

7. Linguistischer
Methodenworkshop
22.2. - 24.2.2016



Korpus I

22.02.2016 | 10.30-13.30 Uhr | Raum 1.305

Carolin Odebrecht

Humboldt-Universität zu Berlin

Kontakt



Carolin Odebrecht

carolin.odebrecht@hu-berlin.de

DFG Projekt LAUDATIO

Forschungsdatenrepository – speziell für historische Korpora

<http://www.laudatio-repository.org>

Humboldt-Universität zu Berlin

Korpuslinguistik und Morphologie

<http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik>

Methodenworkshop

Korpus I

Fachbereiche
Erwartungen
Vorkenntnisse

Überblick

- Die Grundfragen der Korpuslinguistik
 - Einführung und Beispiele
- Korpuserstellung
 - Text
 - Tokenisierung
 - Annotation
- Suche in Korpora
 - Anfragesprache
 - Suche nach Tokenannotation


Forschung mit Korpora

Forschungsfrage!

- Jede korpuslinguistische Untersuchung basiert auf einer Forschungsfrage!
 - Motivation für die Auswahl der Korpora/ Auswahl des sprachlichen Materials
 - Motivation für die Erstellung/ Aufbereitung, z.B. Annotationen
 - Motivation für die Analyse und Auswertung

Beispiele Korpora

- [RIDGES](#) – historisches Deutsch (Odebrecht et al. eingereicht), frei zugänglich
 - Register in German Diachronic Science, Kräuterkundekorpus

Kräutern	Kräutern	Alchimistische Praktik 1603	
Kraut	Kraut	Alchimistische Praktik 1603	
kraut	kraut	Alchimistische Praktik 1603	
Kreutern	Kreutern	Alchimistische Praktik 1603	
Kreutter	Kreutter	Alchimistische Praktik	
Kreüter	kreüter	Fuchs New Kreüterbuch	

4 ⓘ Path: Ridges_Herbology_Version_2.0 > Ridges_v2 > flora.saturnizans.1722

nichts verhalten will , ob **ichs** gleich thun könnte , weil

normalizations

clean	nichts	verhalten	will	,	ob	ichs	gleich	thun	könte	
dipl	nichts	verhalten	will	,	ob	ichs	gleich	thun	könte	
norm	nichts	verhalten	will	,	ob	ich	es	gleich	tun	könnte

Beispiele Korpora

- [BeMaTaC](#) Dialogdaten (Sauer & Lüdeling erscheint), frei zugänglich
 - Berlin Map Task Corpus, gesprochene Sprache

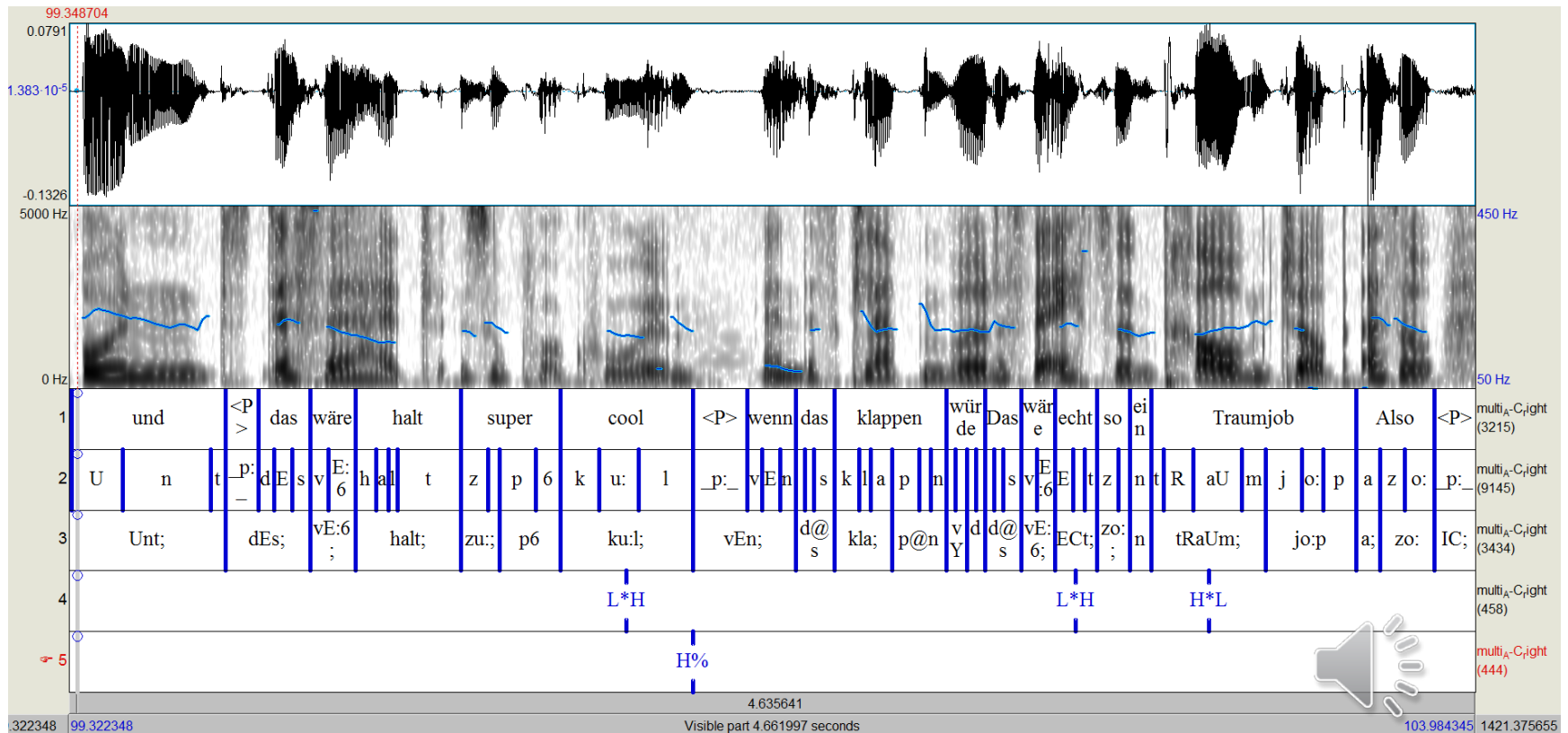
tok	0	1	2	3	4	5	6	7
instructor_dipl		🔊 dann	🔊 legen	🔊 wir	🔊 mal	🔊 los	🔊 mit	🔊 der
instructor_norm		🔊 dann	🔊 legen	🔊 wir	🔊 mal	🔊 los	🔊 mit	🔊 der
instructor_lemma		🔊 dann	🔊 legen	🔊 wir	🔊 mal	🔊 los	🔊 mit	🔊 die
instructor_pos		🔊 ADV	🔊 VFIN	🔊 PPER	🔊 ADV	🔊 ADJD	🔊 APPR	🔊 ART
instructor_utt		🔊 utt						



Trefferreferenzlink <https://korpling.german.hu-berlin.de/annis3/?id=0597f79c-485f-41bb-86b0-9212227040cc>

Beispiele Korpora

- [GECO](#) – (Schweitzer & Lewandowski 2013)
 - spontansprachlicher, freier Dialog



Beispiele Korpora

- [Shenoute.a22](#), Sahidisch (Zeldes & Schroeder 2015), frei zugänglich
 - [Coptic Scriptorium](#)

ε̅ρ̅ρ̅αι̅ ε̅ν̅σ̅υ̅ν̅α̅γ̅ω̅γ̅η̅ τ̅η̅ρ̅ο̅υ̅ μ̅π̅χ̅ο̅ε̅ι̅ς̅ · ε̅τ̅ρ̅ε̅π̅ν̅ο̅υ̅τ̅ε̅ σ̅ω̅ν̅τ̅ ε̅γ̅ν̅ο̅σ̅ μ̅μ̅η̅η̅ω̅ε̅ η̅ϕ̅κ̅τ̅ο̅

⊕ annotations (grid)

⊖ analytic view (document)

V	ADV	PREP	ART	N	CONJ	ADV	PREP	ART	N	ADV	PREP	ART	N		
ρ̅η̅	ε̅ρ̅ρ̅αι̅	ε̅	γ̅	β̅α̅γ̅κ̅α̅λ̅ι̅ο̅ν̅	η̅	ε̅ρ̅ρ̅αι̅	ε̅	γ̅	ω̅ρ̅ω̅ρ̅ο̅υ̅	ε̅ρ̅ρ̅αι̅	ε̅	σ̅ε̅	λ̅α̅α̅γ̅		
PREP	N	CCIRC	PPERS	VSTAT	PREP	PDEM	CREL	ANEGPST	PPERS	V	N	PREP	PPERO		
η̅	ε̅ρ̅ρ̅αι̅	ε̅	ϕ̅	τ̅η̅τ̅ω̅ν̅	ε̅	η̅αι̅	ε̅	μ̅π̅	ο̅γ̅	ο̅γ̅ε̅ρ̅	σ̅α̅ρ̅η̅ε̅	η̅α̅	ϕ̅		
PREP	ART	N	CONJ	ART	N	PREP	ART	N	CREL	V	PREP	V	PREP	ART	N
ρ̅ι̅τ̅μ̅	π̅	ε̅λλ̅ω̅	ε̅ι̅μ̅η̅τ̅ι̅	π̅	μ̅α̅	η̅	η̅	ρ̅ω̅μ̅ε̅	ε̅τ̅	ω̅ρ̅ω̅η̅ε̅	μ̅	μ̅α̅τ̅ε̅	μ̅η̅	η̅	ε̅λλ̅ο̅
APST	PPERS	V	PREP	N	CONJ	PDEM	N	PREP	ACAUS	PPERS	V	PREP	N	PPERO	
α̅	γ̅	ρ̅ε̅α̅ρ̅	η̅	ρ̅ο̅μ̅π̅ε̅	α̅γ̅ω̅	η̅ε̅ι̅	κ̅ο̅ο̅γ̅ε̅	ε̅	τ̅ρ̅ε̅	γ̅	ω̅ρ̅η̅ε̅	η̅	τ̅ο̅ο̅τ̅	ϕ̅	
PREP	ART	N													
μ̅	π̅	ε̅λλ̅ο̅													

.. if one urinates into a vessel with a narrow neck or jar, into any other vessel like these, without having been ordered by the elder, except for the place of those who are very ill and the elders of advanced years and these others, he is to ask the elder.

Korpusreferenzlink
https://corpling.uis.georgetown.edu/annis/scriptorium#_c=c2hlm91dGUuYTIy

Beispiele Korpora

- Dornröschen, [Märchenkorpus](#), frei zugänglich
 - [Textbewegung](#)

1			Path: Maerchenkorpus > grimm_dornroeschen_251-254.tagged (tokens 1 - 7)									
50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,	
@ord@	Dornröschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,	
ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$,	



Trefferreferenzlink <https://korpling.org/annis3/?id=e1be6871-20c3-4968-807f-13156b00e0e2>

Bild:

https://de.wikipedia.org/wiki/Dornr%C3%B6schen#/media/File:Dornr%C3%B6schen_Joseph_Albert_01.jpg

Beispiele Korpora

- Parlamentsreden im Deutschen Bundestag, Zugang benötigt
 - Protokolle/stenographische Berichte

```
1 ⓘ Path: Parlamentsreden_Deutscher_Bundestag > Parlamentskorpus > 16070 (tokens 57209 - 57219)
] : Die EU-Kommission ist doof und die Bundesregierung ist schlau
] : d EU-Kommission sein doof und d Bundesregierung sein schlau
$( $. ART NN VAFIN ADJD KON ART NN VAFIN ADJD
```

Deutscher Bundestag – 18. Wahlperiode – 7. Sitzung. Berlin, Mittwoch, den 15. Januar 2014 331

(A)

(C)

7. Sitzung

Berlin, Mittwoch, den 15. Januar 2014

Beginn: 13:00 Uhr

Vizepräsident Peter Hintze:
Die Sitzung ist eröffnet.

Nach einer interfraktionellen Vereinbarung soll der Jahresbericht 2013 des Wahlprüfungsausschusses auf der Drucksache 17/12050 aus der 17. Wahlperiode federführend dem Verteidigungsausschuss und zur Mitberatung dem Rechtsausschuss überwiesen werden. Sind Sie damit einverstanden? – Ich höre keinen Widerspruch. Dann ist so beschlossen.

Ich rufe den Tagesordnungspunkt 1 auf:

Befragung der Bundesregierung

Die Bundesregierung hat als Thema der heutigen Kabinensitzung mitgeteilt: Migrationsbericht 2013.
Das Wort für einen einleitenden fünfminütigen Bericht hat der Bundesminister des Innern, Herr Dr. Thomas de Maizière. – Herr Bundesminister, bitte.

Dr. Thomas de Maizière, Bundesminister des In-

nen macht insgesamt einen positiven Wanderungssaldo von 370 000 Menschen.

Ich muss kurz erläutern, wie diese Zahlen ermittelt werden; sie stammen übrigens aus der amtlichen Wanderungstatistik. „Fortzug“ heißt: Jemand meldet sich ab und hat keinen anderen Wohnsitz im Inland. „Zuzug“ heißt: Er meldet sich hier an. – Was sich dahinter verbirgt, ergibt sich daraus natürlich nicht per se. Deswegen will ich anhand einiger Hilfszahlen zeigen, was die Zahlen verdeutlichen.

Wenn wir eine große Zuwanderung haben, dann haben wir immer auch eine große Abwanderung. Ein Beispiel ist dabei auch um dieselben Staaten, zum Beispiel Polen. Das heißt: Wir haben in Deutschland eine sehr starke Bewegung – herein und heraus – was alle Formen von Zuwanderung betrifft. Ein Hilfsindikator – das finden Sie auch in dem Bericht –, dass sich mehr als zwei Drittel der fortgezogenen ausländischen Staatsangehörigen wiederum als vier Jahre im Bundesgebiet aufgehalten ha-

Protokoll:

<http://dipbt.bundestag.de/doc/btp/18/18007.pdf>

Korpus

Einordnung

Korpus

Alle Beispiele:

→ Beobachtung authentischer Sprachdaten

Authentisch meint in diesem Zusammenhang tatsächlich geäußerte, nicht weiter künstlich erzeugte oder erdachte Äußerungen.

Das können sein z.B.:

- Chats, Foreneinträge, Blogs
- Briefe, Urkunden, Rechtstexte
- Unterhaltungen, Witze, Lernertexte, Unterrichtsgespräche
- Bibeltexte, Plenarreden ...

Korpus

- Diese Art von Daten werden typischerweise in einem Korpus zusammengefasst aufbereitet!

Korpus:

- Sammlung von digitalisierten, sprachlichen Äußerungen
- enthält Metadaten zu den Daten und
- linguistische Annotationen

(Lemnitzer und Zinsmeister 2006, 7)

- gesprochene/geschriebene Sprache
- Audio/Video/Text
- Muttersprachler/Lerner
- alle möglichen Register

Korpus

→ Bewertung der Eignung von Korpora im Hinblick auf Fragestellung

Je nach Fragestellung Korpora ...

a) mit Texten, die speziell für diese Fragestellung erzeugt/zusammengestellt wurden.

Nutzung des gesamten Korpus, homogen

b) mit Texten, die zu anderen Zwecken erzeugt wurden.

Nutzung von Teilen eines Korpus, heterogen

Korpus

Korpora können...

- **wachsen** oder einen **festen** Umfang haben
 - stetig mehr Daten hinzufügen
- **repräsentativ** oder **spezifisch** sein
 - stellvertretend für einen bestimmten Sprachstand oder spezifisch für ein bestimmtes Register/ bestimmte Sprechergruppen etc.
- **balanciert/heterogen** oder **unbalanciert/homogen** sein
 - Texte aus unterschiedlichen Bereichen oder Texte aus nur einem Register
- **opportunistisch** oder **kontrolliert** sein
 - „alles was ich kriegen kann“ oder Textauswahl nach bestimmten Kriterien

Korpus

Repräsentativität:

- Begriff aus der Statistik:

Man möchte bestimmte Eigenschaften einer Menge (von Personen, Wörtern, Bäumen etc.) untersuchen, die aber zu groß ist, um in ihrer Gesamtheit angeschaut werden zu können.

→ Ziehen einer Stichprobe aus der sogenannten Grundgesamtheit (population)

- Bloß: Was ist die Grundgesamtheit von X, bspw. Sprache?

Korpus

- Was ist Ihre Forschungsfrage?
- Was für ein Korpus benötigen Sie, um diese zu beantworten?

Grundfragen der Korpuslinguistik

Die Grundfragen der Korpuslinguistik

- Wie finde ich, was ich brauche?

→ Erstellen von KORPORA

– Wie bereite ich Daten korpuslinguistisch auf?

- Tokenisierung
- Annotation

➤ Erster Teil des Workshops

➤ Korpuserstellung

Korpuserstellung

Text, Tokenisierung, Annotation

Text

- Wir brauchen einen Text!
- Dornröschen 😊



Bild:

https://de.wikipedia.org/wiki/Dornr%C3%B6schen#/media/File:Dornr%C3%B6schen_Joseph_Albert_01.jpg

Text

Dornröschen

Vor Zeiten war ein König und eine Königin, die sprachen jeden Tag „ach, wenn wir doch ein Kind hätten!“ und kriegten immer keins. Da trug sich zu, als die Königin einmal im Bade saß, daß ein Frosch aus dem Wasser ans Land kroch und zu ihr sprach, „dein Wunsch wird erfüllt werden, ehe ein Jahr vergeht, wirst du eine Tochter zur Welt bringen.“ Was der Frosch gesagt hatte, das geschah, und die Königin gebar ein Mädchen, das war so schön, daß der König vor Freude sich nicht zu lassen wußte und ein großes Fest anstellte. Er ladete nicht blos seine Verwandte, Freunde und Bekannte, sondern auch die weisen Frauen dazu ein, damit sie dem Kind hold und gewogen wären. Es waren ihrer dreizehn in seinem Reiche, weil er aber nur zwölf goldene Teller hatte, von welchen sie essen sollten, so mußte eine von ihnen daheim bleiben. Das Fest ward mit aller Pracht gefeiert, und als es zu Ende war, beschenkten die weisen Frauen das Kind mit ihren Wundergaben: die eine mit Tugend, die andere mit Schönheit, die dritte mit Reichthum, und so mit allem, was auf der Welt zu wünschen ist. Als elfe ihre Sprüche eben gethan hatten, trat plötzlich die dreizehnte herein. [...]

Walter, Maik; Maerchenkorpus (Erste Veröffentlichung des Korpus.) Version: 1.0. Humboldt-Universität zu Berlin. <http://www.textbewegung.de/>.
<http://hdl.handle.net/11022/0000-0000-1F5B-9>

Korpuserstellung

Text, **Tokenisierung**, Annotation

Tokenisierung

- Zerlegen einer 'unorganisierten' Zeichenabfolge (=Ausgangsdaten für Korpus) in kleinste zählbare technische Einheiten
- Tokenisieren (muss **immer** erfolgen)
 - manuell oder automatisch

Tokenisierung

Begriff ‚Token‘:

- kleinste technische im Korpus zählbare Einheit, Einheit mit **beliebiger Größe** (Buchstabe, Silbe, Satz, Absatz, Text, ...)
- Wenn Sie nach Sätzen tokenisieren, können Sie nur ganze Sätze annotieren und finden, nicht aber einzelne Wörter innerhalb von Sätzen!
- meistens aber:
„eine von Leerzeichen (das umfasst Tabulatorzeichen und Zeilenumbrüche) oder Interpunktion begrenzte Folge von Buchstaben oder Ziffern“
(Evert & Fitschen 2001, 371)
≈ graphematisches Wort (?)

Tokenisierung

- Tokenannotation
 - (ausnahmslos) jedem Token wird genau ein Wert zugeordnet
 - automatisch oder manuell
- häufige Tokenannotation für u.a.
 - Lemmatisierung
 - Wortartentagging (Part of Speech Tagging)
 - morphologisches Tagging (Flexion)

Tokenisierung

Dornröschen

Vor Zeiten war ein König und eine Königin, die sprachen jeden Tag „ach, wenn wir doch ein Kind hätten!“ und kriegten immer keins. Da trug sich zu, als die Königin einmal im Bade saß, daß ein Frosch aus dem Wasser ans Land kroch und zu ihr sprach, „dein Wunsch wird erfüllt werden, ehe ein Jahr vergeht, wirst du eine Tochter zur Welt bringen.“ Was der Frosch gesagt hatte, das geschah, und die Königin gebar ein Mädchen, das war so schön, daß der König vor Freude sich nicht zu lassen wußte und ein großes Fest anstellte. Er ladete nicht blos seine Verwandte, Freunde und Bekannte, sondern auch die weisen Frauen dazu ein, damit sie dem Kind hold und gewogen wären. Es waren ihrer dreizehn in seinem Reiche, weil er aber nur zwölf goldene Teller hatte, von welchen sie essen sollten, so mußte eine von ihnen daheim bleiben. Das Fest ward mit aller Pracht gefeiert, und als es zu Ende war, beschenkten die weisen Frauen das Kind mit ihren Wundergaben: die eine mit Tugend, die andere mit Schönheit, die dritte mit Reichthum, und so mit allem, was auf der Welt zu wünschen ist. Als elfe ihre Sprüche eben gethan hatten, trat plötzlich die dreizehnte herein. [...]

Walter, Maik; Maerchenkorpus (Erste Veröffentlichung des Korpus.) Version: 1.0. Humboldt-Universität zu Berlin. <http://www.textbewegung.de/>.
<http://hdl.handle.net/11022/0000-0000-1F5B-9>

Tokenisierung

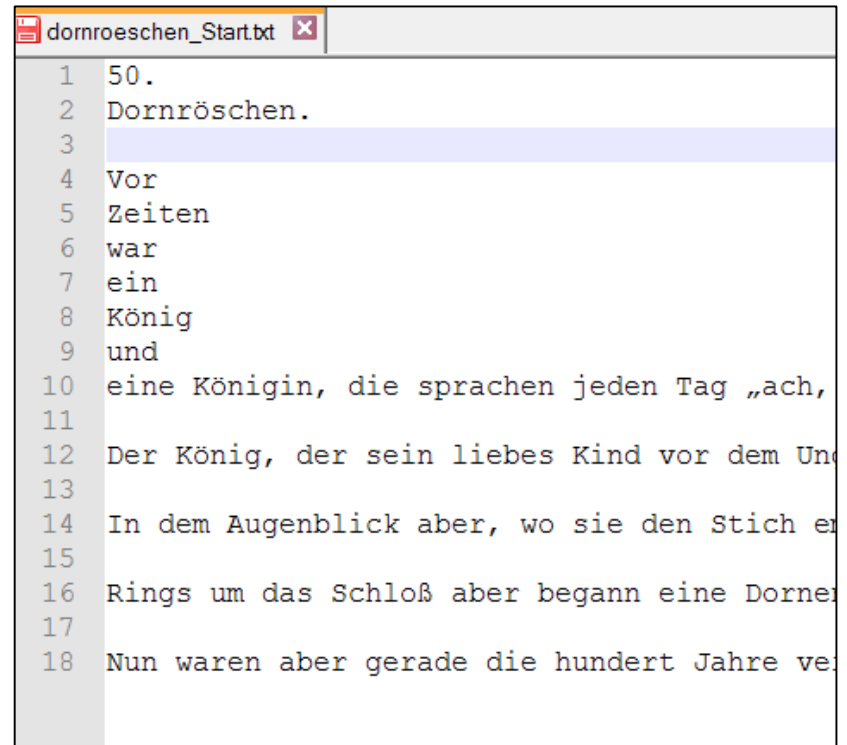
Korpusl.zip

- Download unter <https://www.linguistik.hu-berlin.de/forschung/methodenworkshop/methodenworkshop/Korpora1>
 - Entpacken + Speicherung auf Desktop
- Tokenisieren Sie die Beispieldatei `dornroeschen_Start.txt` in einem Texteditor!
- Das Trennen von Token kann durch ein beliebiges, einheitliches, nicht mehrdeutiges Element erfolgen (Absatz, Tab, Leerzeichen, etc.)
 - Probleme? Fragen?

Tokenisierung

Frau Holle

- hier pro Zeile ein Token
 - Vorteil: nicht ambig wie #/,.(
 - nützlich für Weiterverarbeitung



```
dornroeschen_Start.txt
1 50.
2 Dornröschen.
3
4 Vor
5 Zeiten
6 war
7 ein
8 König
9 und
10 eine Königin, die sprachen jeden Tag „ach,
11
12 Der König, der sein liebes Kind vor dem Un
13
14 In dem Augenblick aber, wo sie den Stich er
15
16 Rings um das Schloß aber begann eine Dorne
17
18 Nun waren aber gerade die hundert Jahre ve
```


Tokenisierung

- Beobachtung:
 - manuelles Tokenisieren ist sehr aufwändig
 - notwendige Entscheidungen hinsichtlich Interpunktion
 - „Trenne Satzzeichen mit Leerzeichen ab.“
- Aber: Sind solche rein graphemisch definierten Tokens immer die Einheit, mit der man weiterarbeiten will?
- Wie definiere ich „Einheit“?
- Wie definiere ich „Token“?

Tokenisierung Probleme

- Manche Einheiten sollen zusammen bleiben:
 - Zahlen:
20 000, 030-2093 9799, BLZ 111 111 11
→ Heuristiken, reguläre Ausdrücke
 - Namen, feste Verbindungen:
New York, Weil der Stadt, en passant, Vereinte Nationen, der Deutsche Bundestag
 - Abkürzungen:
bspw., etc., ...
 - Wie behandeln Sie *z.B.*?

Tokenisierung Probleme

- Partikelverben (Beispiele aus dem Märchenkorpus)
 - [...] *job er ihn als Lehrling **annehmen** wollte.*
 - *Also ward sie **angenommen** zum Küchenmädchen für geringen Lohn.*
 - [...] ***nahm** die alte Hexe die Gestalt der Kammerfrau **an** [...]*
- Ist *nahm ... an* nicht eine Einheit?
- Wie stelle ich sicher, dass ich 3 Formen des Infinitivs *annehmen* zählen kann?
 - Unterscheidung der Bedeutung
 - syntaktische Analyse nötig
 - Tokenisierung kommt hier nicht weiter

Tokenisierung Probleme

- Einige Einheiten **können** aufgeteilt werden, müssen aber nicht (Forschungsfrage?):
 - *beim, zum, gibt's, siehste*
 - Listen, reguläre Ausdrücke, Heuristiken
(Will man Informationen über die ursprüngliche Form behalten?
Stichwort „Normalisierung“)
- bestimmte Sonderzeichen in Formeln u. ä.: Desambiguierung schwierig
 - *42,195 , 8:04:08*
Patienten der WOS-(West of Scotland)-Studie

Tokenisierung

Zusammenfassung

- Entscheidung für eine Definition von „Token“
 - immer Interpretation
 - Fehleranfälligkeit
 - Auswirken auf alle späteren Vorverarbeitungsschritte
- Einige Entscheidungen lassen sich ohne weiteres linguistisches Wissen nicht sicher treffen.
- Ein „dummes“ Verfahren ist jedoch meistens konsequenter als Menschen!

Normalisierung

- bei Nicht-Standardvarietäten sinnvoll, z.B. in
 - gesprochene Sprache
 - Lernersprache
 - historische Sprachstufen
- häufig zusätzlich (und nicht an Stelle der) authentischen sprachlichen Äußerung
 - Annotation
- auch hier, abhängig von der Forschungsfrage
 - verschiedene Definition von Normalisierung
 - Was wird normalisiert?
 - Nach welcher „Norm“ wird normalisiert?

Normalisierung

- Typischer Fall für Normalisierungen: historische Schreibvarianten
 - z.B. Dornröschen Text (1857), Maerchenkorpus

*Als **elfe** ihre Sprüche eben **gethan** hatten, trat plötzlich die dreizehnte herein. Sie wollte sich dafür rächen **daß** sie nicht eingeladen war, und ohne jemand zu grüßen oder nur anzusehen, rief sie mit lauter Stimme „die Königstochter soll sich in ihrem **funfzehnten** Jahr an einer Spindel stechen und **todt** hinfallen.“*

Trefferreferenz: <https://korpling.german.hu-berlin.de/annis3/?id=cc65e453-ee98-4357-8f3a-6617ea0bc363>

Normalisierung

- Typischer Fall für Normalisierungen: Fehler von Lernern einer Sprache
 - z.B. Falko, FalkoEssayL2v2.0 > cbs009_2007_10_L2v2.0
*Sollte man diese Frage **bejaren**, dann **gluabe** ich ganz **erhlich** nicht, dass man verstanden hat, was Feminismus ist [...]*

Trefferreferenz: <https://korpling.german.hu-berlin.de/annis3/?id=a4032d7e-035d-44a6-9962-20d39c2cf705>

Korpuserstellung

Text, Tokenisierung, Annotation

Bislang haben wir

- Text
 - in Einheiten unterteilt
 - nach graphematischen Aspekten
- wir wissen nicht, was „drin steckt“
 - welche Lexeme
 - welche Wortarten
 - welche anderen linguistischen Phänomene

```
dornroeschen_Tokenized.txt
1 50.
2 Dornröschen
3 .
4 Vor
5 Zeiten
6 war
7 ein
8 König
9 und
10 eine
11 Königin
12 ,
13 die
14 sprachen
15 jeden
16 Tag
17 "
18 ach
19 ,
20 wenn
21 wir
22 doch
23 ein
24 Kind
25 hätten
26 !
27 "
```

Wortartentagging

- **„Annotation“** ganz allgemein
 - Interpretationen
 - als Markierung von Dingen, die man später im Korpus systematisch wiederfinden/zählen/auswerten will
 - Informationen so getrennt wie möglich annotieren
 - verschiedene Typen (Art der Zuweisung)
- **„Taggen“** = Zuordnung
(im Prinzip beliebiger) linguistischer Informationen zu (im Prinzip beliebigen) Texteinheiten
- **Wortartentagging**
 - manchmal verkürzt = Wortartzuweisung
 - Part of Speech; pos, PoS o.ä.

Wortartentagging

- auf Token basiert
 - Wiederholung: Token ist die kleinste technische zählbare Einheit im Korpus, meist = graphemisches Wort
- Basis für weitere linguistische Annotationen
- Verwendung als Eingabe für weitere computerlinguistische Anwendungen
(Parser, semantische Verarbeitung, ...)

Wortartentagging

Beispiele:

- Einschränken ambiger Wortformen (Bsp. Parlamentsreden, ANNIS)
 - *wenn Sie **meinen**, ich argumentiere einseitig*
 - *das berücksichtige ich bei **meinen** Wahlentscheidungen*
- Finden aller Kandidaten einer bestimmten Wortart.
- Finden aller Wörter in einer bestimmten Sequenz aufeinanderfolgender Wortarten (Bsp. Märchenkorpus, [ANNIS Suche](#))
 - *Darin stand ein **schöner großer Baum** an dem die herrlichsten Birnen hiengen.*
 - *[...] noch ein **kleines verbuttetes Aschenputtel** da [...]*
 - *[...] nahm der das **schöne weiße Gebein** heraus [...]*
 - *[...] daß die **schöne junge Königin** bald ersticken mußte.*

Wortartentagging

- Ziel: jedes Token erhält ein pos-Tag:
 - *50./ Dornröschen/ ./ Vor/ Zeiten/ war/ ein/ König/ und/ eine/ Königin/ ,/ die/ sprachen/ jeden/ Tag/ "/ ach/ ,/ wenn/ wir/ doch/ ein/ Kind/ hätten/ !/ "/*
- Tagset: Definition einer Menge von pos-Tags
 - Kategorien für Wortartentags
- linguistische Forschungsfrage → Klassifizierung von Wortarten
 - syntaktische Kriterien, morphologische Kriterien

Wortartentagging

- Ziel: jedes Token erhält ein pos-Tag:
 - *50./ADJA Dornröschen/NN ./\$. Vor/APPR Zeiten/NN war/VAFIN ein/ART König/NN und/KON eine/ART Königin/NN ,/\$, die/PRELS sprachen/VVFIN jeden/PIAT Tag/NN "/\$(ach/ITJ ,/\$, wenn/KOUS wir/PPER doch/ADV ein/ART Kind/NN hätten/VAFIN !/\$. "/\$(*
- Tagset: Definition einer Menge von pos-Tags
 - Kategorien für Wortartentags nach dem Stuttgart Tübingen Tag Set (STTS, Schiller et al. 1999)
- linguistische Forschungsfrage → Klassifizierung von Wortarten
 - syntaktische Kriterien, morphologische Kriterien

Wortartentagging

- Entwicklung eines **eigenen** Tagsets
- Nutzung von **bestehenden** Tagsets

Tagsets für deutsche Korpora
(vgl. Rapp & Lezius 2001):

Name der Richtlinien	groß	klein
IBM Heidelberg	689	33
Uni Münster	143	54
STTS (Stuttgart/Tübingen Tag Set)		54
Morphy (Paderborn)	1000	52
...		

Wortartentagging

- Öffnen Sie die dornroeschen_Uebung.exb im EXMARaLDA Partitur Editor!
 - Weisen Sie jedem Token pos-Tags aus der folgenden Liste zu!
 - STTS
 - X[POS]
- Wann welchen Tag zuweisen?
 - Wann trifft die Beschreibung zu?
 - Konkurrierende Bezeichnungen?

STTS

POS =	Beschreibung	Beispiele
ADJA ADJD	attributives Adjektiv adverbiales oder prädikatives Adjektiv	<i>[das] große [Haus]</i> <i>[er fährt] schnell</i> <i>[er ist] schnell</i>
ADV	Adverb	<i>schon, bald, doch</i>
APPR APPRART APPO APZR	Präposition; Zirkumposition links Präposition mit Artikel Postposition Zirkumposition rechts	<i>in [der Stadt], ohne [mich]</i> <i>im [Haus], zur [Sache]</i> <i>[ihm] zufolge, [der Sache] wegen</i> <i>[von jetzt] an</i>
ART	bestimmter oder unbestimmter Artikel	<i>der, die, das,</i> <i>ein, eine</i>
CARD	Kardinalzahl	<i>zwei [Männer], [im Jahre] 1994</i>
FM	Fremdsprachliches Material	<i>[Er hat das mit “]</i> <i>A big fish [” übersetzt]</i>
ITJ	Interjektion	<i>mhm, ach, tja</i>

STTS

KOUI	unterordnende Konjunktion mit “zu” und Infinitiv	<i>um [zu leben], anstatt [zu fragen]</i>
KOUS	unterordnende Konjunktion mit Satz	<i>weil, daß, damit, wenn, ob</i>
KON	nebenordnende Konjunktion	<i>und, oder, aber</i>
KOKOM	Vergleichspartikel, ohne Satz	<i>als, wie</i>
NN	Appellativa	<i>Tisch, Herr, [das] Reisen</i>
NE	Eigennamen	<i>Hans, Hamburg, HSV</i>

STTS

PDS	substituierendes Demonstrativpronomen	<i>dieser, jener</i>
PDAT	attribuierendes Demonstrativpronomen	<i>jener [Mensch]</i>
PIS	substituierendes Indefinitpronomen	<i>keiner, viele, man, niemand</i>
PIAT	attribuierendes Indefinitpronomen ohne Determiner	<i>kein [Mensch], irgendein [Glas]</i>
PIDAT	attribuierendes Indefinitpronomen mit Determiner	<i>[ein] wenig [Wasser], [die] beiden [Brüder]</i>
PPER	irreflexives Personalpronomen	<i>ich, er, ihm, mich, dir</i>
PPOSS	substituierendes Possessivpronomen	<i>meins, deiner</i>
PPOSAT	attribuierendes Possessivpronomen	<i>mein [Buch], deine [Mutter]</i>
PRELS	substituierendes Relativpronomen	<i>[der Hund,] der</i>

STTS

PRELAT	attribuierendes Relativpronomen Relativpronomen	<i>[der Mann ,] dessen [Hund]</i>
PRF	reflexives Personalpronomen	<i>sich, einander, dich, mir</i>
PWS	substituierendes Interrogativpronomen	<i>wer, was</i>
PWAT	attribuierendes Interrogativpronomen	<i>welche [Farbe], wessen [Hut]</i>
PWAV	adverbiales Interrogativ- oder Relativpronomen	<i>warum, wo, wann, worüber, wobei</i>
PROAV	Pronominaladverb	<i>dafür, dabei, deswegen, trotzdem</i>
PTKZU	“zu” vor Infinitiv	<i>zu [gehen]</i>
PTKNEG	Negationspartikel	<i>nicht</i>
PTKVZ	abgetrennter Verbzusatz	<i>[er kommt] an, [er fährt] rad</i>
PTKANT	Antwortpartikel	<i>ja, nein, danke, bitte</i>
PTKA	Partikel bei Adjektiv oder Adverb	<i>am [schönsten], zu [schnell]</i>
TRUNC	Kompositions-Erstglied	<i>An- [und Abreise]</i>

STTS

VVFIN	finites Verb, voll	<i>[du] gehst, [wir] kommen [an]</i>
VVIMP	Imperativ, voll	<i>komm [!]</i>
VVINFINF	Infinitiv, voll	<i>gehen, ankommen</i>
VVIZU	Infinitiv mit "zu", voll	<i>anzukommen, loszulassen</i>
VVPP	Partizip Perfekt, voll	<i>gegangen, angekommen</i>
VAFIN	finites Verb, aux	<i>[du] bist, [wir] werden</i>
VAIMP	Imperativ, aux	<i>sei [ruhig !]</i>
VAINFINF	Infinitiv, aux	<i>werden, sein</i>
VAPP	Partizip Perfekt, aux	<i>gewesen</i>
VMFIN	finites Verb, modal	<i>dürfen</i>
VMINFINF	Infinitiv, modal	<i>wollen</i>
VMPP	Partizip Perfekt, modal	<i>[er hat] gekonnt</i>
XY	Nichtwort, Sonderzeichen enthaltend	<i>D2XW3</i>
\$,	Komma	<i>,</i>
\$.	Satzbeendende Interpunktion	<i>. ? ! ; :</i>
\$(sonstige Satzzeichen; satzintern	<i>– []()</i>

Wortartentagging Tools

- STTS: wird oft zum Taggen von Korpora anhand freiverfügbarer Programme („Tagger“) verwendet:
 - *TreeTagger* (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>)
 - *RFTagger* (<http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/>)
 - *TNT* (<http://www.coli.uni-saarland.de/~thorsten/tnt/>)

Wortartentagging Probleme

- bei diesem automatischen Annotationsverfahren gibt es systematische Fehler
- Beispiele:
 - *nach 13 Jahren **Kohl**/NN*
[Mannheimer Korpus, TreeTagger]
 - *ich schmiede mir **ne**/FM **schicke**/VVFIN rubinklinge und kauf mir für viel gold beim örtlichen alchimisten ein **gift**/ADJD um meine neue **waffe**/FM noch besser zu machen , und was passiert als ich das **zeug**/ADJD **auftrage**/NN ?* [www.worldofgothic.de, TreeTagger]

Korpuserstellung

Text, Tokenisierung, Annotation

Lemmatisierung

- Beispiel
 - Herausfinden, wie häufig kommt in Korpus X das Verb **meinen** im Gegensatz zu **sagen** vor.
 - Notwendig: Formen
 - *meinst, meintest, meinst, meinen (!), meine (!), ...*
 - Wie findet man alle, aber keine anderen Formen?
 - **Lemmatisieren** = auf ein Lemma zurückführen

Lemmatisierung

Kleine Definition:

- **Lemma:** abstrakte Grundform, Lexikoneintrag
- **Wortform:** bestimmte Form in einem Paradigma (Problem: Synkretismus)
- **getaggtetes Token:** Wortform mit Annotation (Lemma, Wortart, Flexionsmorphologie, ...)
- aus dem Beispiel: *meine* als *meinen* oder als *mein*

Lemmatisierung

- Öffnen Sie die dornroeschen_Uebung.exb im EXMARaLDA Partitur Editor!
- Lemmatisieren Sie jedes Token!
 - X[lemma]
 - Wie sähen die Lemmata für die vorhandenen Wortformen und ihre Wortarten aus? Welche würden Sie festlegen?

Lemmatisierung

Lemmatisierung ist (bei flektierbaren Wörtern) nicht anhand sprachinhärenter Eigenschaften entscheidbar, sondern ist arbiträr!

- *50./@ord@ Dornröschen/Dornröschen ./ Vor/vor Zeiten/Zeit war/sein ein/ein König/König und/und eine/ein Königin/Königin ,/ die/PRELS/d sprachen/sprechen jeden/jed Tag/Tag "/" ach/ITJ/ach ,/\$/, wenn/wenn wir/wir doch/doch ein/ART/ein Kind/Kind hätten/haben !/! "/„ [...]*
- häufig Konventionen: für Verben > Infinitiv Präsens, für Nomen > Nominativ Singular

Wortartentagging & Lemmatisierung Probleme

- trotz dieser Annotationen
 - Ambiguitäten können bleiben
 - *Bank*
 - *lying*
 - *verlassen*
 - immer problematisch: unbekannte Wörter
 - Partikelverben nicht aufgelöst
- Also ward sie **angenommen** zum Küchenmädchen für geringen Lohn.*
- [...] **nahm** die alte Hexe die Gestalt der Kammerfrau **an** [...]*
- Sind *haben* und *sein* immer Hilfsverben?
 - morphologische und syntaktische Kategorien

Wortartentagging & Lemmatisierung Probleme

- Überlegen Sie:
- Wie sollten die unterstrichenen Formen
ein Glas voller Wasser
ein voller Eimer Wasser
lemmatisiert werden?
- Welche Wortart hat hier *verrückt*?
Er ist verrückt.
- Wortart und Lemma einer Wortform müssen eine
homogene Analyse ergeben.

Automatisches Wortartentagging

- STTS: wird oft zum Taggen von Korpora anhand freiverfügbarer Programme („Tagger“) verwendet:
 - *TreeTagger* (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>)
- hier: Treetagger (Schmid 1994)
 - Anleitung: Korpuserstellung siehe Anhang

Vergleich eigene und automatische Lemmatisierung und pos-Tagging

File Edit View Transcription Tier Event Timeline Format Help

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
X [txt]	50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,	die	sprachen	jeden	Tag	"	ach	,
X [pos]																			
X [lemma]																			

File Edit View Transcription Tier Event Timeline Format Help

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
X [txt]	50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,	die	sprachen	jeden	Tag	"	ach	,
X [pos]	ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$,	PRELS	VVFIN	PIAT	NN	\$(ITJ	\$,
X [lemma]	@ord@	Dornröschen	.	vor	Zeit	sein	ein	König	und	ein	Königin	,	d	sprechen	jed	Tag	"	ach	,

Methodische Bemerkung

Annotationen

- Im Prinzip: Alles ist annotierbar!
 - Annotation als Basis für jede weitere (linguistische) Analyse
 - immer auch (linguistische) Interpretation
 - mit Fehlern (eines Tools aber auch eines Annotators) muss gerechnet werden
 - Überprüfung von Annotationen
 - Motivation eines jeden Aufbereitungsschrittes durch Fragestellung
(vgl. Kübler & Zinsmeister 2015)
- **Was nicht annotiert ist (direkt wie indirekt), kann auch nicht gefunden werden!**
- **Dokumentation der Datenaufbereitung ist essentiell!**

Korpuserstellung

- Dornröschen enthält
 - Text, tokenisiert
 - nach graphematischen Aspekten
 - Wortarten und Lemmata
- Auswertung der Annotationen
 - Suche nach linguistischen Phänomenen
 - Suche nach Annotationen
- Korpus in einem Suchtool mit Hilfe der Suche nach Annotationen analysieren!
- Dornröschen ist ein Dokument im Märchenkorpus!

1	50.	ADJA	@ord@	
2	Dornröschen	NN	Dornröschen	
3	.	\$.	.	
4	Vor	APPR	vor	
5	Zeiten	NN	Zeit	
6	war	VAFIN	sein	
7	ein	ART	ein	
8	König	NN	König	
9	und	KON	und	
10	eine	ART	ein	
11	Königin	NN	Königin	
12	,	\$,	,	
13	die	PRELS	d	
14	sprachen	VVFIN	sprechen	
15	jeden	PIAT	jed	
16	Tag	NN	Tag	
17	"	\$("	
18	ach	ITJ	ach	
19	,	\$,	,	
20	wenn	KOUS	wenn	

Grundfragen der Korpuslinguistik

Die Grundfragen der Korpuslinguistik

- Wie finde ich, was ich brauche?

→ Finden von KORPORA

- Welche Datengrundlagen gibt es?
- Wo sind diese aufgeführt?
- Welche Daten wurden verwendet?
- Welche Annotationen wurden verwendet?

Finden von Korpora in ...

- Suchinterfaces für Korpora wie z.B.
 - ANNIS Such- und Visualisierungstool
 - Corpus Query Processor (CQP)
 - Corpus Search, Management and Analysis System (COSMAS II)
 - TigerSearch
- Projekt-Homepages von Korpora wie
 - siehe Beispiele aus Folien 8-14
- Forschungsdatenrepositorien oder -archiven wie z.B.
 - LAUDATIO-Repository für historische Korpora
 - Textgrid Repository für historische Korpora
 - Virtual Language Observatory (VLO), als Metasuche

Märchenkorpus

<http://www.laudatio-repository.org>

LAUDATIO-Repository

Home » View » Märchenkorpus

- Home
- Documentation
- View
- Search

Märchenkorpus Version 1.0

2014-01-20 15:21:04 ▾

Märchenkorpus Version 1.0, Humboldt-Universität zu Berlin, 1.0, 295880 Tokens, Erste Veröffentlichung des Korpus.

Formats: [txt](#), [treetaggeroutput](#), [reIANNIS](#)

Always quote citation when using data!

Walter, Maik; Märchenkorpus (Erste Veröffentlichung des Korpus.) Version: 1.0. Humboldt-Universität zu Berlin. <http://www.textbewegung.de/hdl.handle.net/11022/0000-0000-1F5B-9>

- ▶ **Corpus Märchenkorpus Version 1.0**
- ▶ **Documents**
- ▶ **Annotation**
- ▶ **PreparationStep**

Korpus

Märchenkorpus

- Walter, Maik; Maerchenkorpus (Version 1.0), Humboldt-Universität zu Berlin. <http://www.textbewegung.de/>.
<http://hdl.handle.net/11022/0000-0000-8211-9>

Grundfragen der Korpuslinguistik

Die Grundfragen der Korpuslinguistik

- Wie finde ich, was ich brauche?

→ Suchen in KORPORA

- Wie kann ich auf Korpora zugreifen?
- Wie kann ich in Korpora nach Annotationen suchen?



ANNIS

Suchen in Korpora

Suche in Korpora

Allgemeine Informationen zu Suchtools:

- nicht zu verwechseln mit Annotationstools, z.B.
 - EXMARaLDA Partitur Editor = Annotieren von ling. Kategorien (Teil 1 des Workshops)
 - ANNIS Such- und Visualisierung= Suche nach Annotationen in Korpora (Teil 2 des Workshops)
- enthalten häufig eine Menge an Korpora unterschiedlicher Herkunft
- besitzen eigene Suchabfragesprachen
- Sprachenlogik und Komponenten innerhalb der Sprachen z. T. sehr ähnlich
- Konzeption der Tools für unterschiedliche Daten
 - große vs. kleine Datenmengen
 - flache/einfache vs. tiefe/komplexe Annotationen

Suchen in Korpora und nach Annotationen mit ...



ANNIS

ANNotation of **Information Structure**

Search and Visualization in Multilevel Linguistic Corpora

<https://korpling.org/annis3/>



ANNIS

- Browser basiertes Such- und Visualisierungstool für Mehrebenenkorpora
 - Serverinstallation, lokale Installation
- generisches Datenmodell Salt (Zipser & Romary 2010)
 - Graph basiert (Krause & Zeldes 2014)
- generischen Anfragesprache ANNIS Query Language (AQL)
 - Zwei Schreibweisen: Klauselschreibweise und verkürzte Schreibweise
 - grundsätzlich gilt:
 - Gefunden werden kann nur das, was auch annotiert ist!**
 - kein Annotations- oder NLP-Tool



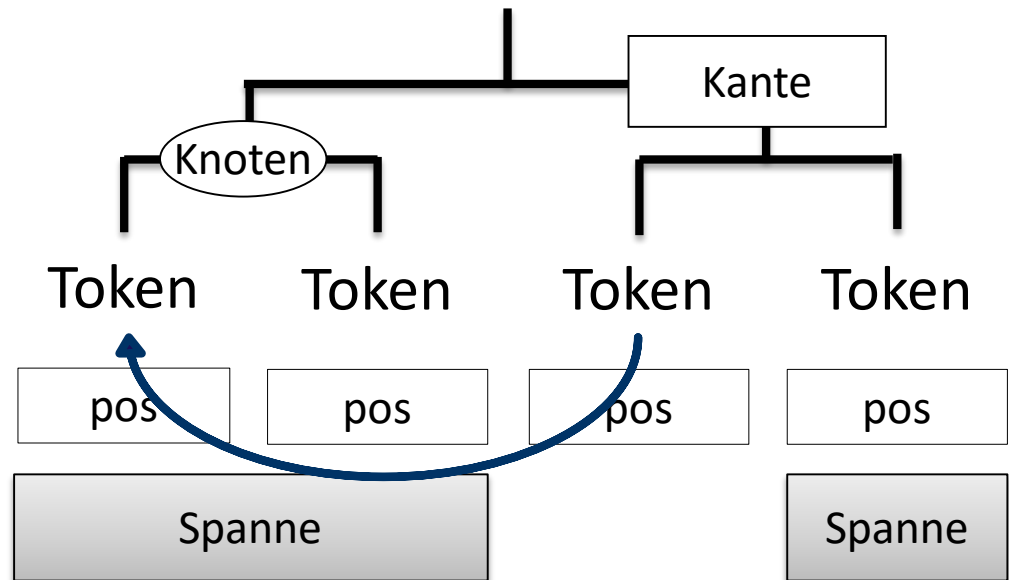
Korpora Annotationen

- Suche und Visualisierung diverser Annotationskonzepte, z.B.
 - Token
 - Spannen
 - Bäume
 - Filterung nach Metadaten
- unabhängig von der Bedeutung der Annotationen (z.B. Tagset)

Formate

- Zur Frage: Wie bekomme ich mein Korpus in ANNIS?
- Konverter Framework Pepper (Zipser & Romary 2010)
 - gemeinsames Datenmodell mit ANNIS → Salt
 - Unterstützung von u.a.

TEI XML,
 MMAX,
 EXMARaLDA,
 ANNIS,
 TIGER XML,
 TCF,
 PAULA





Unser Zugang

- HU – Instanz
 - öffentlicher Zugang: <https://korpling.org/annis3/>
 - kein Login benötigt
- Suche mit
 - Korpusreferenz Märchenkorpus:
https://korpling.org/annis3/#_c=TWFlcmNoZW5rb3JwdXM
 - für alle Referenzlinks letzter Zugriff am 22.02.2016



Beispielkorpus Märchenkorpus

- Texte aus dem 19. Jahrhundert, knapp 300.000 Token
- Historisches Deutsch, Märchen
- Tokenannotationen:
 - Lemmata, Wortarten

[* Show in ANNIS search interface](#)

50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,	die	sprachen	jeden	Tag	„ach
@ord@	Dornröschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,	die	sprechen	jede	Tag	<unknown>
ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$.	PRELS	VFIN	PIAT	NN	ADJD

Trefferreferenz: <https://korpling.org/annis3/?id=fc707243-e26b-4d9f-bcd2-d3a3844c671b>



Token

- 1) Als Token bezeichnet man häufig die **kleinste (technische) Einheit** in einem Korpus.
- 2) Ein Token entspricht oft (aber nicht immer) einem **graphemischen Wort** oder **Satzzeichen**.
- 3) Nach diesen Einheiten kann man in ANNIS **suchen**.

[* Show in ANNIS search interface](#)

50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,	die	sprachen	jeden	Tag	„ach
@ord@	Dornroschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,	die	sprechen	jede	Tag	<unknown>
ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	,\$	PRELS	VFIN	PIAT	NN	ADJD

■ für ein Token



ANNIS INTERFACE



korpling.org/annis3

[About ANNIS](#) | [Report Problem](#) | [Help us to make ANNIS better!](#) | not logged in | [Login](#)

Please enter AQL query

[Query Builder](#)

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

Corpus List | Search Options

Visible:

Name	Texts	Tokens		
a5.hausa.news	4	2.017		
a5.hausa.umarnin.uwa	47	10.194		
abraham.our.father	7	7.671		
AliBaba.6.all	1	202		

[Help/Examples](#)

[Tutorial](#)

[Example Queries](#)

Example Query	Description	open corpus browser
Q /[/j]Jahr/	Search for the "jahr" with upper or lower-case "j" (regular expression)	KAJUK
Q /[Ee]vangelika/	Search for the "evangelika" with upper or lower-case "e" (regular expression)	a5.hausa.news
Q "."	search for the word "."	a5.hausa.umarnin.uwa
Q norm="νοϋτε"	search for the normalized word νοϋτε	abraham.our.father
Q pos="NPROPN"	search for proper names	abraham.our.father
Q /[//]/	Search for the "/" with upper or lower-case "/" (regular expression)	HSJ-Briefe
Q "zu"	search for the word "zu"	HSJ-Maese_Nissima
Q /[Ff]un/	Search for the "fun" with upper or lower-case "f" (regular expression)	HSJ-Maese_Nissima
Q "schwester"	search for the word "schwester"	HSJ-Schevet_Jehude
Q /[Dd]as/	Search for the "das" with upper or lower-case "d" (regular expression)	HSJ-Schevet_Jehude
Q "Aine"	search for the word "Aine"	HSJ-Varia
Q /[Aa]iner/	Search for the "ainer" with upper or lower-case "a" (regular expression)	HSJ-Varia
Q = "	Findet einen Junktor	KAJUK
Q "in"	search for the word "in"	KAJUK



Interface

The screenshot shows the ANNIS web interface. At the top left, there are links for 'About ANNIS' and 'Report Problem'. The main header includes 'Help us to make ANNIS better!' and a 'not logged in' status with a 'Login' button. On the left side, there is a 'Query Builder' section with a text input field for AQL queries and buttons for 'Search', 'More', and 'History'. Below this is a welcome message: 'Welcome to ANNIS! A tutorial is available on the right side.' The 'Corpus List' section shows a table of available corpora. The main content area on the right contains a 'Help/Examples' section with a 'Tutorial' link and a table of 'Example Queries'. Three red arrows point to the 'Tutorial' link, the 'Example Queries' table, and the 'Visible: All' dropdown menu in the corpus list.

Example Query	Description	open corpus browser
/[/j]ahr/	Search for the lower-case "j"	
/[Ee]vangelika/	Search for the or lower-case "e" (regular expression)	a5.hausa.news
".	search for the word "."	a5.hausa.umarnin.uwa
norm="noyɛ"	search for the normalized word noyɛ	abraham.our.father
pos="NPROP"	search for proper names	abraham.our.father
/[//]/	Search for the "/" with upper or lower-case "/" (regular expression)	HSJ-Briefe
"zu"		jima
/[Fflun]/		jima
"schwest"		lude
/[Dd]as/	Search for the "das" with upper or lower-case "d" (regular expression)	HSJ-Schevet Jehude
"Aine"	search for the word "Aine"	HSJ-Varia
/[Aa]iner/	Search for the "ainer" with upper or lower-case "a" (regular expression)	HSJ-Varia
j="j"	Findet einen Junktor	KAJUK
"in"	search for the word "in"	KAJUK

Name	Texts	Tokens		
a5.hausa.news	4	2.017		
a5.hausa.umarnin.uwa	47	10.194		
abraham.our.father	7	7.671		
AliBaba.6.all	1	202		



Tutorial

Help us to make ANNIS better! not lo

Help/Examples

Tutorial

Choose topic ▾ [Print](#)

- ANNIS interface >
- ANNIS Query language >**

- Searching for Word Forms
- Searching for Annotations
- Searching using Regular Expressions
- Searching for Trees
- Searching for Pointing Relations
- Exporting Results
- Frequency Analysis
- Complete List of Operators

ANNIS INTERFAC

Using the ANNIS

The ANNIS interface i
search form and the

The Search Form

```
head_1_pos=/V.  
>dep[func="dobj"]
```

Example Queries

Thema wählen



Suchfenster

Fehler/Fragen
an das
ANNIS-Team

Shortcut Suche:
Strg + Enter

Anzeige:

- a) Anzahl der Treffer
- b) Fehler in der Anfrage



i-Button Metadaten

ANNIS interface showing a list of corpora and a table of example queries. A red arrow points to the 'i' icon in the corpus list, which is labeled "i" Zugriff auf Korpusmetadaten".

Visible: All

Name	Texts	Tokens	i	📄
a5.hausa.news	4	2.017	ⓘ	📄
a5.hausa.umarnin.uwa	47	10.194	ⓘ	📄
abraham.our.father	7	7.671	ⓘ	📄
AliBaba.6.all	1	202	ⓘ	📄

Example Query	Description	open corpus browser
Q /[/j]Jahr/	Search for the "jahr" with upper or lower-case "j" (regular expression)	KAJUK
Q /[/Ee]vangelika/	Search for the "evangelika" with upper or lower-case "e" (regular expression)	a5.hausa.news
Q ". "	search for the word "."	a5.hausa.umarnin.uwa
Q norm="νοϋτε"	search for the normalized word νοϋτε	abraham.our.father
Q pos="NPROP"	search for proper names	abraham.our.father
Q /[/]/	Search for the "/" with upper or lower-case "/" (regular expression)	HSJ-Briefe
Q "zu"	search for the word "zu"	HSJ-Maese_Nissima
Q /[/Ff]un/	Search for the "fun" with upper or lower-case "f" (regular expression)	HSJ-Maese_Nissima
Q "schwester"	Search for the word "schwester"	HSJ-Schevet_Jehude
Q /[/Dd]as	Search for the word "das" with upper or lower-case "d" (regular expression)	HSJ-Schevet_Jehude
Q "Aine"	Search for the word "Aine"	HSJ-Varia
Q /[/Aa]liner/	Search for the "ainer" with upper or lower-case "a" (regular expression)	HSJ-Varia
Q = "	Findet einen Junktor	KAJUK
Q "in"	search for the word "in"	KAJUK



i-Button Korpusmetadaten

Corpus information for RIDGES_Herbology_Version4.1 (ID: 11636)

Metadata		Available annotations	
Select corpus/document:	RIDGES_Herbology_Version4.1	Node Annotations	
Name	Value	Name	Example (click to use query)
Homepage	http://korpling.german.hu-berlin.de/ridges/index_en.html	atLeast	atLeast="1,000000"
annotators	Abed-Ali, Ilham; Andresen, Silke; Ast, Henriette; Belz, Malte; Christen, Doreen; Dayal, Mascha; Dittberner, Antonia; Döhn, Cora; Driemel, Imke; Efremova, Olja; Eichhorn, Gill-Maria; Esser, Judith; Gerlach, Annegret; Giesel, Linda; Kiraga, Sebastian; Kolbik, Ewa Anna; Kovács, Kornél; Krüger, Daisi; Lehmann, Anna-Maria; Lober, Maria; Lueders, Laura; Lüdelling, Anke, Maniscalco, Samuele; Metzsig, Manuel; Meyer, Alexander; Müller, Sandra; Müller, Vinzent; Murphy, Andrew; Mursell, Johannes; Okuda, Akiko; Perlitz, Laura; Reinig, Katharina; Riesler, Ina; Rosin, Lena; Sachse, Franz-Josef; Sapronova, Anna; Sauer, Simon; Schmidt, Claudia; Sorokovska, Iryna; Springmann, Uwe; Stephan, Kristina; Tiemann, Juliane; Tóth, Anna; Tóth, Réka; Turtureanu, Alexander; Wekel, Juliana; Zuchewicz, Karolina et al.	atMost	atMost="2,000000"
projectDesc	RIDGES Project (Register In Diachronic German Science), funded by the Google Digital Humanities Research Award	attr_gen	attr_gen="gpost"
respStmt	Compiled by Thomas Krause, Anke Lüdelling, Carolin Odebrecht, Amir Zeldes. The RIDGES team consists of Anke Lüdelling, Amir Zeldes, Carolin Odebrecht, Vivian Voigt and Laura Perlitz.	author_ref	author_ref="pron1sg"
sponsor	Google Inc.	bemerkung	bemerkung="grammatisch stellt es sich als verbalsubstantiv mit (idg.) ti-suffix zu got. siukan 'krank sein'. "
version	4.1	brace	brace="brRight"
		brace_dir	brace_dir="left"
		clean	clean="/"
		definition	definition="fig"
		dipl	dipl="/"
		disease	disease="di"
		div1	div1="div1"
		div1_tvne	div1_tvne="div1_tvne"
		Edge Annotations	
		Edge Types	
		Meta Annotations	

Link to corpus: https://korpling.org/annis3/#_c=UkIER0VTX0hlcmljYvG9neV9WZXJzaW9uNC4x

Korpusreferenzlink



i-Button

Verfügbare Annotationen

The screenshot shows the ANNIS interface with a search query `attr_gen="gpost"` in the search bar. A red arrow points from this query to the `attr_gen` row in the 'Available annotations' table. The table lists various annotations with their names, example values, and URLs.

Name	Example (click to use query)	URL
atLeast	atLeast="1,000000"	
atMost	atMost="2,000000"	
attr_gen	attr_gen="gpost"	
author_ref	author_ref="pron1sg"	
bemerkung	bemerkung="grammatisch stellt es sich als verbalsubstantiv mit (idg.) ti-suffix zu got. siukan 'krank sein'."	
brace	brace="brRight"	
brace_dir	brace_dir="left"	
clean	clean="/"	
definition	definition="fig"	
dipl	dipl="/"	
disease	disease="di"	
div1	div1="div1"	
div1_tvne	div1_tvne="book"	

Link to corpus: https://korpling.org/annis3/#_c=UkIER0VTX0hlcjvjbG9neV9WZXJzaW9uNC4x

Annotationsebenen und Beispiele für Werte. Doppelklick fügt das Beispiel in die Suchmaske ein.



i-Button

Dokumentmetadaten

Info for salt:/RIDGES_Herbology_Version4.1/Alchemisti... + ×

Metadata

document: AlchemistischePraktik_1603

Name	Value
annis:doc	AlchemistischePraktik_1603
default_ns:author	Andreas Libavius
default_ns:bibl	Libavius, Andreas (1603) Alchimistische Praktik. Frankfurt am Main. Johann Saur. 4-26.
default_ns:date	1603
default_ns:pubPlace	Frankfurt
default_ns:publisher	Johann Saur
default_ns:title	Alchimistische Praktik
default_ns:version	4.1

corpus: RIDGES_Herbology_Version4.1



History

The screenshot shows the ANNIS web interface. At the top left, there is a search input field containing the query `attr_gen="gpost"`. Below the search field, there are buttons for 'Search', 'More', and 'History'. A red arrow points to the 'History' button. A dropdown menu is open below the 'History' button, showing the query `attr_gen="gpost"` and a 'Show more details' button. To the right of the search field, there is a 'Query Builder' button and a keyboard icon. The main content area displays search results for the query. The first result is for the document 'AlchemistischePraktik_1603 (dipl 206 - 218)' with the text 'alfo nicht dørffen fleißig der heimlichkeit der natur nachforchen . Wolten lieber /'. The second result is for 'AlchemistischePraktik_1603 (dipl 478 - 490)' with the text 'din gen erftlichen die namen der Glæfer / Infrumenten / Ofen vnd'. The interface also shows a 'Corpus List' section with 'Visible: Demo-Hamburg' and a 'Filter' section with 'Name' and 'Texts' columns. The bottom of the interface shows a table with columns for document name, ID, and other details.

History-Abfrage:
Liste aller Abfragen einer Sitzung



Interface

Anfrage Optionen

The screenshot shows the ANNIS search interface. At the top, there are navigation links for 'About ANNIS' and 'Report Problem'. The main search area contains the query `attr_gen="gpost"`. Below the search bar, there are buttons for 'Search', 'More', and 'History'. The results section shows '1489 matches in 29 documents'. The 'Search Options' tab is active, displaying several dropdown menus: 'Left Context' (5), 'Right Context' (5), 'Show context in' (dipl), 'Results Per Page' (10), and 'Order' (Ascending). Red arrows point from the 'Show context in' dropdown to a red box containing the question 'Welche Transkriptionsebene soll in der Trefferliste angezeigt werden?' and from the 'Left Context' dropdown to another red box containing the text 'Größe des Kontexts festlegen'.

Größe des Kontexts festlegen

Welche Transkriptionsebene soll in der Trefferliste angezeigt werden?



Trefferkonkordanz Märchenkorpus

The screenshot displays the ANNIS search interface. On the left, a sidebar shows a corpus list with 'Märchenkorpus' selected, indicating 211 texts and 295,880 tokens. The main search results area shows a list of 737 matches across 70 documents. The top result is highlighted, showing the path 'Maerchenkorpus > grimms_allerleirauh_353-359' and the text 'Es war einmal ein König, der hatte eine Frau'. The word 'König' is highlighted in red. Red arrows point to various UI elements with labels:

- i-Button für Dokumentmetadaten**: Points to the 'i' icon in the first result row.
- Suchreferenzlink**: Points to the search icon in the top right of the result row.
- Dokumentname**: Points to the path 'Path: Maerchenkorpus > grimms_allerleirauh_353-359'.
- Trefferanzahl**: Points to the text '737 matches in 70 documents'.
- Trefferreferenzlink**: Points to the search icon in the third result row.
- Treffer**: Points to the word 'König' in the third result row.



ANNIS QUERY LANGUAGE



AQL

ANNIS Query Language

- Prinzip I
 - Attribut-Wert-Paar
 - Prinzip II
 - Relationen zwischen Attribut-Wert-Paaren
- Gilt für alle Annotationsarten und Korpora in ANNIS!



Prinzip I

Attribut-Wert-Paar

- 1) Voraussetzung ist das **Vorhandensein einer Ebene namens „tok“**. (Metadaten!)
- 2) Erwartetes Ergebnis ist es, **exakt alle Vorkommen** dieser Zeichenkette in „tok“ im ausgewählten Korpus zu finden.

tok= /König/

Suchreferenz: <https://korpling.org/annis3/?id=963dcfeb-be31-43dd-887e-fb5c51a90edd>

Attribut

(Layer, Tier, Ebene ...)

Wert

(Wort, Lemma, Satz, Wortart ...)

tok	50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,
lemma	@ord@	Dornröschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,
pos	ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$,



Prinzip I

Attribut-Wert-Paar

- 1) Voraussetzung ist das **Vorhandensein einer Ebene namens „pos“**. (Metadaten!)
- 2) Erwartetes Ergebnis ist es, **exakt alle Vorkommen** dieser Zeichenkette in „pos“ im ausgewählten Korpus zu finden.

pos= /NN/

Suchreferenz: <https://korpling.org/annis3/?id=f41c2ede-4b22-40c1-962b-ca655a7bcd39>

Attribut

(Layer, Tier, Ebene ...)

Wert

(Wort, Lemma, Satz, Wortart ...)

tok	50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,
lemma	@ord@	Dornröschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,
pos	ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$,



Prinzip I

Attribut-Wert-Paar

- 1) Voraussetzung ist das **Vorhandensein einer Ebene namens „lemma“**. (Metadaten!)
- 2) Erwartetes Ergebnis ist es, **exakt alle Vorkommen** dieser Zeichenkette in „lemma“ im ausgewählten Korpus zu finden.

lemma=**/sein/**

Suchreferenz: <https://korpling.org/annis3/?id=9f527f76-dfc1-4561-99f5-c4126d599801>

Attribut

(Layer, Tier, Ebene ...)

Wert

(Wort, Lemma, Satz, Wortart ...)

tok	50.	Dornröschen	.	Vor	Zeiten	war	ein	König	und	eine	Königin	,
lemma	@ord@	Dornröschen	.	vor	Zeit	sein	eine	König	und	eine	Königin	,
pos	ADJA	NN	\$.	APPR	NN	VAFIN	ART	NN	KON	ART	NN	\$,



EXKURS: REFERENZIERUNG IN ANNIS



Korpusreferenz

- Aufrufen von ANNIS + ein Korpus ist ausgewählt
 - **Märchenkorpus in ANNIS**
 - https://korpling.org/annis3/#_c=TWFlcmNoZW5rb3JwdXM
 - Korpusmetadaten

Metadata		Available annotations		
Select corpus/document: Maerchenkorpus		Node Annotations		
Name	Value	Name	Example (click to use query)	URL
Herausgeber	Maik Walter	lemma	lemma="-<unknown>"	
Kontakt	walter@textbewegung.de	pos	pos="NN"	
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 Brüdern Grimm herausgegebenen 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, Fächdiskurse & Labore).			
Titel	Märchenkorpus			
		Edge Annotations		
		Erichs Types		
		Meta Annotations		
Link to corpus: https://korpling.org/annis3/#_c=TWFlcmNoZW5rb3JwdXM				



Suchreferenz

- Aufrufen von ANNIS + ein Korpus ist ausgewählt + eine Anfrage ist gestellt
 - **Märchenkorpus in ANNIS + Suche tok=/König/**
 - <https://korpling.org/annis3/?id=963dcfeb-be31-43dd-887e-fb5c51a90edd>

Base text ▾ Token Annotations ▾

1 2 / 74 > >| Displaying Results 1 - 10 of 737 Result for: tok=/König/

1 ⓘ ↻ Path: Maerchenkorpus > grimm_allerleirauh_353-359 left context: 5 ▾ right context: 5 ▾ ^
(tokens 3 - 13)

. Es war einmal ein König , der hatte eine Frau
. es sein einmal eine König , die haben eine Frau
\$. PPER VAFIN ADV ART NN \$, PRELS VAFIN ART NN

2 ⓘ ↻ Path: Maerchenkorpus > grimm_allerleirauh_353-359 left context: 5 ▾ right context: 5 ▾
(tokens 50 - 60)

würde , rief sie den König und sprach „wenn du nach
werden , rufen sie die König und sprechen <unknown> du nach
VAFIN \$, VAFIN PPER ART NN KON VAFIN ADJD PPER APPR



Trefferreferenz

- Aufrufen von ANNIS + ein Korpus ist ausgewählt + eine Anfrage ist gestellt
 - **Märchenkorpus in ANNIS + Suche tok=/König/ + 1.Treffer**
 - <https://korpling.org/annis3/?id=761cd050-e32a-4bec-a3c7-5f23abb592ad>

Base text ▾ Token Annotations ▾

1 / 74 Displaying Results 1 - 10 of 737 Result for: tok=/König/

1 Path: Märchenkorpus > grimm_allerleirauh_353-359 left context: 5 right context: 5 (tokens 3 - 13)

. Es war einmal ein **König**, der hatte eine Frau
. es sein einmal eine König, die haben eine Frau
\$. PPER VAFIN ADV ART NN \$, PRELS VAFIN ART NN

2 Path: Maerchenkorpus > grimm_allerleirauh_353-359 left context: 5 right context: 5 (tokens 50 - 60)

würde, rief sie den **König** und sprach „wenn du nach
werden, rufen sie die König und sprechen <unknown> du nach
VAFIN \$, VFIN PPER ART NN KON VFIN ADJD PPER APPR



Trefferreferenz Einstellung

Match reference link

Share your match: 1. Choose the visualization to share. 2. Copy the generated link or code. 3. Share this link with your peers or include the code in your website.

Select visualization

Display Name
kwic

Link for publications

<https://korpling.org/annis3/?id=761cd050-e32a-4bec-a3c7-5f23abb592ad>

Code for embedding visualization into web page

```
<iframe height="300px" width="100%" src="https://korpling.org/annis3/?id=761cd050-e32a-4bec-a3c7-5f23abb592ad"></iframe>
```

Preview

[Show in ANNIS search interface](#)

. Es war einmal ein König , der hatte eine
. es sein einmal eine König , die haben eine
↑ DDFD VAFIN ADV ART MN # DDFD VAFIN ART

Close

Auswahl der
Annotationsanzeige

Links für
Publikationen und
Homepages

Vorschau



Trefferreferenz Ausführen

Link zur Suchanfrage

[* Show in ANNIS search interface](#)

. Es war einmal ein König , der hatte eine Frau
. es sein einmal eine König , die haben eine Frau
\$. PPER VAFIN ADV ART NN \$, PRELS VAFIN ART NN



ZURÜCK ZU AQL



Prinzip II

Relationen

- Suchen Sie alle Vorkommen des Lemmas *was*, das als Relativpronomen verwendet wird!



Prinzip II

Relationen

- Verknüpfung von zwei Attribut-Wert-Paaren
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Wissen, welche Annotationsebenen welche Art der Beziehung zwischen einander besitzen können!
 - **Es ist immer eine Verbindung zwischen AW-Paaren notwendig!**
 - heute verkürzte Schreibweise

tok	fragte	er	die	beiden	Stieftöchter	was	er	ihnen	mitbringen	sollte	?
lemma	fragen	er	die	beide	Stieftochter	was	er	sie	mitbringen	sollen	?
pos	VVFIN	PPER	ART	PIAT	NN	PRELS	PPER	PPER	VVINF	VMFIN	\$.

tok	was	sollen	wir	euch	geben	?
lemma	was	sollen	wir	ihr	geben	?
pos	PWS	VMFIN	PPER	PPER	VVINF	\$.

Trefferreferenz: <https://korpling.org/annis3/?id=5ae7932e-f462-4dc3-b2ff-6dcf155b6347>

Trefferreferenz: <https://korpling.org/annis3/?id=7a297068-b6c2-439d-a37a-546909bae750>



Prinzip II

Relationen

- Suchen Sie alle Vorkommen des Lemmas *was*, das als Relativpronomen verwendet wird!

lemma=/was/

AW-Paar #1

_ = _

Identische Überlappung

pos=/PRELS/

AW-Paar #2

- 290 Treffer in 122 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=c90bdd98-f5aa-437a-87bc-ff913eca0b01>



AUFGABE

- Suchen Sie nach nominalen Phrasen in Form von Wortartenabfolgen.
 - NP: Artikel attributives Adjektiv Nomen
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?

pos → ART ADJA NN



AUFGABE

- Verknüpfung von zwei Attribut-Wert-Paaren
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Wissen, welche Annotationsebenen welche Art der Beziehung zwischen einander besitzen können!
 - **Es ist immer eine Verbindung zwischen AW-Paaren notwendig!**
 - heute verkürzte Schreibweise



tok	Es	war	aber	keine	in	der	ganzen	Welt	zu	finden	,
lemma	es	sein	aber	keine	in	die	ganz	Welt	zu	finden	,
pos	PPER	VAFIN	ADV	PIAT	APPR	ART	ADJA	NN	PTKZU	VVINF	\$,

Trefferreferenz: <https://korpling.org/annis3/?id=57cbef66-7f79-4214-938f-aaf648160b27>



AUFGABE

- Suchen Sie nach nominalen Phrasen in Form von Wortartenabfolgen.
 - NP: Artikel attributives Adjektiv Nomen
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?

pos=/ART/

AW-Paar #1

.

Direkte Präzedenz

pos=/ADJA/

AW-Paar #2

.

Direkte Präzedenz

pos=/NN/

AW-Paar #3

- 3489 Treffer in 204 Dokumenten
- Bedingung: Token präzedieren Token. Präzedenz zwischen allen anderen Ebenen (hier pos) über die Token!
- AW-Paar #1 und #3 sind nur indirekt verbunden

Suchreferenz: <https://korpling.org/annis3/?id=16c38b81-58b2-4092-8d20-a939e57db92a>

Syntax Highlighting

- AW-Paare in verschiedenen Farben
 - immer gleiche Reihenfolge – Nummer der AW-Paare, wie die Farben vergeben werden (z.B. erst rot, dann lila, dann grün)
 - Treffer dann genau in diesen Farben
 - unabhängig von der Schreibweise, den Annotationsebenen und Operatoren

The screenshot shows the ANNIS (Annotation-based Network for Information Systems) interface. On the left, a search query is entered: `pos=/ART/ . pos=/ADJA/ . pos=/NN/`. Below the query, it indicates "3489 matches in 204 documents". The main area displays search results for the path "Maerchenkorpus > grimm_allerleirauh_353-359". The results are shown in two rows, each with a table of tokens and their corresponding POS tags. The tokens are highlighted in colors corresponding to the query: red for ART, purple for ADJA, and green for NN.

Token	POS
und	KON
dachte	VFIN
nicht	PTKNEG
daran	PAV
,	\$
eine	ART
zweite	ADJA
Frau	NN
zu	PTKZU
nehmen	VVINF
.	\$.
Endlich	ADV
sprachen	VFIN

Token	POS
suchen	VVINF
,	\$.
die	PRELS
an	APPR
Schönheit	NN
der	ART
verstorbenen	ADJA
Königin	NN
ganz	ADV
gleich	ADJD
käme	VVFIN
.	\$.
Es	PPER



ANNIS-Tutorials

- erste Tutorialvideos zum
 - Interface
 - Annotationsarten und -anzeigen
 - Suche nach Wortartenannotation („PoS“)
 - unter <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/corpus-tools/annis-tutorials>
- Weitere Videos sind geplant!
- User-Dokumentation: [http://corpus-tools.org/annis/resources/ANNIS User Guide 3.3.5.pdf](http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.3.5.pdf)
- Wünsche und Anmerkungen an korpling@hu-berlin.de

Zusammenfassung

- Heute im Workshop erarbeitet
 - **Allgemeines**
 - Grundbegriffe aus der Korpuslinguistik
 - Grundfragen der Korpuslinguistik
 - Immer wieder neu hinterfragen und motivieren!
 - **Korpuserstellung**
 - Textauswahl, Tokenisierung, Annotation
 - Problematisierungen und erste Vorgehensweisen
 - Erste Schritte um eigene Korpora zu erstellen oder Korpora dritter zu verstehen!
 - **Suche nach und Analyse von Annotationen**
 - Anfragesprache
 - Abbildungen von Korpus und Annotationskonzept
 - Erste Schritte korpuslinguistischer Analyse und Auswertung eigener oder Korpora dritter!
- Materialien, Tools und (fast alle) Korpora sind alle frei verfügbar!

Kooperation

- unterstützt durch

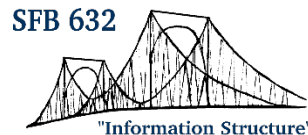
Humboldt-Universität zu Berlin



Georgetown University Washington



Sonderforschungsbereich 632



Deutsche Forschungsgemeinschaft



und viele weitere!

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT!



Referenzen

Links

Referenzierung durch Links

- Alle Links in der Präsentation wie im Anhang sind, wenn nicht anders angegeben, geprüft am 22.02.2016!
 - Korpusreferenzlinks
 - Suchreferenzlinks
 - Trefferreferenzlinks
 - Referenzen für Software und Korpora



ANNIS

Tutorialvideos



ANNIS-Tutorials

- erste Tutorialvideos zum
 - Interface
 - Annotationsarten und -anzeigen
 - Suche nach Wortartenannotation („PoS“)
 - unter <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/corpus-tools/annis-tutorials>
- Weitere Videos sind geplant!
- User-Dokumentation: [http://corpus-tools.org/annis/resources/ANNIS User Guide 3.3.5.pdf](http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.3.5.pdf)
- Wünsche und Anmerkungen an korpling@hu-berlin.de



ANNIS

Suchaufgaben



AUFGABE 1

- Suchen Sie nach allen Vorkommen der Wortform *Sohn!*
- Attribut: tok
- Wert: Sohn



AUFGABE 1

- Suchen Sie nach allen Vorkommen der Wortform *Sohn*!
- Attribut: tok
 - Wert: Sohn

tok =/Sohn/

Ergebnis: 155 Treffer in 46 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=365e147f-001b-44fc-a0b1-0da424b6b0f2>



AUFGABE 2

- Suchen Sie nach allen Vorkommen des Lemmas *Sohn!*
- Was erwarten Sie zu finden, im Vergleich zur ersten Suche in Aufgabe 1?



AUFGABE 2

- Suchen Sie nach allen Vorkommen des Lemmas *Sohn!*
- Was erwarten Sie zu finden, im Vergleich zur ersten Suche in Aufgabe 1?
 - Attribut: lemma
 - Wert: Sohn

lemma =/Sohn/

Ergebnis: 191 Treffer in 55 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=db6a0017-b8f3-40cd-8d32-029a666a71ca>



Operatoren für die Mustersuche

- . Ein beliebiges Zeichen
- ? 0 oder 1 Zeichen (des vorherigen Elementes)
- * 0 bis unendlich viele Zeichen (d. vorh. E.)
- + 1 bis unendlich viele Zeichen (d. vorh. E.)
- \\ wörtlich (folgendes Zeichen)
- ! nicht
- (a | b) a oder b (auch: [ab])



Operatoren Joker .

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



Operatoren ? und * +

da[↷]s?

das vorherige Zeichen ist optional

→ ϕ , s → *da, das*

da[↷]s*

das vorh. Zeichen kommt 0- bis ∞ mal vor

→ ϕ , s, ss, ... → *da, das, dass, dassssss*

da[↷]s+

das vorh. Zeichen kommt 1- bis ∞ mal vor

→ s, ss, ... → *das, dass, dassssss*



AUFGABE 3

- Welche Treffer erwarten Sie bei folgender Mustersuche?
- Welche Funktion besitzt der `.` ?
 - Attribut: tok
 - Wert: g.b.

`tok=/g.b./`



AUFGABE 3

- Welche Treffer erwarten Sie bei folgender Mustersuche?
- Welche Funktion besitzt der `.` ?
 - Attribut: tok
 - Wert: g.b.

tok=/g.b./

Treffer: *gibt, gebe, gäbe, gebs*

Ergebnis: 72 Treffer in 46 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=5549c4b5-37a0-45cb-b346-07b65b3309d8>



AUFGABE 4

- Welche Treffer erwarten Sie für folgende Mustersuchen?
 - tok=/de.en/
 - tok=/(S|H)ohn/
 - tok=/ihren?/



AUFGABE 4

- Welche Treffer erwarten Sie für folgende Mustersuchen?
 - tok=/de.en/
 - <https://korpling.org/annis3/?id=effbe27e-b504-42df-a812-6249a819c8ee>
 - *deren, denen, dehen, deden*
 - tok=/(S|H)ohn/
 - <https://korpling.org/annis3/?id=7da860dc-3744-45c3-869c-ffec0f72aa6a>
 - *Sohn, Hohn*
 - tok=/ihren?/
 - <https://korpling.org/annis3/?id=52804fa1-9d3c-44c3-ad02-1383b797e9cc>
 - *ihren, ihre*



AUFGABE 5

- Suchen Sie alle Wortformen, die auf *mann* enden!
- Welche Operatoren werden benötigt?
- Welche Treffer erwarten Sie?



AUFGABE 5

- Suchen Sie alle Wortformen, die auf *mann* enden!
- Welche Operatoren werden benötigt?
- Welche Treffer erwarten Sie?
- Attribut: tok

tok = /.+mann/

Treffer: *jedermann, Bettelmann, Spielmann, Kaufmann* etc.

Ergebnis: 187 Treffer in 46 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=590148a7-5fb1-4896-b826-a9879737d813>



AUFGABE 6

- Suchen Sie alle Wortformen, die mit *Wunder* anfangen!
- Welche Operatoren werden benötigt?
- Welche Treffer erwarten Sie?



AUFGABE 6

- Suchen Sie alle Wortformen, die mit *Wunder* anfangen!
- Welche Operatoren werden benötigt?
- Welche Treffer erwarten Sie?
- Attribut: tok

tok = /Wunder.+/

Treffer: *Wunderkraft, Wundergarten, Wunderdinge, etc.*

Ergebnis: 15 Treffer in 12 Dokumenten

Suchreferenz: <https://korpling.org/annis3/?id=a7340ba2-4917-4828-9a76-86b3956cf621>



AUFGABE 7

- Suchen nach Alternativen!
- Welche Treffer erwarten Sie?
- Wie unterscheiden sich diese Mustersuchen?

tok = /(gut|schlecht)/

tok=/(guter|gutes)/

tok=/bes(ser|t)/



AUFGABE 7

- Suchen nach Alternativen!
- Welche Treffer erwarten Sie?
- Wie unterscheiden sich diese Mustersuchen?

tok = /(gut|schlecht)/

tok=/(guter|gutes)/

tok=/bes(ser|t)/

- Erste Suche nach alternativen Wörtern.
- Zweite Suche nach alternativen Wortformen.
- Dritte Suche nach alternativen Zeichen.

Suchreferenz: <https://korpling.org/annis3/?id=aa1a4442-2f8e-40c0-baf4-2e97b7d674bd>

Suchreferenz: <https://korpling.org/annis3/?id=731c39db-6aa9-46f8-83bd-74b873d843b9>

Suchreferenz: <https://korpling.org/annis3/?id=85a6a67f-2e54-492d-b5c4-4a6a98deac5c>



AUFGABE 8 a

- Finden Sie alle Vorkommen der konjugierten Formen des Verbs *meinen* im Präsens, aber keine anderen!
- Welches Attribut benötigen Sie?
- Welche Werten wollen Sie finden und welche nicht?



AUFGABE 8 a

- Finden Sie alle Vorkommen der konjugierten Formen des Verbs *meinen* im Präsens, aber keine anderen!
- Welches Attribut benötigen Sie?
- Welche Werten wollen Sie finden und welche nicht?

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

- Findet auch Infinitiv und Formen des Possessivpronomens.
- Informationen aus anderen Annotationsebenen benötigt!

Suchreferenz: <https://korpling.org/annis3/?id=69bd64df-9435-4dc2-9ae3-6c778261a1d1>



AUFGABE 8 b

- Verknüpfung von zwei Attribut-Wert-Paaren
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Wissen, welche Annotationsebenen welche Art der Beziehung zwischen einander besitzen können!
 - **Es ist immer eine Verbindung zwischen AW-Paaren notwendig!**
 - heute verkürzte Schreibweise
 - für Aufgabe 8 a:

tok	ich	bringe	damit	meinen	Vater	von	seinen	bösen	Gedanken
lemma	ich	bringen	damit	mein	Vater	von	sein	böse	Gedanke
pos	PPER	VFIN	PAV	PPOSAT	NN	APPR	PPOSAT	ADJA	NN

tok	was	meinst	du	damit	?
lemma	was	meinen	du	damit	?
pos	PWS	VFIN	PPER	PAV	\$.

Trefferreferenz: <https://korpling.org/annis3/?id=49f993df-6920-4d5b-8e7b-06b96b450352>

Trefferreferenz: <https://korpling.org/annis3/?id=d9e7763b-b1fc-4277-a8de-f089714e8f12>



AUFGABE 8 b

- Verknüpfung von zwei Attribut-Wert-Paaren
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Wissen, welche Annotationsebenen welche Art der Beziehung zwischen einander besitzen können!
 - **Es ist immer eine Verbindung zwischen AW-Paaren notwendig!**
 - heute verkürzte Schreibweise
 - für Aufgabe 8 a:

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

AW-Paar #1

_ = _

identische Überlappung

pos=/VVFIN/

AW-Paar #2

Suchreferenz: <https://korpling.org/annis3/?id=7e7eea18-f493-482c-b3b6-9cf6ae2b700f>



AUFGABE 9

- Suchen Sie alle Vorkommen des Lemmas *was*, das **nicht** als Relativpronomen verwendet wird!



AUFGABE 9

- Suchen Sie alle Vorkommen des Lemmas *was*, das **nicht** als Relativpronomen verwendet wird!

lemma=/was/

AW-Paar #1

=

identische Überlappung

pos!=/PRELS/

AW-paar #2

- 592 Treffer in 161 Dokumenten
- Negation durch !

Suchreferenz: <https://korpling.org/annis3/?id=6edda9d4-497d-4e87-b417-17917e0a6a04>



AUFGABE 10

- Suchen Sie nach allen Vorkommen der Wortform *schöne*, direkt gefolgt von einem Nomen!
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?



AUFGABE 10

- Suchen Sie nach allen Vorkommen der Wortform *schöne*, direkt gefolgt von einem Nomen!
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?

tok=/schöne/

AW-Paar #1

.

Direkte Präzedenz

pos=/NN/

AW-Paar #2

- 94 Treffer in 54 Dokumenten
- Noch keinen Überblick, welche Nomen das Attribut *schöne* erhalten haben!

Suchreferenz: <https://korpling.org/annis3/?id=8fb795e5-0454-4cf0-8abb-ff559387a4da>



AUFGABE 11

- Suchen Sie nach allen Vorkommen der Wortform *schöne*, direkt gefolgt von einem Nomen in dem Märchen Dornröschen!
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?



AUFGABE 11

- Suchen Sie nach allen Vorkommen der Wortform *schöne*, direkt gefolgt von einem Nomen in dem Märchen Dornröschen!
 - Welche Annotationen helfen bei dieser Aufgabe?
 - Welche Art Operator benötigen Sie?

tok=/schöne/

AW-Paar #1

.

Direkte Präzedenz

pos=/NN/

AW-Paar #2

&

Verknüpfung

meta::Titel=/Dornroeschen/ Metadatum

- 2 Treffer in 1 Dokumenten
- Metadaten in AQL: **meta::** müssen nicht relatiert, aber verknüpft (mit **&**) werden!

Suchreferenz: <https://korpling.org/annis3/?id=a9acbac6-c9a8-4a93-83d8-bfc21306a741>

Übersicht Operatoren für die Mustersuche

- `.` Ein beliebiges Zeichen
- `?` 0 oder 1 Zeichen (des vorherigen Elementes)
- `*` 0 bis unendlich viele Zeichen (d. vorh. E.)
- `+` 1 bis unendlich viele Zeichen (d. vorh. E.)
- `\\` wörtlich (folgendes Zeichen)
- `!` nicht
- `[abc]` Menge (oder `[^abc]`=alles außer abc)
- `(a|b)` a oder b (auch: `[ab]`)
- `a{2,3}` a 2 bis 3mal

Übersicht Operatoren für die Relationen zwischen AW-Paaren

- **&** verbindet Suchanfragen
- #1 **.** #2 direkte Präzedenz
- #1 **.*** #2 indirekte Präzedenz
- #1 **.3,5** #2 #2 folgt mit 3 bis 5 Einheiten Abstand
- #1 **_=_** #2 #1 und #2 identische Überlappung
- #1 **_o_** #2 #1 und #2 überlappen sich
- #1 **_i_** #2 #1 inkludiert #2

Korpuserstellung

Automatisches Lemmatisieren und Taggen

Wortartentagging & Lemmatisierung

Selber taggen

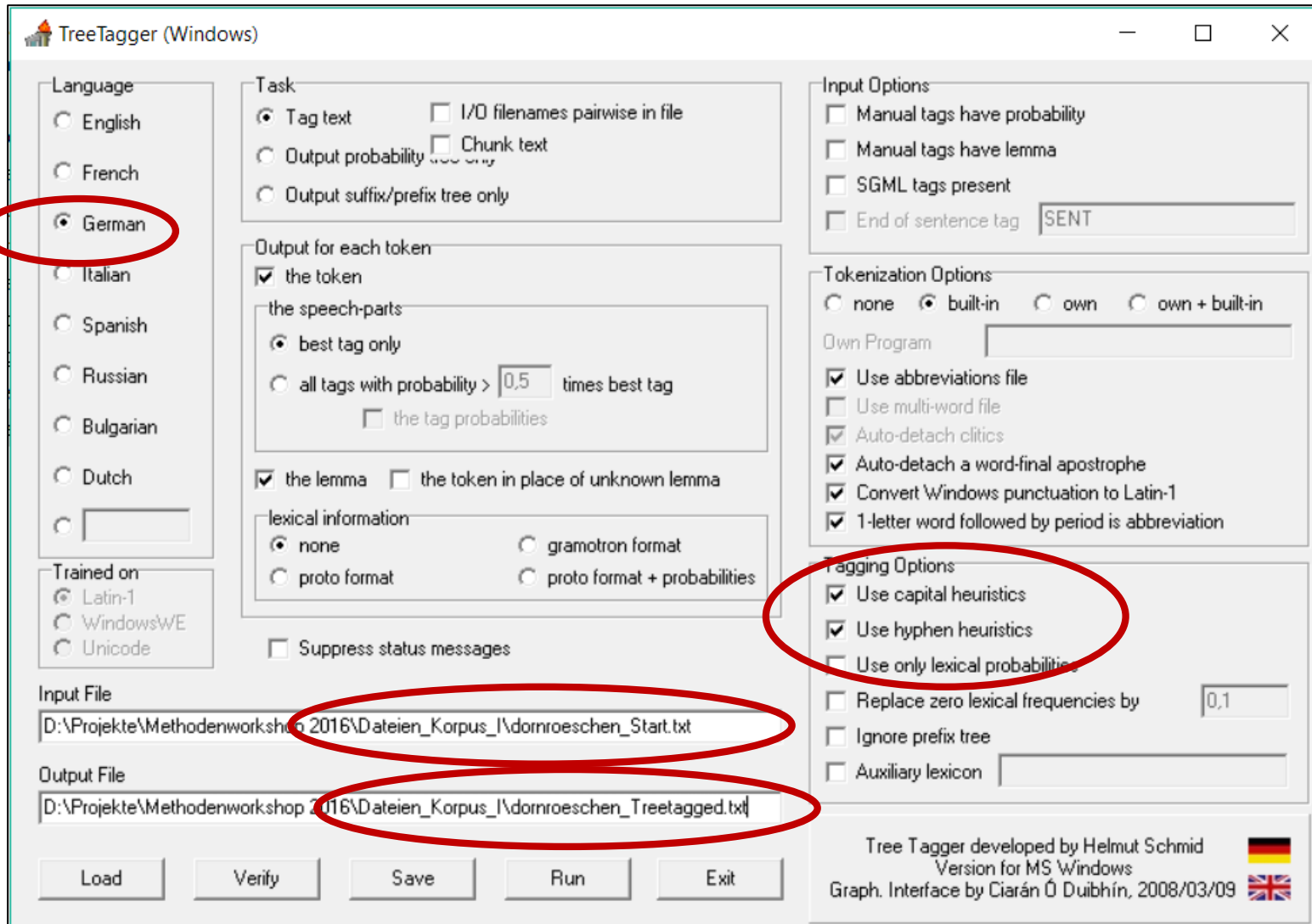
- STTS-Guidelines
<http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>
- Treetagger
<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- Graphisches Interface
<http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/wintreetagger.exe>
- Sie brauchen als Grundlage:
 - Textdatei (.txt; ANSI-Kodierung)
- Weiterverarbeitung:
 - z.B. EXMARaLDA (www.exmaralda.org;
besitzt Treetagger-Importer, über „File > Import ...“ im Partitur Editor)

Wortartentagging & Lemmatisierung

Selber (tree-) taggen

- Der einfachste Weg zur TreeTagger-Installation ist der folgende:
 - zip-Datei runterladen
 - Inhalt des zip-Ordners in den Ordner C:\Program Files\TreeTagger kopieren (die Ordnerstruktur müssen Sie so anlegen).
 - Im "bin"-Ordner können Sie jetzt das Programm wintreetagger.exe starten.
 - Im Interface "German" auswählen, am besten rechts unten "Use capital heuristics" und "Use hyphen heuristics" auswählen.
 - unter "Input File" ins Weiße klicken: Rohtext (beliebiger fortlaufender Text in ANSI-Kodierung) auswählen.
 - unter "Output File" ins Weiße klicken: Zieldatei definieren (diese muss noch nicht existieren). Geben Sie die Dateiendung .txt mit an oder fügen Sie diese später der erzeugten Datei hinzu.
 - "Run"

Tree-Taggen mit FrauHolle_Start.txt



Referenzen

Literatur, Software, Korpora

Einführungen /Handbücher Korpuslinguistik

- **Kuebler, Sandra; Zinsmeister, Heike** (2015): Corpus Linguistics and Linguistically Annotated Corpora. London: Bloomsbury Academic.
- **Lüdeling, Anke; Kytö, Merja** (2009) (Hrsg.): Corpus Linguistics. An International Handbook. Vol 2. (Reihe Handbücher zur Sprach- und Kommunikationswissenschaft) Berlin; Mouton de Gruyter.
- **Lemnitzer, Lothar; Zinsmeister, Heike** (2006): Korpuslinguistik – Eine Einführung Tübingen; Gunter Narr Verlag.

Referenzen

- Evert, Stefan; Fitschen, Arne** (2001): Textkorpora. In: Carstensen et al. (Hrsg) *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Spektrum Akademischer Verlag, Heidelberg, 369 – 376.
- Krause, Thomas; Zeldes, Amir** (2014): ANNIS3: A new architecture for generic corpus query and visualization. in: Digital Scholarship in the Humanities 2014 <http://dsh.oxfordjournals.org/cgi/content/abstract/fqu057?ijkey=GJBr0LhNfKW1g8i&keytype=ref> [letzter Zugriff: 20.10.15]
- Kuebler, Sandra; Zinsmeister, Heike** (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Academic.
- Leech, Geoffrey** (1993): Corpus Annotation Schemes. *Literary and Linguistic Computing* 8(4), 275–281.
- Lemnitzer, Lothar; Zinsmeister, Heike** (2006): *Korpuslinguistik – Eine Einführung* Tübingen; Gunter Narr Verlag.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik** (2008): Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache* 2(2008), 67-73.
- Odebrecht, Carolin; Belz, Malte; Zeldes, Amir; Lüdeling, Anke** (eingereicht) RIDGES Herbology - Designing a Diachronic Multi-Layer Corpus. <https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiterinnen/anke/pdf/odebrechtetalridges-submitted.pdf>
- Rapp, Reinhard; Lezius, Wolfgang** (2001): Statistische Wortartenannotierung für das Deutsche. *Sprache und Datenverarbeitung* 25(2), 5–21.
- Sauer, Simon; Lüdeling, Anke** (erscheint) Flexible Multi-Layer Spoken Dialogue Corpora. Erscheint in *International Journal of Corpus Linguistics, Special Issue on Spoken Corpora* (herausgegeben von John Kirk und Gisle Andersen) <https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiterinnen/anke/pdf/sauerluedelintoappear.pdf>
- Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine** (1999): Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical Report. Institut für maschinelle Sprachverarbeitung, Stuttgart.
- Schmid, Helmut** (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing. Manchester und UK, S. 44–49.
- Schweitzer, Antje; Lewandowski, Natalie** (2013): Convergence of Articulation Rate in Spontaneous Speech. In: Proceedings of Interspeech, S. 525–529.
- Zeldes, Amir ; Schroeder, Caroline T.** (2015): Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities* 31(1), 164-176. https://corpling.uis.georgetown.edu/amir/pdf/Computational_Methods_for_Coptic_prepub.pdf [letzter Zugriff 20.10.2015]
- Zipser, Florian; Romary, Laurent** (2010): A Model Oriented Approach to the Mapping of Annotation Formats using Standards. Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC-2010 . Valletta, Malta, pp. 7–18.

Referenzen

Korpora

- Berlin Map Task Corpus
 - <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/bematac/bematac>
- Fehlerannotiertes Lernerkorpus des Deutschen
 - <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko>
- GECO
 - <http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.html>
- Märchenkorpora
 - Walter, Maik; Maerchenkorpus (Version 1.0), Humboldt-Universität zu Berlin.
<http://www.textbewegung.de/>
<http://hdl.handle.net/11022/0000-0000-8211-9>
- Register in Diachronic German Science
 - Lüdeling, Anke; Odebrecht, Carolin; Zeldes, Amir; RIDGES-Herbology (Version 4.1), Humboldt-Universität zu Berlin. <http://korpling.german.hu-berlin.de/ridges/>.
<http://hdl.handle.net/11022/0000-0000-8253-F>
- Shenoute.a22
 - urn:cts:copticLit:shenoute.a22, Projekt Coptic Scriptorium (Amir Zeldes, Caroline T. Schroeder)
https://corpling.uis.georgetown.edu/annis/scriptorium#_c=c2hlbm91dGUuYTly

Referenzen

Software

- Suche- und Visualisierungstool für Korpora: ANNIS
<http://corpus-tools.org/annis>
- Suchtool für Korpora: COSMAS II
<http://www.ids-mannheim.de/cosmas2/>
- Suchtool für Korpora: CQP
<https://korpling.german.hu-berlin.de/cqpwi/login.php>
- Annotationstools für Korpora: EXMARaLDA
<http://www.exmaralda.org/>
- Repositorium für historische Korpora: LAUDATIO
<http://www.laudatio-repository.org/repository/>
- Repositorium für historische Korpora: Textgrid
<http://www.textgrid.de/>
- Suchtool für Korpora: TIGERSearch
<http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html>
- Suche nach Informationen über Korpora und Tools: Virtual Language Observatory
<http://catalog.clarin.eu/vlo/>