
pos	KOUS	<b>ART</b>	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<sub>3</sub>  
("Lemma")

Wert

<b>word</b>	Sofern	das	System	herrscht
pos	KOUS	<b>ART</b>	NN	VVFIN
lemma	sofern	d	System	herrschen

# Prinzip I: Variablen-Wert-Paare



■ lemma = "d"

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

# beliebig erweiterbar...

■ **ZH1S** = "S"

**Variable<sub>4</sub>**  
("Satz auf ZH1-Ebene")

**Wert**

<b>word</b>	Sofern	das	System	herrscht
pos	KOUS	ART	NN	VVFIN
lemma	sofern	d	System	herrschen
ZH1S	<b>S</b>			

# beliebig erweiterbar...



■ ZH1S = "S"

...findet alle Sätze wie  
*Sofern das System herrscht*

auf der Suche Ebene "ZH1S"



# Suche nach Strings

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

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)

- Annis<sub>2</sub> 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<sup>↷</sup>s?

das vorherige Zeichen ist optional

→  $\phi$ , s → *da, das*

das<sup>↷</sup>s\*

das vorh. Zeichen kommt 0- bis  $\infty$  mal vor

→  $\phi$ , s, ss, ... → *da, das, dass, dassssssssss*

das<sup>↷</sup>s+

das vorh. Zeichen kommt 1- bis  $\infty$  mal vor

→ s, ss, ... → *das, dass, dassssssssssssss*

# Aufgabe

- Was passiert, wenn Sie die Operatoren kombinieren?

word = /Frau.??/

word = /Frau.\*//

word = /Frau.+//



# 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 ( ) und |

- Frau(en)?     findet *Frau* und *Frauen*
- (e|E)r     findet *er* und *Er*  
→ wichtig bei Satzanfängen
- lemma=/(Frau|Mann|Kind)/  
findet diese drei Lemmata  
mit einer Suche

# Aufgabe

- Finden Sie mit der Lemmasuche Wörter, deren Grundform auf *-ling* oder *-lang* endet

→ lemma = /.\*(i|a)ng/

# 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

...

# Stuttgart-Tübingen-Tagset (STTS)

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

# Stuttgart-Tübingen-Tagset (STTS)

<b>VERB</b>	<b>Vollverb</b>	Auxiliar	Modalverb
finit	<b>VV</b> FIN	VAFIN	VMFIN
Imperativ	<b>VV</b> IMP	VAIMP	
infinit	<b>VV</b> INF	VAINF	VMINF
Infinitiv mit <i>zu</i>	<b>VV</b> IZU		
Partizip 2	<b>VV</b> PP	VAPP	VMPP



# Aufgabe

- Suchen Sie nach Possessivpronomen

pos =/PPOS(S|AT)/

# Prinzip II: Relationen

- Einzelne **Variable-Wert-Paare** nacheinander geschrieben
- Zwischen den Paaren muss **IMMER eine Beziehung hergestellt** werden
- Die Beziehung wird mittels entsprechender Operatoren ausgedrückt (nächste Folie)

Variable<sub>1</sub> = Wert<sub>1</sub>

"Beziehung"

Variable<sub>2</sub> = Wert<sub>2</sub>



Ausdruck 1



Beziehungsoperator



Ausdruck 2

# Liste möglicher Beziehungen ...



⌵
About ANNIS
Report Problem
[Help us make ANNIS better!](#)

Please enter AQL query

Q Search
More ▾
History ▾

Welcome to ANNIS! A tutorial is available on the right side.

Corpus List
Search Options

Visible: Falko

Name	Texts	Tokens		
CLEG13	729	285.286	ⓘ	📄
FalkoEssayL1v2.0	94	70.110	ⓘ	📄
falkoEssayL1v2.3	95	70.615	ⓘ	📄
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	ⓘ	📄
FalkoSummaryL1v1.2	57	21.211	ⓘ	📄
FalkoSummaryL2v1.2	106	40.638	ⓘ	📄
FalkoWHIGL2v2.1	196	130.949	ⓘ	📄
KanDeL_cross_cohort_v2	425	73.920	ⓘ	📄

Help/Examples

Help

☰
✎
🔍
ANNIS Help

## Complete List of Operators

AQL currently includes the following operators:

Operator	Description	Illustration	Notes
.	direct precedence	<b>AB</b>	For non-terminated most terms multiple segments which consist of specified words
.*	indirect precedence	<b>A xyz B</b>	For specific words .n,m can be used between 3 and 4 tokens maximum can be used. As above, specified, e.g.
^	directly near	<b>AB or BA</b>	Same as precedence. In corpora the layer or may be specified
^.*	indirectly near	<b>A xyz B or B xyz A</b>	Like indirect precedence. The form ^ between 3 and 4 tokens default maximum

# Prinzip II: Relationstypen

pos = /PPOS(S|AT)/



Ausdruck 1

\_=\_



Beziehungsoperator: "Abdeckung"

lemma = "meinen"



Ausdruck 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 function 'object, accusative'
		A 	For specific distance of dominance, >n,r

# Aufgabe

- Finden Sie möglichst viele Vorkommen von Partikelverben mit der Lemmasuche und der Wortartensuche verknüpft

→ lemma = /(ver|be|ent|zer|).\*/  
& pos=/N.\*/

# Aufgabe

- Finden Sie nun Vorkommen von **word = "meinen"** , die ausschließlich finite Vollverben sind

**word = "meinen"      =**  
**pos= "VVFIN"**

# 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" &
```

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



word = "zu"

pos = "NN"

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./

# 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	,
word	weil	sie	entdeckt	eine	Aspekte	der	Gesellschaft		,

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

# Edit Tags

<b>ZHDiff</b>	<b>Operation in Zielhypothese</b>
<b>INS</b>	<b>Token eingefügt</b>
<b>DEL</b>	<b>Token gelöscht</b>
<b>CHA</b>	<b>Token geändert</b>
<b>MERGE</b>	<b>mehrere Token verbunden</b>
<b>SPLIT</b>	<b>Token in mehrere aufgespalten</b>
<b>MOVS</b>	<b>Token von hier bewegt</b>
<b>MOVT</b>	<b>Token hierhin bewegt</b>

# Aufgabe

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

ZH1pos="PRF"

  <sup>=</sup>  

ZH1Diff="INS"

# Aufgabe

- ...alle indefiniten Artikel, die in den Lernertexten fehlen

**Lösung:**

ZH1lemma="d"

    
=

ZH1Diff="INS"

# Metadaten (Informationen über den jeweiligen Text) finden



Please enter AQL query

Query Builder

Help/Examples falkoEssayL2v2.4 x

Filter documents by name

Corpus information for falkoEssayL2v2.4 (ID: 7121)

Metadata		Available annotations	
Name	Value	Name	Example (click to use query)
projectName	Falko Essay L2 v2.4	ctok	ctok=","
		ctoklemma	ctoklemma="d"
		ctokpos	ctokpos="NN"
		fm	fm="fm:eng"
		lemma	lemma="d"
		lemma.	lemma.="d"
		macro	macro="title"
		pos	pos="NN"
		pos.	pos.="NN"
		topo1	topo1="MF"
		topo2	topo2="VCE"
		TXTstructure	TXTstructure="PAR"
		verbfehlertyp	verbfehlertyp="flex"

Alle Annotationen und Metadaten anzeigen

text- & lernerbezogene Metadaten anzeigen

Visible: Falko

Name	Texts	Tokens		
CLEG13	729	285.286	i	
FalkoEssayL1v2.0	94	70.110	i	
falkoEssayL1v2.3	95	70.615	i	
FalkoEssayL2v2.0	248	132.066	i	
FalkoEssayL2v2.3	248	131.628	i	
<b>falkoEssayL2v2.4</b>	<b>248</b>	<b>144.619</b>	<b>i</b>	<b></b>
FalkoEssayL2WHIGv2.0	195	130.187	i	
FalkoGeorgetownL2v1.0	92	78.151	i	
FalkoSummaryL1v1.2	57	21.211	i	

Link to corpus: [https://korpling.german.hu-berlin.de/falko-suche/#\\_c=ZmFsa29Fc3NheUwydjluNA](https://korpling.german.hu-berlin.de/falko-suche/#_c=ZmFsa29Fc3NheUwydjluNA)

45

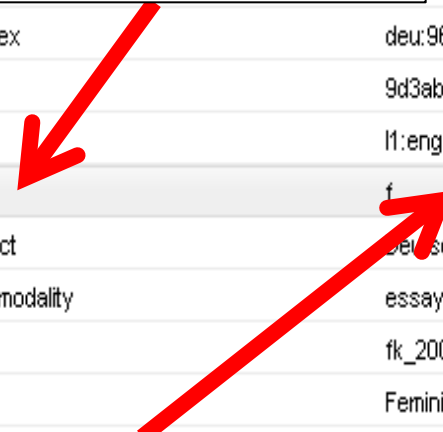
# Metadaten finden

- Metadaten:
  - Variablen und Werte für
    - Text
    - Lerner
- Klicken Sie Beispiele an, um sie ins Suchfenster zu überführen

Meta Data for id 916

SPK0:I2_1_duration	96
SPK0:I2_1_langschool	N/A
SPK0:I2_1_school	N/A
SPK0: <b>Variable</b>	
SPK0:L2index	deu:96:N/A:N/A:N/A:N/A:N/A
SPK0:name	9d3abd29757a2f3f0244c7f26c50b440e7b302a1
SPK0:reg	1:eng,1:N/A,1:N/A,1:N/A,1:N/A,1:N/A,1:N/A,1:N/A
SPK0:sex	f
major-subject	ber_schintensiv
production-modality	essay
subcorpus	fk_2006_07
topic	Feminismus
transcription-d	27.07.06
transcriptionN	fk007_2006_07
transcriptor	MMMV
projectName	FALKO Essay Corpus L2 2.0
projectURL	<a href="#">Falko project site</a>

**Wert**



# Nach Metadaten filtern

- Nach Metadaten sucht man mit `meta::Variable = "Wert"`

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

(Die Variable für Geschlecht ist `sex [m,f]`).

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



# Nach Metadaten filtern

- Nach der Muttersprache 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.*/
```

# 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.\*/\*

# Zusammenfassung: Operatoren (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 *außer* die Menge)
- (a|b)** a oder b
- a{2,3}** a 2 bis 3 mal

# Zusammenfassung: Operatoren für Tokenrelationen in ANNIS

#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

# Sprachkürzel in Falko (Auswahl)

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

**Herzlichen Dank!**

# Referenzen

- **Vorstellung des Falko-Korpus:**

Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik (2008): Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache* 45 (2), S. 67–73.

- **Ausformulierung der Erhebungs- und Korpusannotationsrichtlinien:**

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].

- **Vorstellung des ANNIS-Suchsystems:**

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*,<sub>35</sub>