Korpuslinguistik

Plan für den Vormittag

- Organisatorisches
- Einordnung: Welche Korpora gibt es, was können sie?
- Beispiele für korpusbasierte Forschung
- Übungen zu DWDS

Organisatorisches

Hausarbeiten

- Thema frei wählbar

- Mögliche Themen z.B.
  - eines der Themen aus dem unmittelbar folgenden kleinen Forschungsüberblick
  - Zweifelsfälle (vgl. Klein 2003)
  - Grammatikalisierungsphänomene (vgl. Szczepaniak 2009)

Hausarbeiten

- eigene (kleine) Korpusrecherche sehr erwünscht, aber nicht obligatorisch

- Allerdings sollte auf korpuslinguistische Aspekte eingegangen werden:
  - Welche Korpora wurden in der einschlägigen Literatur gewählt und warum?
  - Ist diese Wahl sinnvoll? Wie könnte man die Ergebnisse replizieren oder ergänzen?

Hausarbeiten

- Umfang: nach Prüfungsordnung – bei empirischen Arbeiten kann der Umfang aber i.d.R. sehr gering sein
- "Prototypischer" Aufbau:
  - Kurzer Forschungsüberblick
  - Formulierung von Hypothesen
  - Methode (welches Korpus, welche Suchabfrage)
  - Auswertung
  - Fazit

Korpuslinguistik:
Einige Anwendungsbeispiele

Anwendungsbeispiele

1. Graphematik: Funktionserweiterung des Apostrophs
2. Morphologie: Wortbildungswandel
3. Zweifelsfälle: Fugen-s
4. Syntax-Semantik-Schnittstelle: *ein bisschen* und *ein wenig*
5. Beispiel für exploratives / induktives Arbeiten: Distributionale Semantik

Apostroph (Scherer 2013)

- phonographischer Apostroph: *hab's, gibt's*
- morphographischer (grenzmarkierender) Apostroph: *Moni's Friseursalon, Dienstag's Schnitzeltag*

Apostroph (Scherer 2013)

- Fragestellung: Wie häufig ist der morphographische Apostroph im geschriebenen Deutschen, und welche Faktoren steuern seine Verwendung?

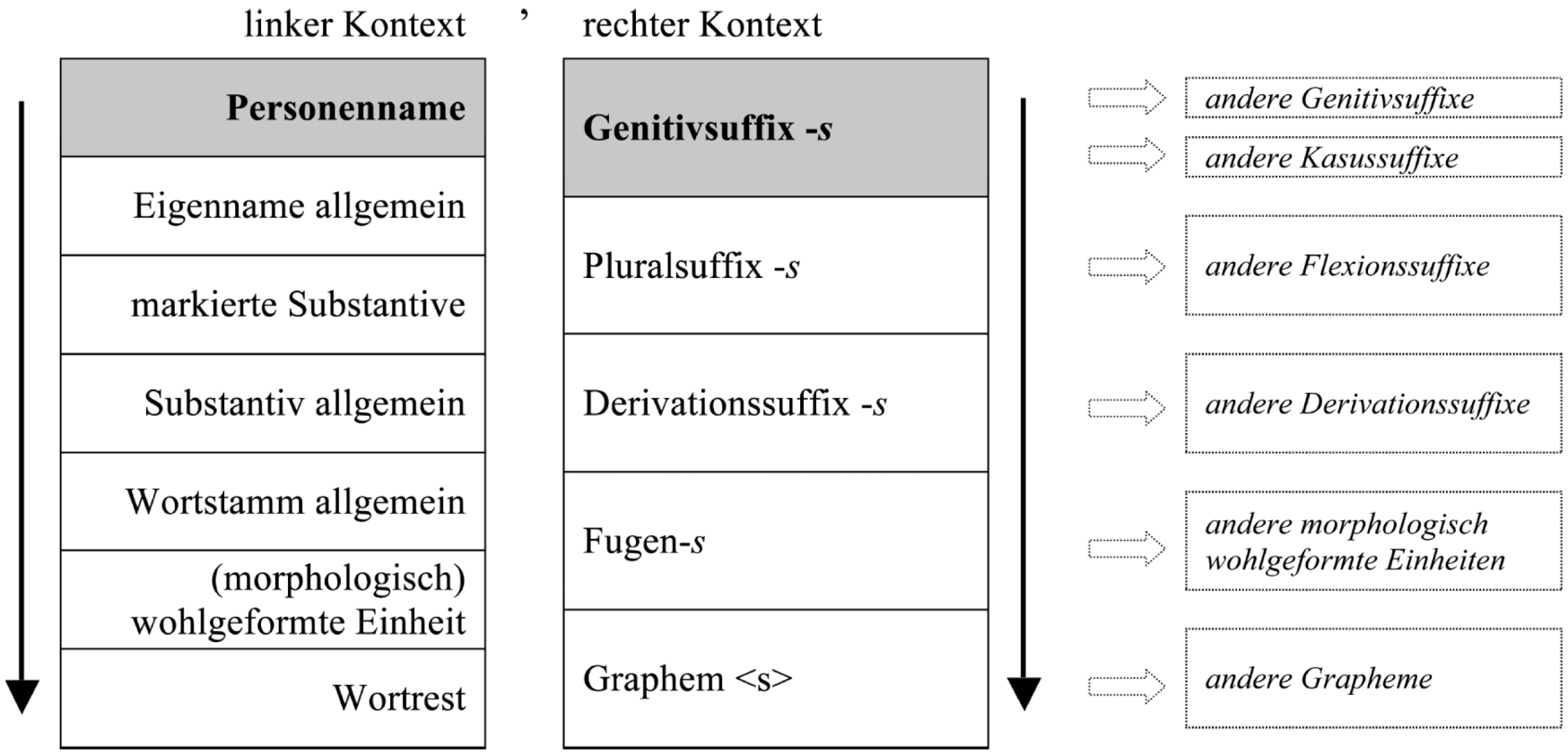
- Korpus: u.a. DWDS-Kernkorpus

Ergebnis

- 10-20 % der Apostrophe im Korpus sind morphographisch
- morphographischer Apostroph tritt v.a. zur Genitivmarkierung bei Personennamen auf; auch historisch älteste Belege in dieser Funktion
  → prototypischer Kontext

Ergebnis

- diachroner Funktionszuwachs des Apostrophs



Morphologie: Wortbildungswandel

Was ist Wortbildungswandel?

Scherer (2006):

Wortbildungswandel als Wandel von Wortbildungs**beschränkungen**, der sich im Wandel morphologischer Produktivität niederschlägt.

Ein Beispiel

Watergate

Nipplegate

Hosen-Gate

Flach, Kopf & Stefanowitsch (2018)

- Was sind die Fragestellungen des Aufsatzes?
- Welche Hypothese(n) untersuchen F, K & S?
- Welche Untersuchungsmethoden werden verwendet?

Exkurs: Namen

- *-gate* als **onymisches** Konfix
- mit *-gate* werden **Eigennamen** abgeleitet
- Im Gegensatz zu Appellativen zeichnen sich Eigennamen aus durch
  - Monoreferenz: Sie referieren auf genau eine Entität (z.B. *Angela Merkel*)
  - Direktreferenz: kein "Umweg" über eine potentielle / prototypische Bedeutung

-gate als Konfix

- Konfixe teilen Eigenschaften mit Affixen und freien Wörtern:
  - wie Affixe sind sie an einen Wortstamm gebunden,
  - wie freie Wörter tragen sie lexikalische Bedeutung.

Übersicht 6: Einheiten der Wortbildung

Einheiten / Merkmale	Wortstamm	Konfix	Affix
bedeutungstragend	ja	ja	nein
wortfähig	ja	nein	nein

(Fleischer & Barz 2012: 64)

Konfix vs. Affixoid

- Affixoid als Einheit zwischen Wort und Affix
- im Gegensatz zum Konfix zeichnet sich das Affixoid durch Desemantisierung / entkonkretisierte Bedeutung aus
- Beispiel: ***Riesen****krach* (nicht *'Krach eines Riesen'), *Laub****werk*** (kein 'Werk', sondern Kollektivum)
- jedoch: enorm umstrittenes Konzept (vgl. z.B. Schmidt 1987, Stevens 2005)

-gate als onymisches Suffix

- aus *Watergate* reanalysiert
- schon im Jahr des Geschehens (1972/73) erste -*gate*-Bildungen im Englischen
- wurde in den vergangenen Jahren auch im Deutschen produktiv, z.B. *Hosen-Gate*

(vgl. Flach, Kopf & Stefanowitsch im Ersch.)

Beispiele (aus Wortwarte)

- "Schnell war von " **Guacamole-Gate** " die Rede . Die Debatte nahm beinahe Loriot'sche Dimensionen an , frei nach dem Motto : Die Erbse bleibt draußen !"
- Falls hier eine Trennwand geplant war, fehlt für ihre Installation der nötige Platz. Unter Mitarbeitern des russischen Außenministeriums kursiert noch eine zweite Erklärung, wie es zum "**Toiletten-Gate**" kommen konnte.

Entwicklung

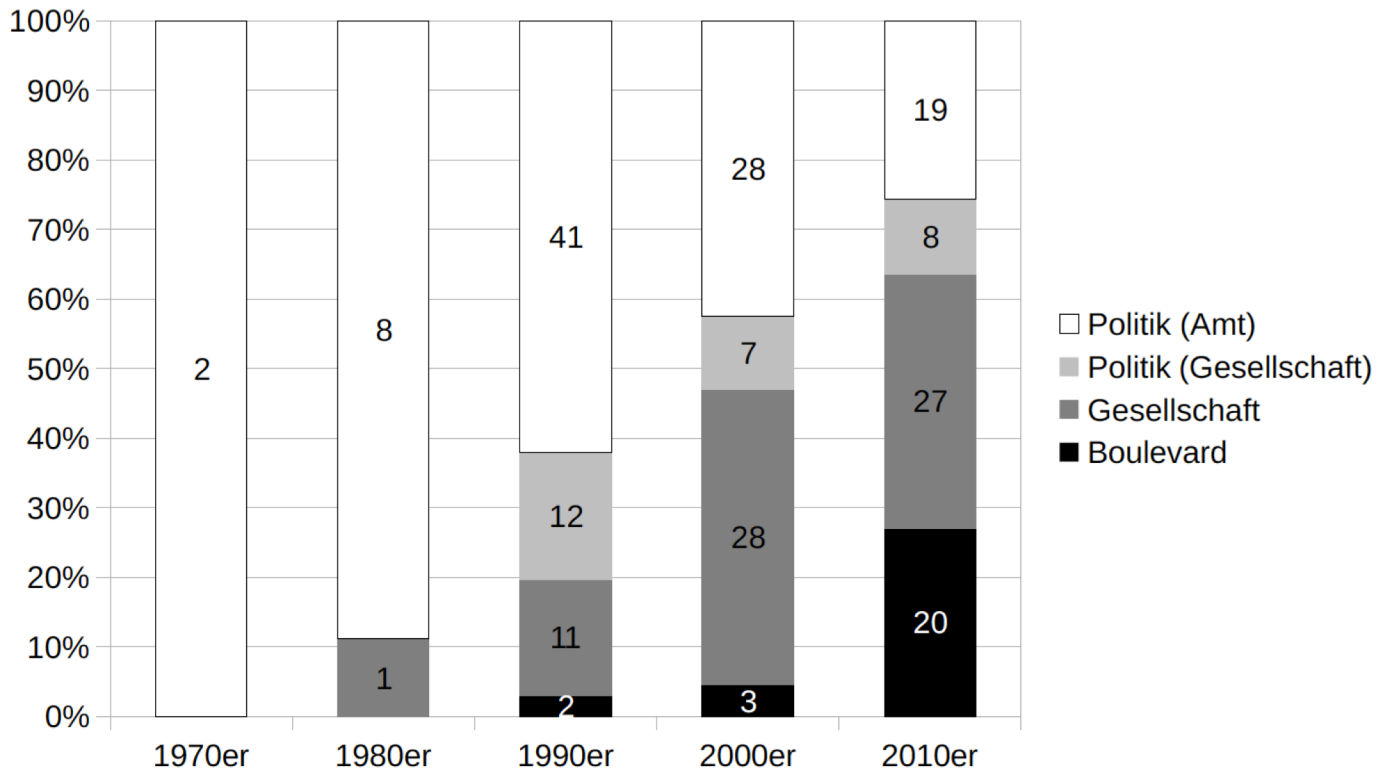


Abb. 5a. Deutsche Erstbelege (Entlehnungen und dt. Bildungen) in DeReKo/ZEIT nach Skandalfeld (n = 217).

aus Flach, Kopf & Stefanowitsch (im Ersch.)

Aufgabe

- Wie können wir nach *-gate*-Bildungen im DWDS suchen?

- Wie können wir Fehltreffer von vornherein aus unseren Ergebnissen ausschließen?

Bedeutungsverschiebung von *-gate*

- Flach et al. (2018) zeigen, dass *-gate* eine "Trivialisierung" erfährt: von großen politischen zu kleinen boulevardesken Skandalen
- Gibt es ähnliche Entwicklungen in anderen Bereichen?

"X-phemism mill"

- Expressive Bedeutungen schleifen sich mit der Zeit quasi ab
- vgl. *scheiße*: Enttabuisierung im Laufe des 20. Jahrhunderts
- Allan & Burridge (2006): "X-phemism mill" – Euphemismen und Dysphemismen verlieren an expressiver Bedeutung und werden durch neue ersetzt
- im Bereich der Euphemismen vgl. z.B. mhd. *kranc* 'schwach' > nhd. *krank*

Expressivität

- Begriff wird häufig recht vage benutzt
- Traugott & Dasher (2002: 94) führen ihn v.a. auf Traugott (1982) zurück
- Dort bezieht er sich v.a. auf die "interpersonale" Komponente von Sprache nach Halliday & Hasan (1976)

Halliday, M.A.K. & Ruqaia Hasan. 1976. *Cohesion in English*. London: Longman.

Traugott, Eliabath Closs. 1982. From propositional to textual and expressive meanings; some semantic–pragmatic aspects of grammaticalization. In Winfred P. Lehmann and Yakov Malkiel, eds., Perspectives on Historical Linguistics, 245–271. Amsterdam: Benjamins

Traugott, Elizabeth Closs & Richard B. Dasher. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Expressivität

- Traugott & Dasher (2002) benutzen *Expressivität* quasi-synonym mit Subjektivität
- diachrone Ausbildung expressiven Gehalts als *subjectification*

Subjectification: "the development of a grammatically identifiable expression of speaker **belief** or speaker **attitude** to what is said" (Traugott 1995)

Expressivität

- In der Literatur zur expressiven Morphologie (z.B. Zwicky & Pullum 1987): Expressivität eng verknüpft mit Pejoration
- zentral ist aber auch hier Evaluation seitens der Sprecherin / des Sprechers
- Beispiel bei Zwicky & Pullum: engl. *shm*-reduplication

Beispiel: *shm*-reduplication

- eignet sich gut für Korpusuntersuchung, weil es im Engl. nur wenige Wörter gibt, die mit <s(c)hm> beginnen.
- Datengrundlage hier: ENCOW16BX (Schäfer & Bildhauer 2012); TV News Archive

Beispiel

- https://archive.org/details/KGO_20180323_063500_Jimmy_Kimmel_Live/start/74.2/end/120

Beispiel: *shm*-reduplication

- **Reboot, schmeboot**, I want season six (http://www.aintitcool.com/node/55397)
- **Issues schmissues!** In this primary race , they do n't matter . (http://madvilletimes.com/2014/05/wismer-a-moderate-kinda-sorta-and-thats-why-dems-should-vote-lowe)
- **Productivity Shmoductivity.** Microsoft's assertion that productivity will suffer is complete tosh. (http://www.pcpro.co.uk/news/387199/microsoft-government-switch-to-open-source-will-cause-dissatisfaction)

Morphologie: Wortbildungswandel

- Bsp.: *ung*-Nominalisierung

(1) kein Fluch / **Murmelung** noch Ungedult wu̍rde bey ihnen geſpu̍rt (Grimmelshausen, Simplicissimus, 1699, DTA)

(2) daz herze ir in dem lîbe spielt/ von sender **jâmerunge**. (Konrad von Würzburg, Herzmaere, 13th century, MHG Dictionary)

Beispiel *ung*-Nominalisierung

- Fragestellung: Wie hat sich die Produktivität der *ung*-Nominalisierung diachron entwickelt?

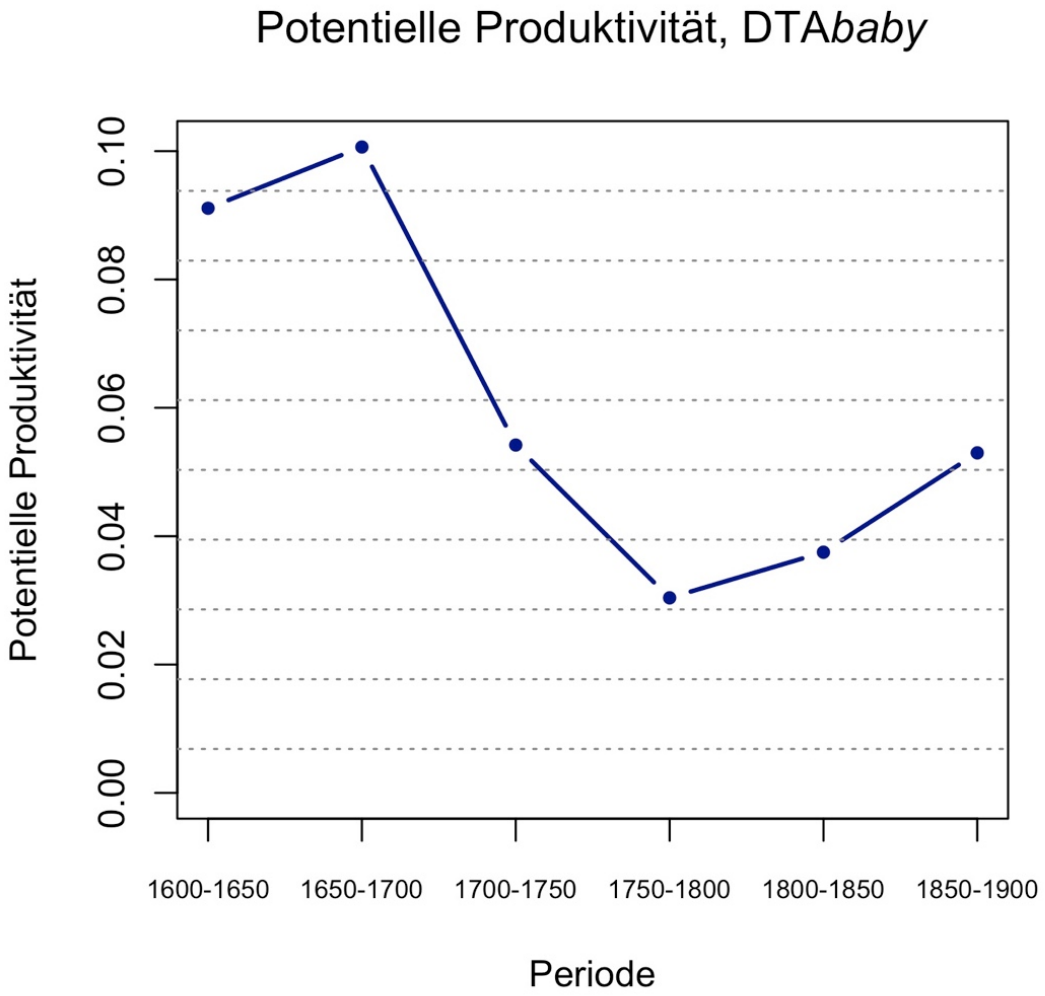
- Datenbasis: u.a. DTA

Potentielle Produktivität

- Einfache Formel:

$$\frac{\text{Anzahl der Hapaxe, die zum Wortbildungsmuster gehören (z.B.: alle Hapaxe auf -heit/-keit)}}{\text{Gesantzahl der Belege, die zum Wortbildungsmuster gehören (z.B.: alle Belege auf -heit/-keit)}}$$

Beispiel *ung*-Nominalisierung

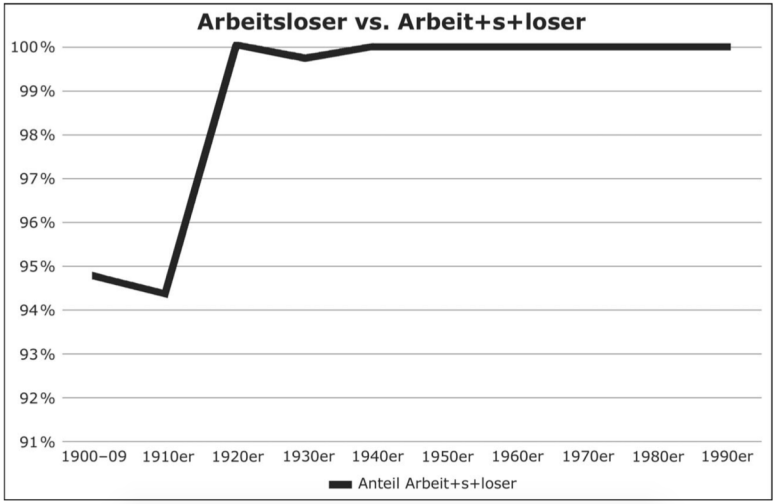


Beispiel Fugenelemente

- Hypothese: Die diachron häufiger werdende s-Fuge markiert "schlechte" phonologische Wörter

- Datenbasis: DWDS-Korpus

Beispiel Fugenelemente



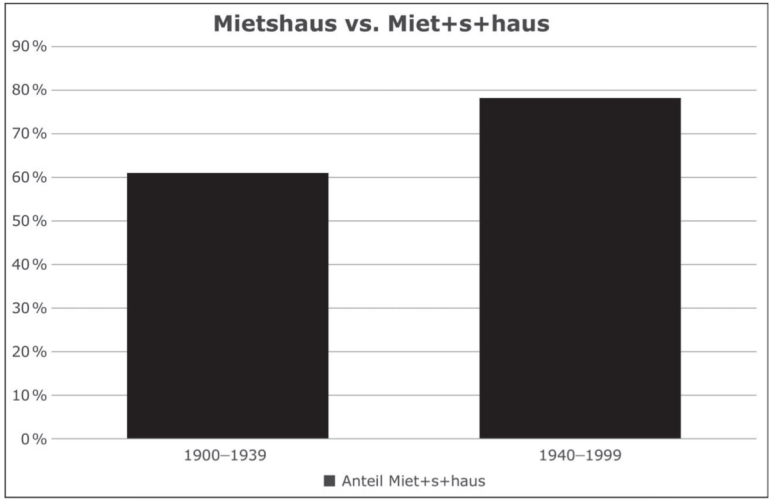


Tabelle 6: *Die Häufigkeit des Fugen*-s *nach präfigierten Erstgliedern.*

	Erstglied enthält:	
	unbetontes Präfix (Typ *Berúf*-)	betontes Präfix (Typ *Ánruf*-)
Tokens:	85 % (von 495.887 Komposita)	36 % (von 324.503 Komposita)
Types:	82 % (von 17.999 Komposita)	37 % (von 11.325 Komposita)

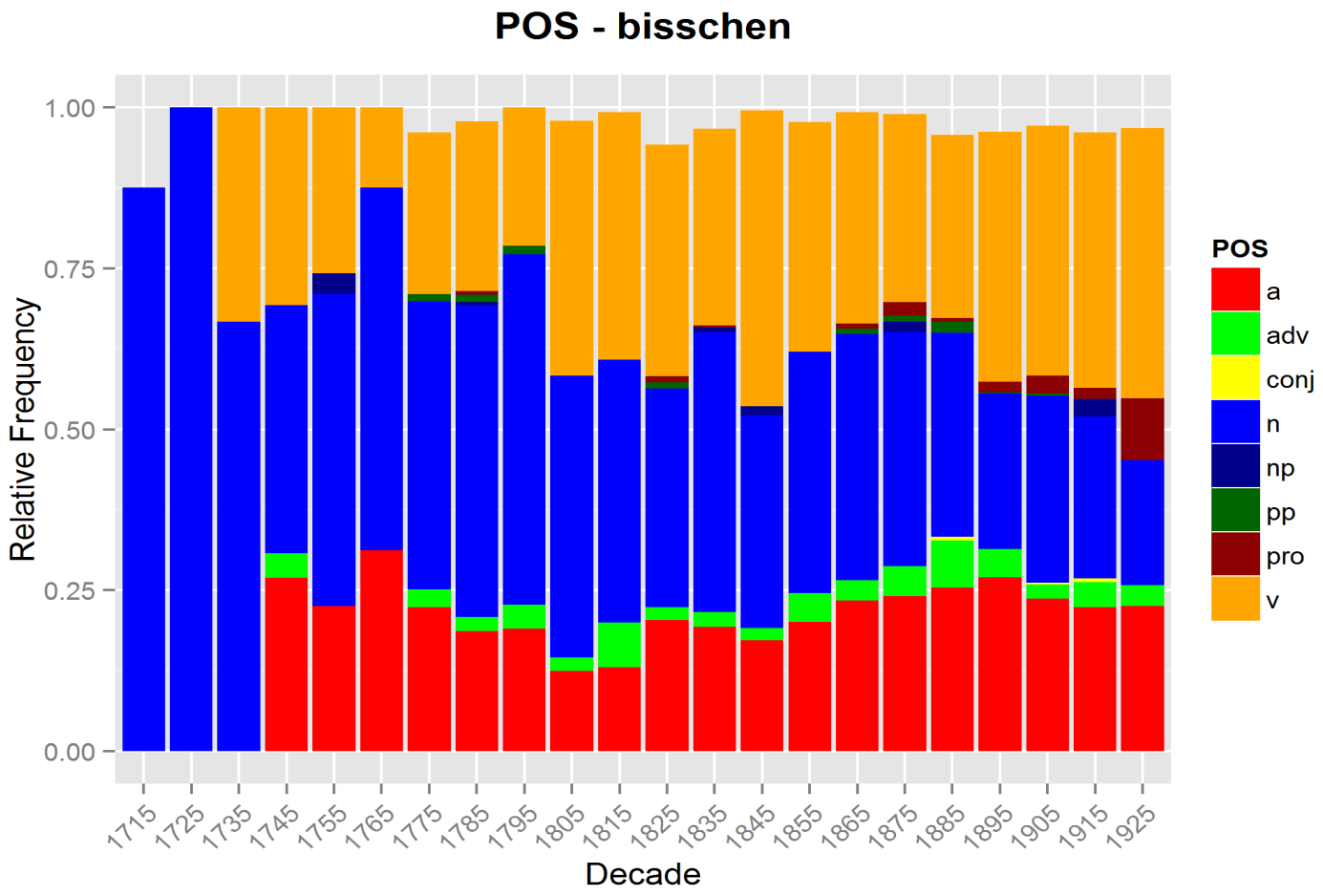
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)

- Traugott (2007): Grammatikalisierungspfad für englische Konstruktionen wie *a bit*

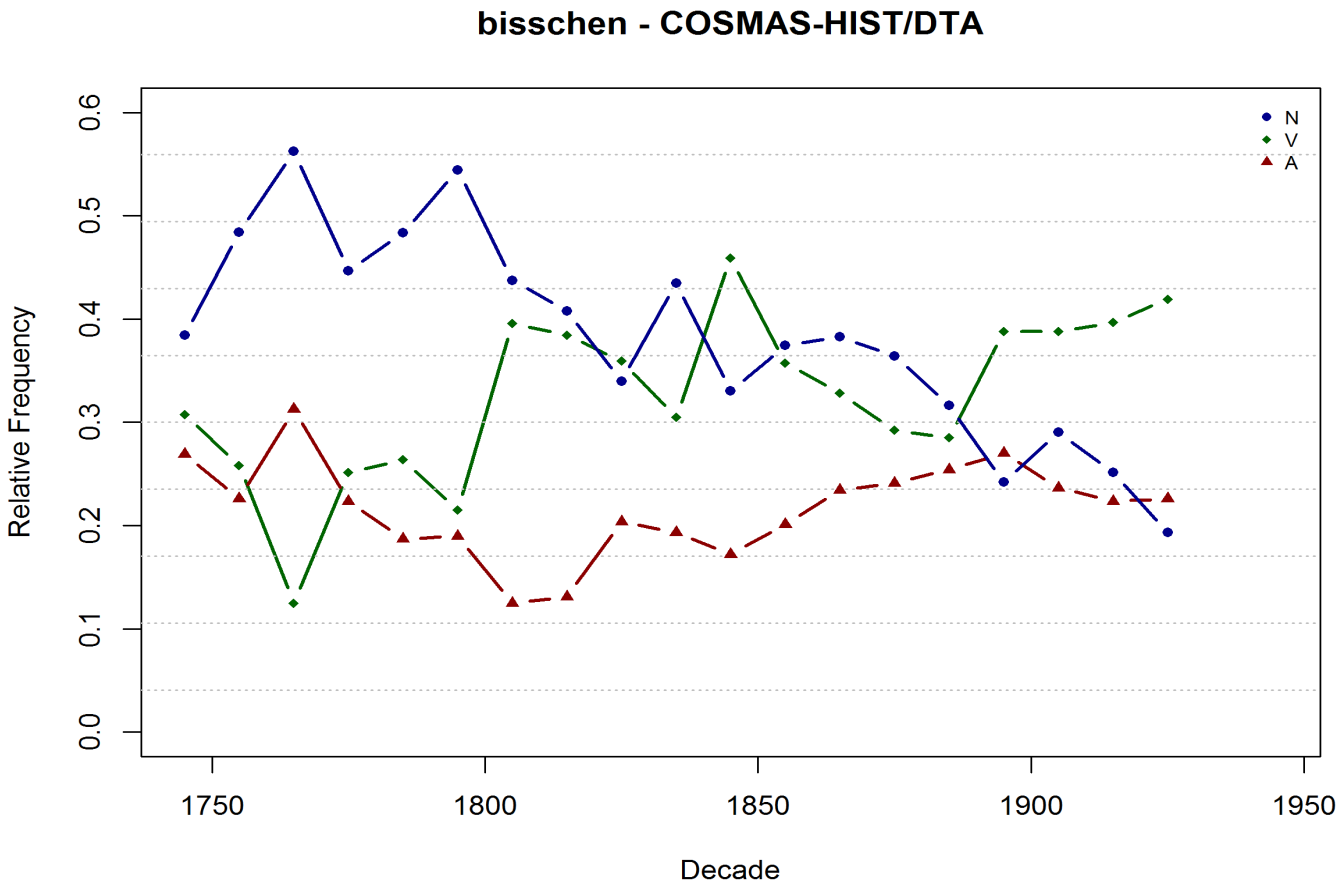
- pre-partitive > partitive > quantifier > degree modifier > free adverb

- Fragestellung: Haben sich die deutschen Konstruktionen *ein bisschen* und *ein wenig* ähnlich entwickelt?
- Datengrundlage: DeReKo-HIST und DTA

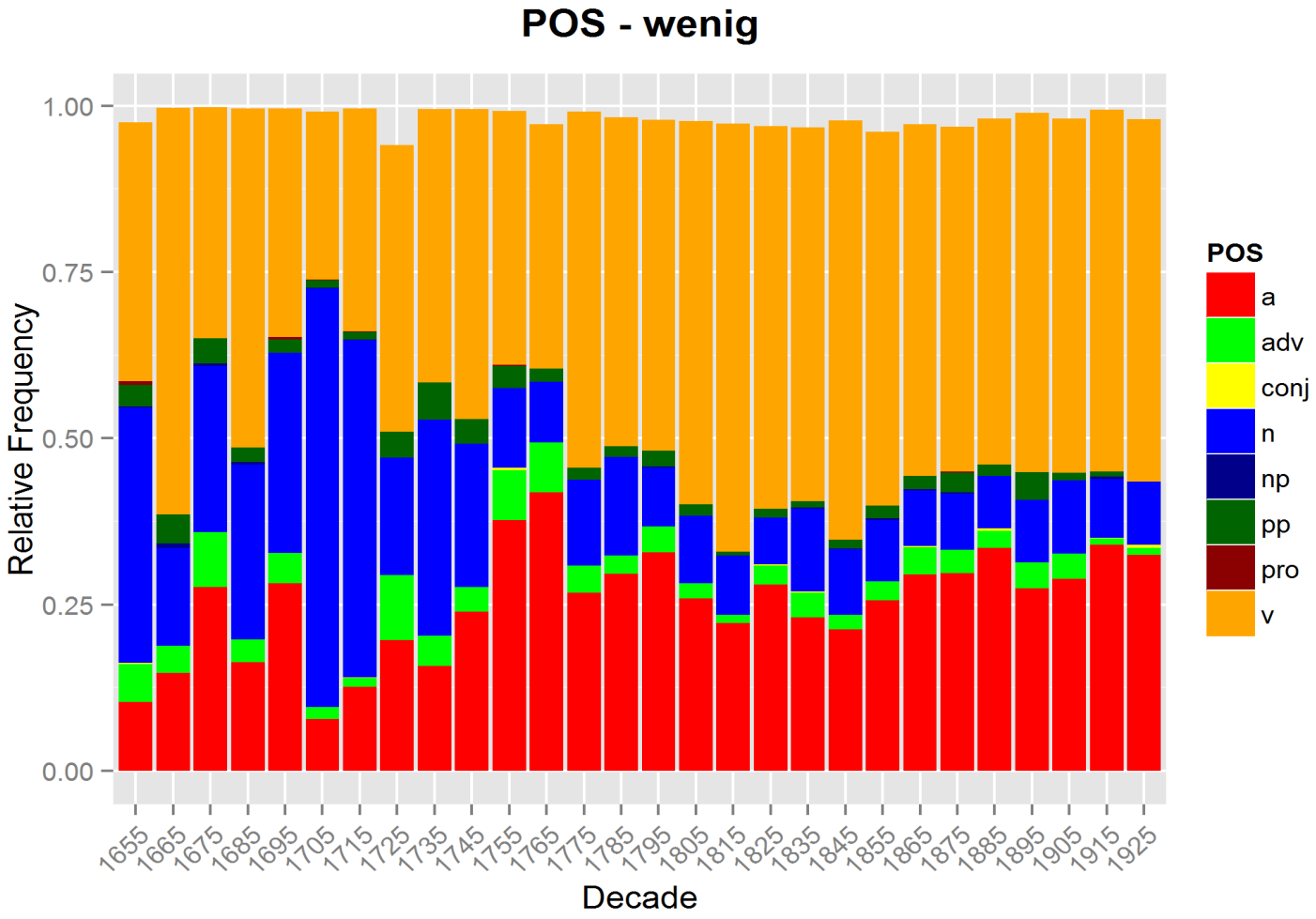
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



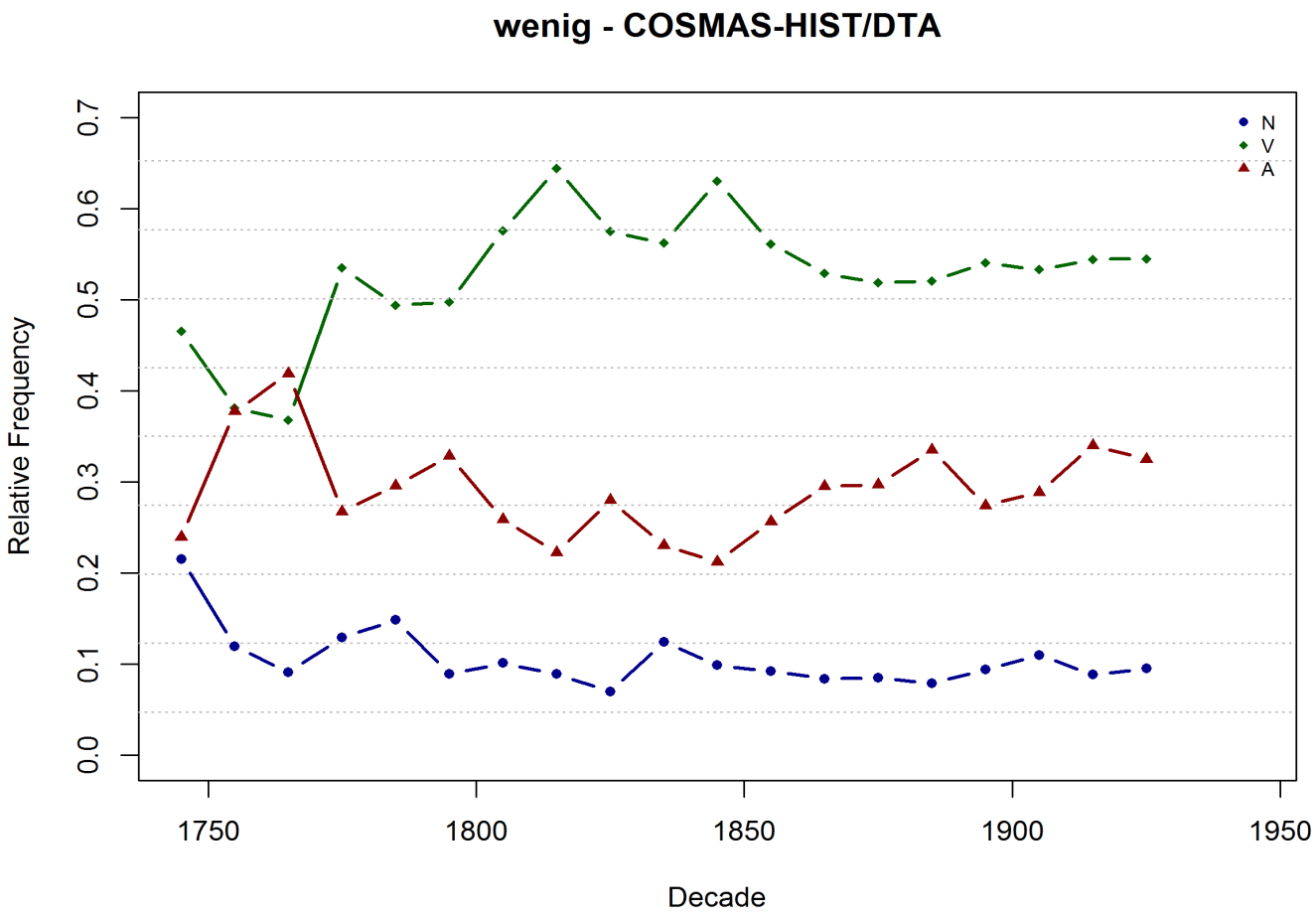
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



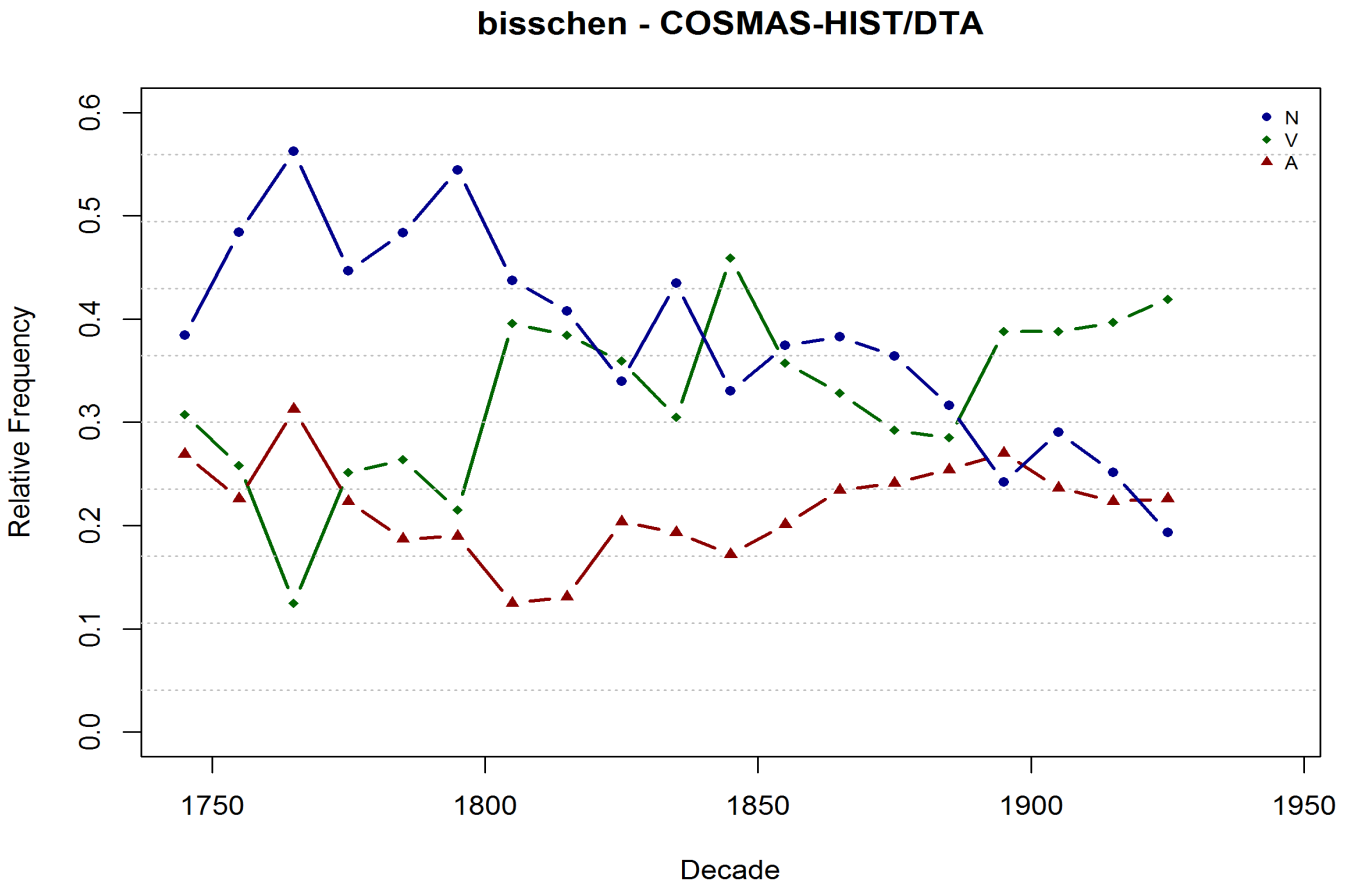
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Übung: DWDS

Etwas komplexer - "Rattenfängerkonstruktion":

- *die Frage, die zu stellen er sich nicht getraut hat*

vs.

- *die Frage, die er sich nicht zu stellen getraut hat*

Eigene Fragestellungen…

DWDS und DTA

Anwendungsbeispiel

- Erste Gehversuche mit einem historischen Korpus des Deutschen

Deutsches Textarchiv

durchsuchbar über

- www.deutschestextarchiv.de
- www.dwds.de

Deutsches Textarchiv

- DTA ist **getaggt** und **lemmatisiert**

```
<tokens>
  <token ID="w1">D.</token>
  <token ID="w2">Henrici</token>
  <token ID="w3">Ca&#x017F;paris</token>
  <token ID="w4">Abelii</token>
  <token ID="w5">,</token>
  <token ID="w6">Wohlerfahrner</token>
  <token ID="w7">Leib-Medicus</token>
  <token ID="w8">Der</token>
  <token ID="w9">Studenten</token>
  <token ID="wa">,</token>
  <token ID="wb">welcher</token>
  <token ID="wc">So</token>
  <token ID="wd">wohl</token>
  <token ID="we">allen</token>
  <token ID="wf">auf</token>
  <token ID="w10">Schulen</token>
  <token ID="w11">Gymna&#x017F;iis</token>
  <token ID="w12">und</token>
  <token ID="w13">Univer&#x017F;ita&#x0364;ten</token>
  <token ID="w14">Lebenden</token>
  <token ID="w15">oder</token>
  <token ID="w16">auf</token>
  <token ID="w17">Rei&#x017F;en</token>
  <token ID="w18">begriffenen</token>
  <token ID="w19">gelehrten</token>
  <token ID="w1a">Per&#x017F;onen</token>
  <token ID="w1b">/</token>
  <token ID="w1c">als</token>
  <token ID="w1d">auch</token>
  <token ID="w1e">allen</token>
  <token ID="w1f">Men&#x017F;chen</token>
  <token ID="w20">insgemein</token>
  <token ID="w21">die</token>
  <token ID="w22">no&#x0364;thig&#x017F;ten</token>
  <token ID="w23">Reguln</token>
  <token ID="w24">und</token>
  <token ID="w25">herrlich&#x017F;ten</token>
  <token ID="w26">Artzeneyen</token>
  <token ID="w27">mittheilet</token>
```

```
<tag tokenIDs="w1e90">ADV</tag>
<tag tokenIDs="w1e91">NN</tag>
<tag tokenIDs="w1e92">APPR</tag>
<tag tokenIDs="w1e93">NN</tag>
<tag tokenIDs="w1e94">ART</tag>
<tag tokenIDs="w1e95">ADJA</tag>
<tag tokenIDs="w1e96">KON</tag>
<tag tokenIDs="w1e97">ADJA</tag>
<tag tokenIDs="w1e98">NN</tag>
<tag tokenIDs="w1e99">$(</tag>
<tag tokenIDs="w1e9a">PRELS</tag>
<tag tokenIDs="w1e9b">APPR</tag>
<tag tokenIDs="w1e9c">PRF</tag>
<tag tokenIDs="w1e9d">ADV</tag>
<tag tokenIDs="w1e9e">ART</tag>
<tag tokenIDs="w1e9f">NN</tag>
<tag tokenIDs="w1ea0">ART</tag>
<tag tokenIDs="w1ea1">NN</tag>
<tag tokenIDs="w1ea2">VAFIN</tag>
<tag tokenIDs="w1ea3">APPR</tag>
<tag tokenIDs="w1ea4">PDAT</tag>
<tag tokenIDs="w1ea5">VAFIN</tag>
<tag tokenIDs="w1ea6">PIAT</tag>
<tag tokenIDs="w1ea7">NN</tag>
<tag tokenIDs="w1ea8">ART</tag>
<tag tokenIDs="w1ea9">NN</tag>
<tag tokenIDs="w1eaa">VVFIN</tag>
<tag tokenIDs="w1eab">$(</tag>
<tag tokenIDs="w1eac">KON</tag>
<tag tokenIDs="w1ead">KOUS</tag>
<tag tokenIDs="w1eae">PPER</tag>
<tag tokenIDs="w1eaf">PTKNEG</tag>
<tag tokenIDs="w1eb0">VVFIN</tag>
```

```
<lemma tokenIDs="wc05a">/</lemma>
<lemma tokenIDs="wc05b">jedoch</lemma>
<lemma tokenIDs="wc05c">aber</lemma>
<lemma tokenIDs="wc05d">d</lemma>
<lemma tokenIDs="wc05e">beide</lemma>
<lemma tokenIDs="wc05f">/</lemma>
<lemma tokenIDs="wc060">d</lemma>
<lemma tokenIDs="wc061">äußerlich</lemma>
<lemma tokenIDs="wc062">und</lemma>
<lemma tokenIDs="wc063">verbergen</lemma>
<lemma tokenIDs="wc064">Verstand</lemma>
<lemma tokenIDs="wc065">sich</lemma>
<lemma tokenIDs="wc066">in</lemma>
<lemma tokenIDs="wc067">d</lemma>
<lemma tokenIDs="wc068">Kontext</lemma>
<lemma tokenIDs="wc069">geschickt</lemma>
<lemma tokenIDs="wc06a">erweisen</lemma>
<lemma tokenIDs="wc06b">mögen</lemma>
<lemma tokenIDs="wc06c">/</lemma>
<lemma tokenIDs="wc06d">damit</lemma>
<lemma tokenIDs="wc06e">beide</lemma>
<lemma tokenIDs="wc06f">d</lemma>
<lemma tokenIDs="wc070">Geheimnis</lemma>
<lemma tokenIDs="wc071">nicht</lemma>
<lemma tokenIDs="wc072">merken</lemma>
<lemma tokenIDs="wc073">/</lemma>
<lemma tokenIDs="wc074">und</lemma>
<lemma tokenIDs="wc075">doch</lemma>
<lemma tokenIDs="wc076">auch</lemma>
<lemma tokenIDs="wc077">verstehen</lemma>
<lemma tokenIDs="wc078">werden</lemma>
<lemma tokenIDs="wc079">.</lemma>
<lemma tokenIDs="wdf00">schwarz</lemma>
<lemma tokenIDs="wdf01">Brief</lemma>
<lemma tokenIDs="wdf02">zu</lemma>
<lemma tokenIDs="wdf03">schreiben</lemma>
<lemma tokenIDs="wdf04">/</lemma>
<lemma tokenIDs="wdf05">daß</lemma>
```

Deutsches Textarchiv

- Zur Suchsyntax (DDC) s. die "Hilfe zur Suche" auf der Website

Deutsches Textarchiv und DWDS

- Annotationsebenen:

$w	Wortebene (im DTA: approximierter Latin-1-Text)
$l	Lemma (unflektierte Form)
$p	Part-of-Speech-Analyse

bei DTA außerdem:

$u	Originaltext, UTF-8-kodiert
$v	CAB-normalisierte Wortform

Aufgabe

Bitte notieren Sie die korrekte Syntax für folgende Anfragen:

1. Die (genaue) Wort**form** *König* (also nicht *Könige, Königs …*)
2. Das **Lemma** *laufen*
3. Die Pluralform *Wagen* vs. *Wägen*
4. Die Konstruktion ADJ *werden* (z.B. *verrückt werden*)
5. Die Abfolge *weil* + Personalpronomen + Verb (z.B. *weil ich sag das halt so*)
6. Apostroph bei Genitivformen von Wörtern, die auf *-s* enden (*des Korpus'*)
7. Verben im Infinitiv vs. Verben im *zu*-Infinitiv
8. Frequenz von *ward* vs. *wurde* im 17., 18. und 19. Jahrhundert

~~Dies da~~

D*

D* (Dstar)

- alterantives Interface für die BBAW-Korpora
- v.a. für Frequenzabfragen geeignet
- allerdings wenig dokumentiert

D* (Dstar)

- Hilfe zur Suche u.a. in diesem Tutorial von Andreas Blombach: http://sprachwissenschaft.fau.de/personen/daten/blombach/korpora.pdf
- und in diesem Blogpost von Frank Wiegand: https://sprache.hypotheses.org/723

D* (Dstar)

- Grundmuster für Frequenzanfragen:

COUNT (hier normale DDC-Abfrage einsetzen)

- Beispiel:

COUNT($p = /NN/g)

(zählt alle Substantive)

D* (Dstar)

- by-Operator: nach $l (Lemma), $p (POS) etc.

COUNT (hier normale DDC-Abfrage einsetzen)

- Beispiel:

COUNT($p = /NN/g) #BY($l)

(zählt alle Substantive nach Lemma)

D* (Dstar)

- Grundmuster für Frequenzanfragen:

COUNT (hier normale DDC-Abfrage einsetzen)

- Beispiel:

```
count( "$w=/[Jj]e/g $w=/.*er/g=1" &&
       "$w=/desto/g $w=/.*er/g=2" )
```

Die Mutter aller (deutschsprachigen) Korpora: DeReKO / COSMAS II

DeReKo

- existiert seit 1964
- größte Sammlung gegenwartssprachlicher Korpora des Deutschen
- kein ausgewogenes Korpus – jedoch lassen sich eigene "virtuelle Korpora" erstellen, die man auf die für die jeweilige Fragestellung relevanten Variablen hin ausbalancieren kann

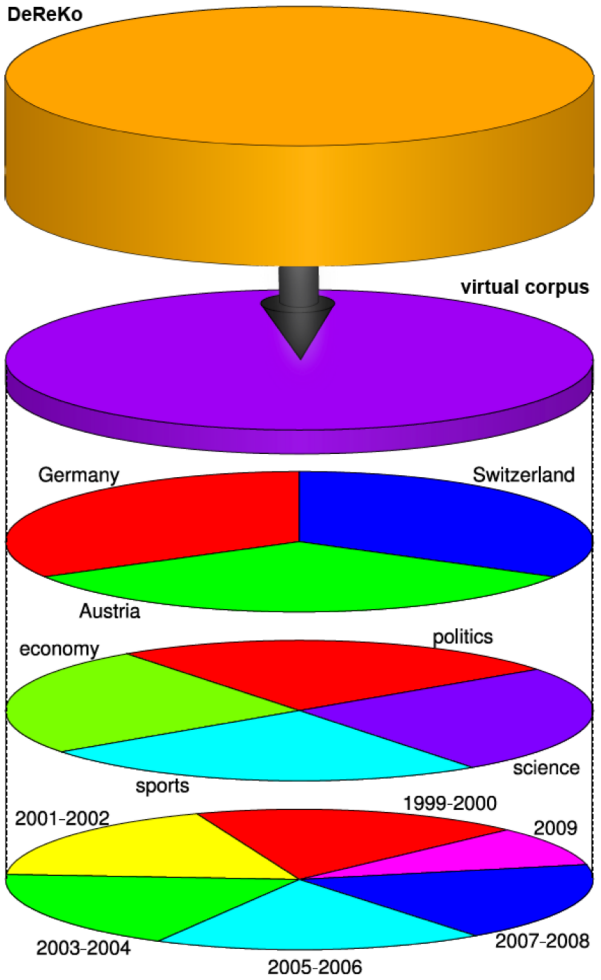
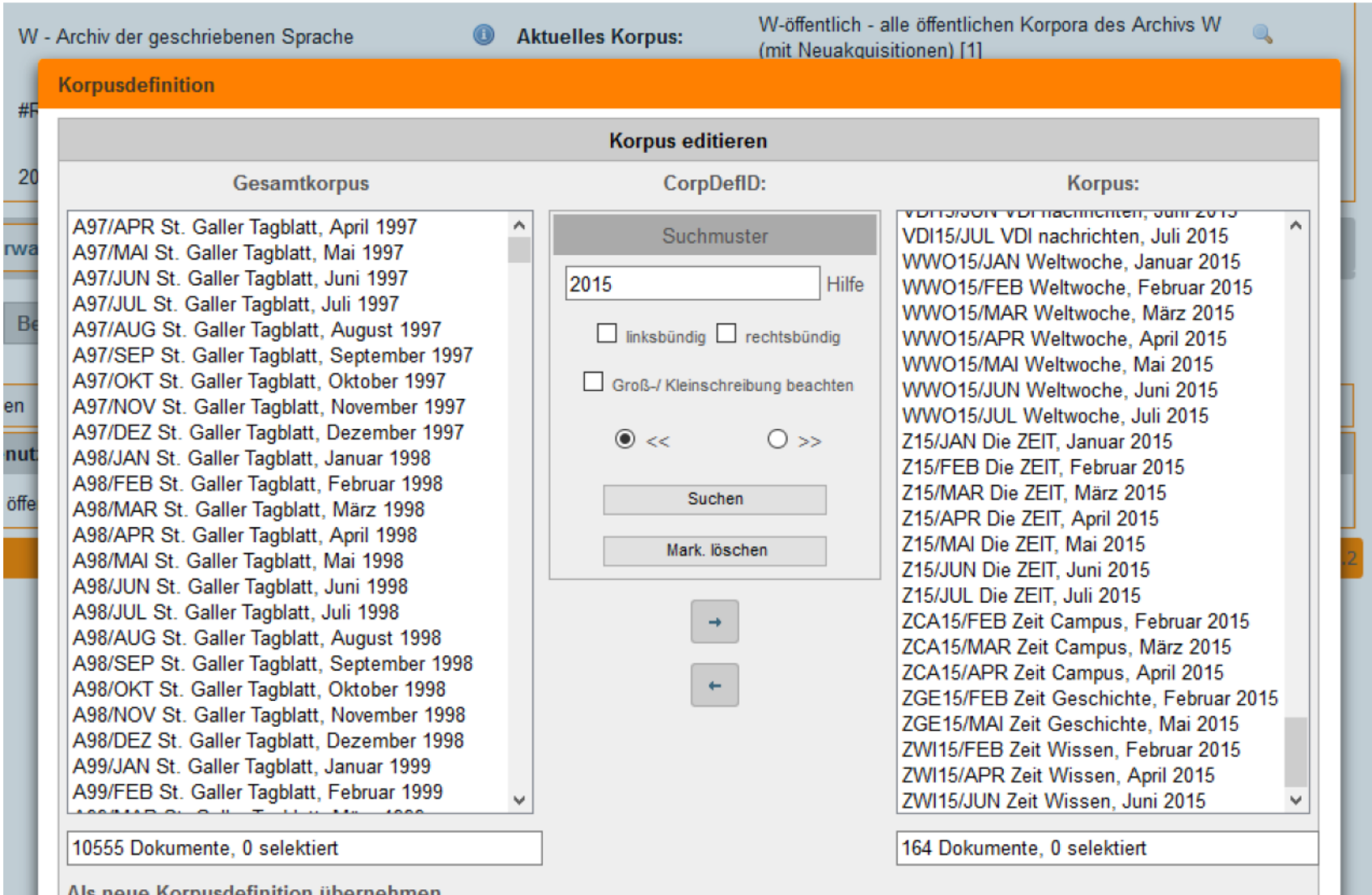


Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

(aus Kupietz et al. 2010)

Virtuelle Korpora erstellen



Zur Konzeption

- DeReKo ist konzipiert als "Urstichprobe" (Kupietz 2010)
- d.h. keine fertig verwendbare Stichprobe, sondern Datenpool, aus dem BenutzerIn sich nach eigenen Kriterien eine (auf bestimmte Kriterien hin ausgewogene) Stichprobe zusammenstellen kann.

Zur Terminologie

Das DeReKo benutzt folgende Bezeichnungen:

- **Dokument:** Es enthält mindestens einen Text (z.B. einen Roman) oder mehrere Texte (z.B. einen Monat Zeitungsartikel des St. Galler Tagblatts.
- **Korpus:** Dieses enthält mehrere Dokumente, z.B. alle Dokumente des St. Galler Tagblatts.
  Virtuelles Korpus: Mehrere Dokumente oder Korpora können aber auch zu virtuellen Korpora zusammengefasst werden.
- **Archiv:** Archive sind die oberste Ebene des DeReKo. Das "Archiv der geschriebenen Sprache" enthält also alle Korpora von geschriebener Sprache. Darin liegt auch das Korpus des St. Galler Tagblatts.

Quelle: https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_www_erste.html

DeReKo

- Zugang nur über Schnittstelle COSMAS II
- Vorteil: relativ großer Funktionsumfang
- Nachteil: beschränkte Exportmöglichkeiten (max. 10000 Belege)

Reguläre Ausdrücke in COSMAS

Platzhalter (Quelle: http://www.ids-mannheim.de/cosmas2//win-app/hilfe/suchanfrage/eingabe-grafisch/syntax/WORT.html)

- Der Platzhalter ***** steht für 0, 1, 2, … beliebige Zeichen.
- Der Platzhalter **+** steht für 0 oder 1 beliebiges Zeichen.
- Der Platzhalter **?** steht für genau 1 beliebiges Zeichen.
- Die Platzhalter können mehrmals innerhalb einer Wortform eingesetzt werden,
- und sie können an jede beliebige Stelle einer Wortform plaziert werden.
- Beim Einsatz des Platzhalters ***** müssen mindestens zwei Buchstaben spezifiziert werden.
- Die Platzhalterfunktion kann aufgehoben werden, indem ein **** vorangestellt wird.

Zur Annotation von COSMAS

- Das große Archiv W ist nicht POS-getaggt
- Jedoch gibt es ein Subkorpus mit getaggten Texten
- Tagged-C und Tagged-C2 (ab 2010): getaggt mit Connexor
- Tagged-T und Tagged-T2 (ab 2010): getaggt mit TreeTagger
- Tagged-Archive haben immer noch > 1 Mrd. laufende Wortformen

Aufgabe

Bitte formulieren Sie Suchausdrücke für

- Wörter mit und ohne Fugenelement
- wegen + NP
- weil + Personalpronomen + Verb (*weil ich sag das halt so)*
- Frequenz von *ward* vs. *wurde* in historischen Texten

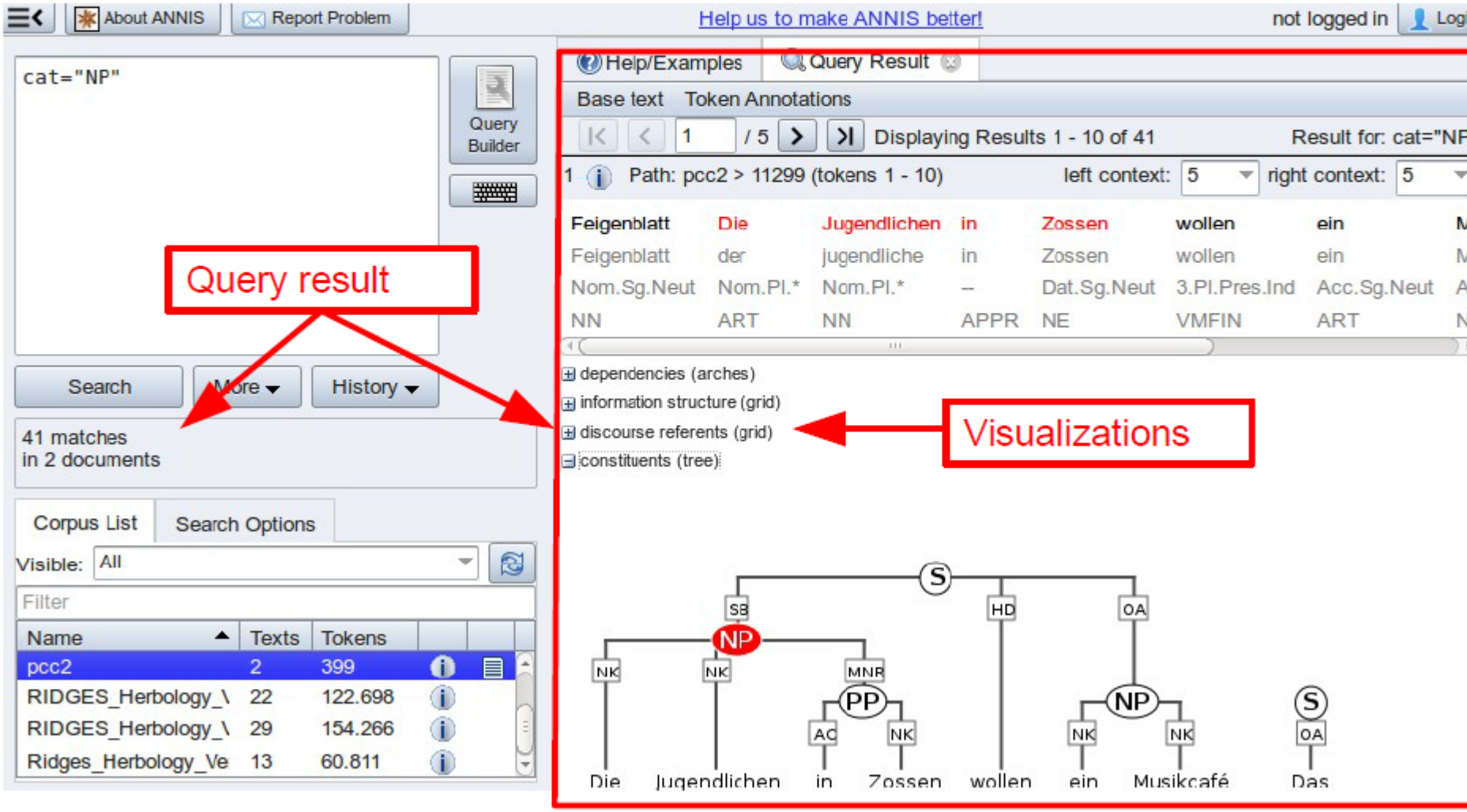
Mehr Aufgaben (Bubenhofer 2019)

ANNIS

ANNIS

- Noch ein Korpusabfragesystem…
- entwickelt an der HU Berlin
- durchsucht existierende, bereits annotierte Korpora (keine eingebaute Annotationsfunktion)
- verschiedene Visualisierungsmöglichkeiten

ANNIS: Search and Visualization in Multilayer Linguistic Corpora



Quelle: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/carolin/annis_workshop2014.pdf

ANNIS

zu den über ANNIS verfügbaren Korpora gehören z.B.

- Deutsch Diachron Digital (im Entstehen)
- RIDGES (historische Kräutertexte)
- Potsdamer Kommentarkorpus (Zetungskommentare)
- KiezDeutschKorpus
- seit kurzem: Bonner Frühneuhochdeutschkorpus

ANNIS Query Language (AQL)

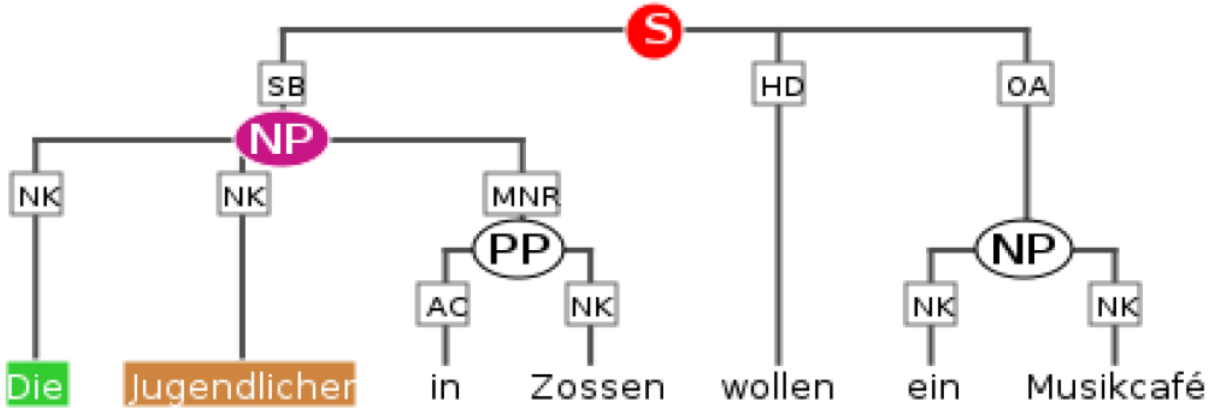
- Ähnlich logisch wie CQL…
- …aber (m.E.) etwas komplizierter.
- Vorteil: extrem detaillierte syntaktische Suchanfragen möglich.
- vgl. "Cheat Sheet für ANNIS"

ANNIS: Search and Visualization in Multilayer Linguistic Corpora

Simplified syntax

- Question:

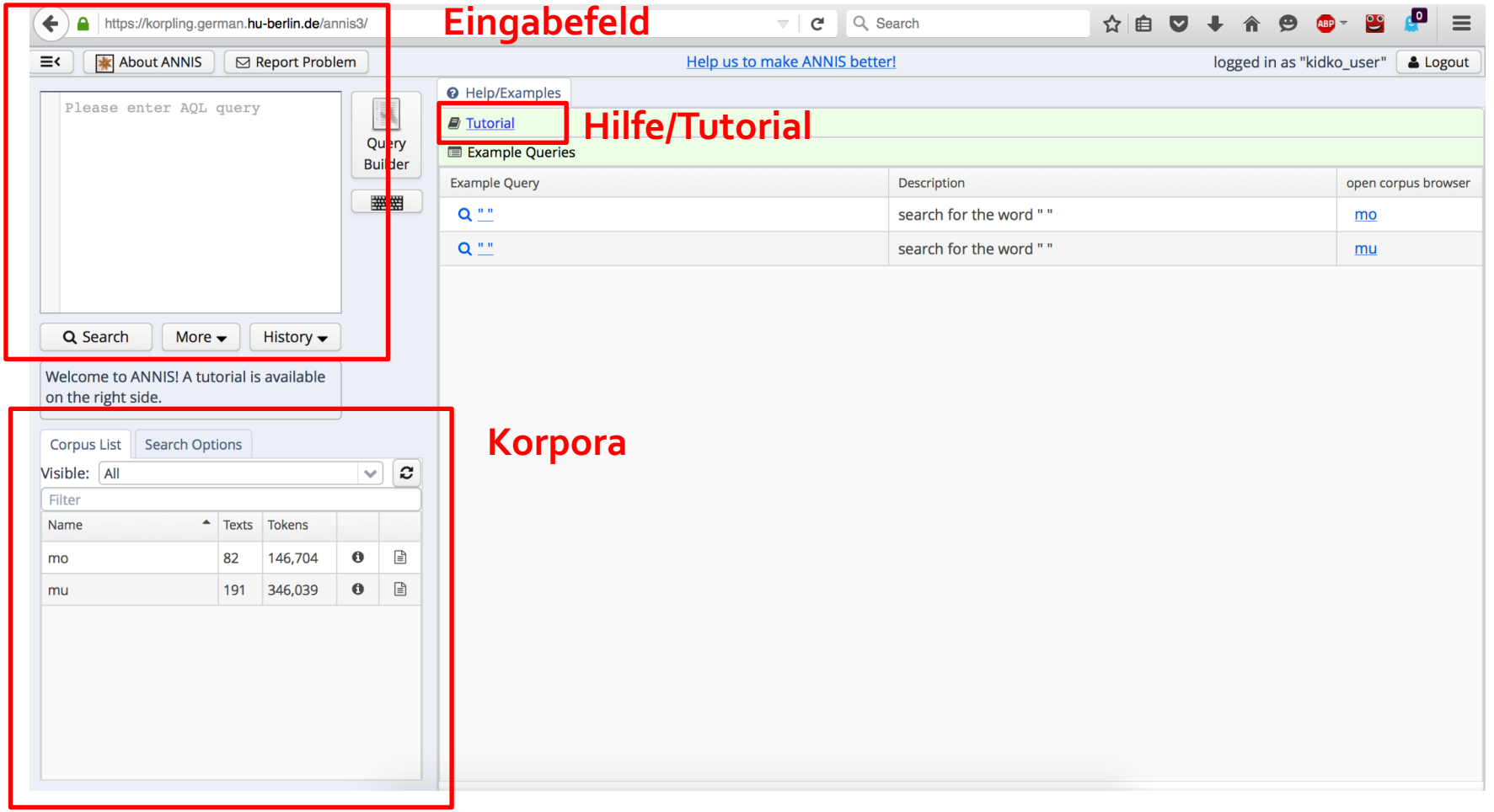
  „Die“ followed by „Jugendlichen“ both being dominated by a prepositional phrase which is dominated by a sentence



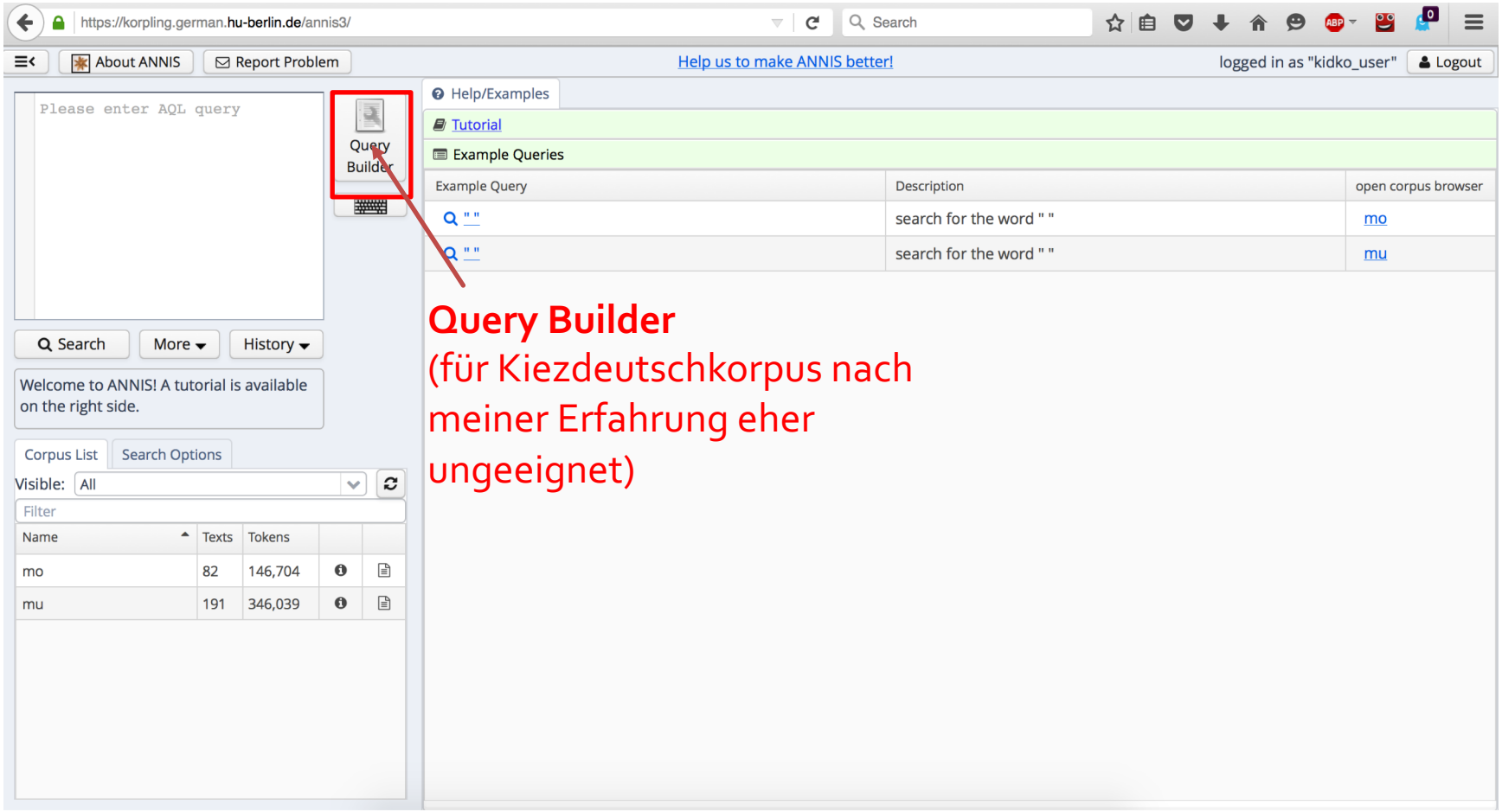
So far:
cat="S" & cat="NP" & "Die" & "Jugendlichen" & #1 > #2 & #2 > #3 & #2 > #4 & #3 . #4

Quelle: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/carolin/annis_workshop2014.pdf

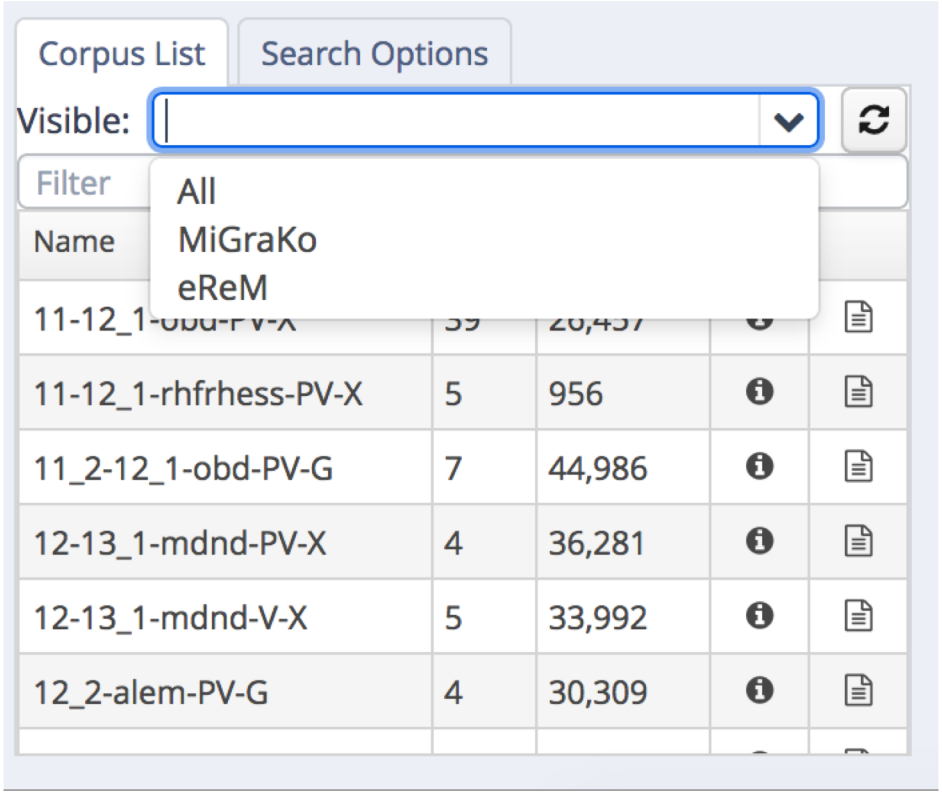
ANNIS Query Language



ANNIS Query Language



REM: Ein Korpus wählen



- MiGraKo: Korpus der mhd. Grammatik
- eReM: Ergänzungstexte zum MiGraKo

Einzelne Tokens suchen

- sehr einfach – aber nicht immer ergiebig, z.B.

"so"

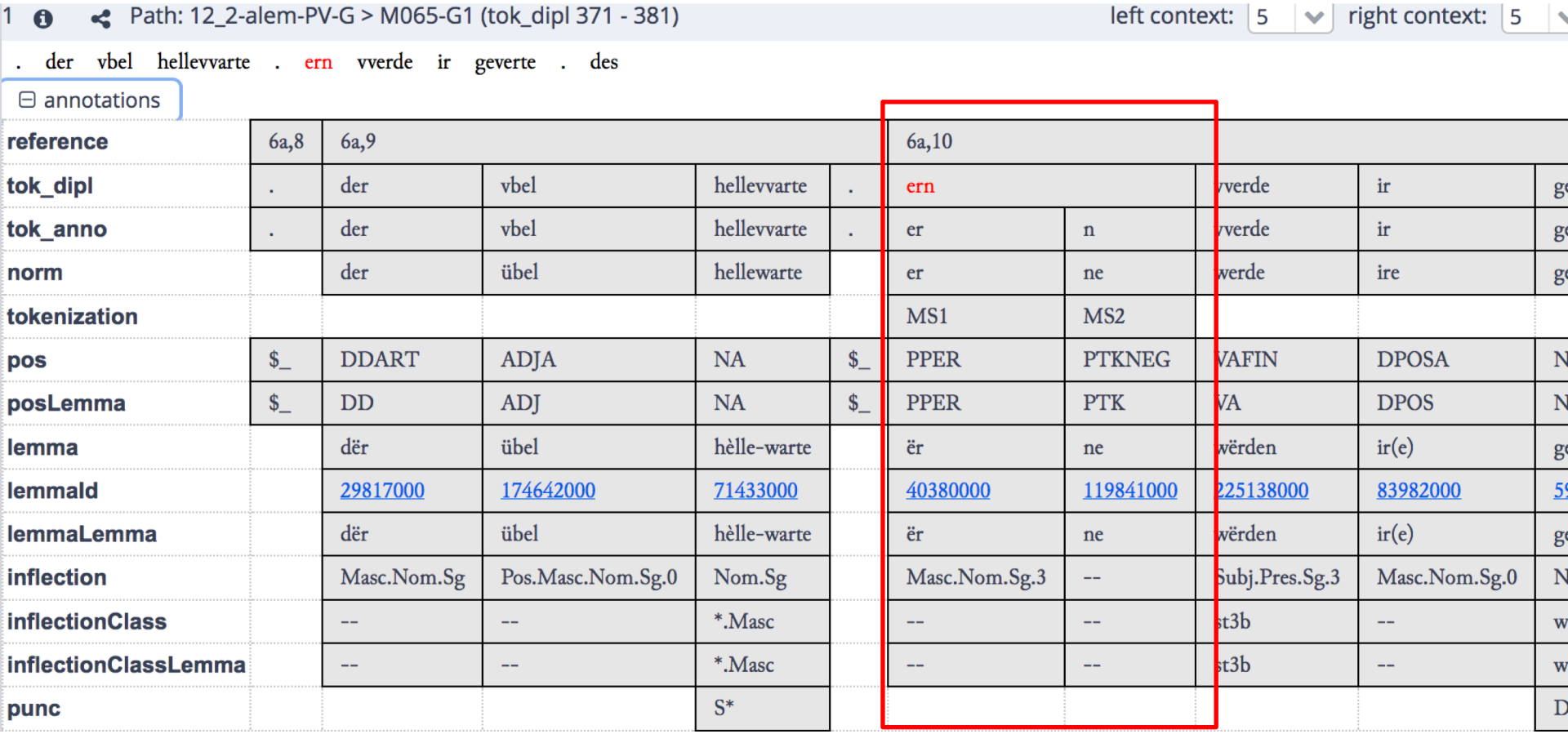
"mann"

"vrouwe" – keine Treffer

Lemmasuche

- Lemmas in REM sind mhd. Lemmas (wie im Mhd. Wörterbuch)
- im Zweifelsfall unter mhdwb.de nach richtigem Lemma suchen.

Ebenen in REM



1 Path: 12_2-alem-PV-G > M065-G1 (tok_dipl 371 - 381) left context: 5 right context: 5

. der vbel hellevvarte . ern vverde ir geverte . des

annotations

reference	6a,8	6a,9				6a,10				
tok_dipl	.	der	vbel	hellevvarte	.	ern		vverde	ir	ge
tok_anno	.	der	vbel	hellevvarte	.	er	n	vverde	ir	ge
norm		der	übel	hellewarte		er	ne	werde	ire	ge
tokenization						MS1	MS2			
pos	$_	DDART	ADJA	NA	$_	PPER	PTKNEG	VAFIN	DPOSA	N
posLemma	$_	DD	ADJ	NA	$_	PPER	PTK	VA	DPOS	N
lemma		dër	übel	hèlle-warte		ër	ne	wërden	ir(e)	ge
lemmaId		29817000	174642000	71433000		40380000	119841000	225138000	83982000	59
lemmaLemma		dër	übel	hèlle-warte		ër	ne	wërden	ir(e)	ge
inflection		Masc.Nom.Sg	Pos.Masc.Nom.Sg.0	Nom.Sg		Masc.Nom.Sg.3	--	Subj.Pres.Sg.3	Masc.Nom.Sg.0	N
inflectionClass		--	--	*.Masc		--	--	st3b	--	w
inflectionClassLemma		--	--	*.Masc		--	--	st3b	--	w
punc				S*						D

Übung: Negation im Mhd.

- Wir wollen die Entwicklung der Negation im Mhd. untersuchen:

- "doppelte" Negation mit *en* und *niht* vs. einfache Negation
- Wonach müssen wir suchen?

Übung: Konversion im Mhd.

- Wir wollen alle nominalisierten Infinitive im REM finden
- (z.B. nhd. *das Singen, das Tanzen*)
- Hierfür können wir die pos- und posLemma-Annotation nutzen…

Einzelne Tokens suchen

- Groß- und Kleinschreibung beachten!
- *so* findet nur klein geschriebene Belege, *So* nur groß geschriebene
- Oft wird *SO* als Fokusmarker in der Transkription auch komplett groß geschrieben.
- Wie finden wir alle drei Formen?
- Logisch: Operatoren!

Einzelne Tokens suchen

"so" | "So" | "SO"

Einzelne Lemmas suchen

- Wie bei CQP, gilt auch bei ANNIS: Attribute können unterschiedliche Namen haben!
- Wir erinnern uns: "lemma" heißt im BNC aus unerfindlichen Gründen "hw"
- Beim Kiezdeutsch-Korpus gibt es keine Lemmatisierung, nur Normalisierung…

n

Ebenen

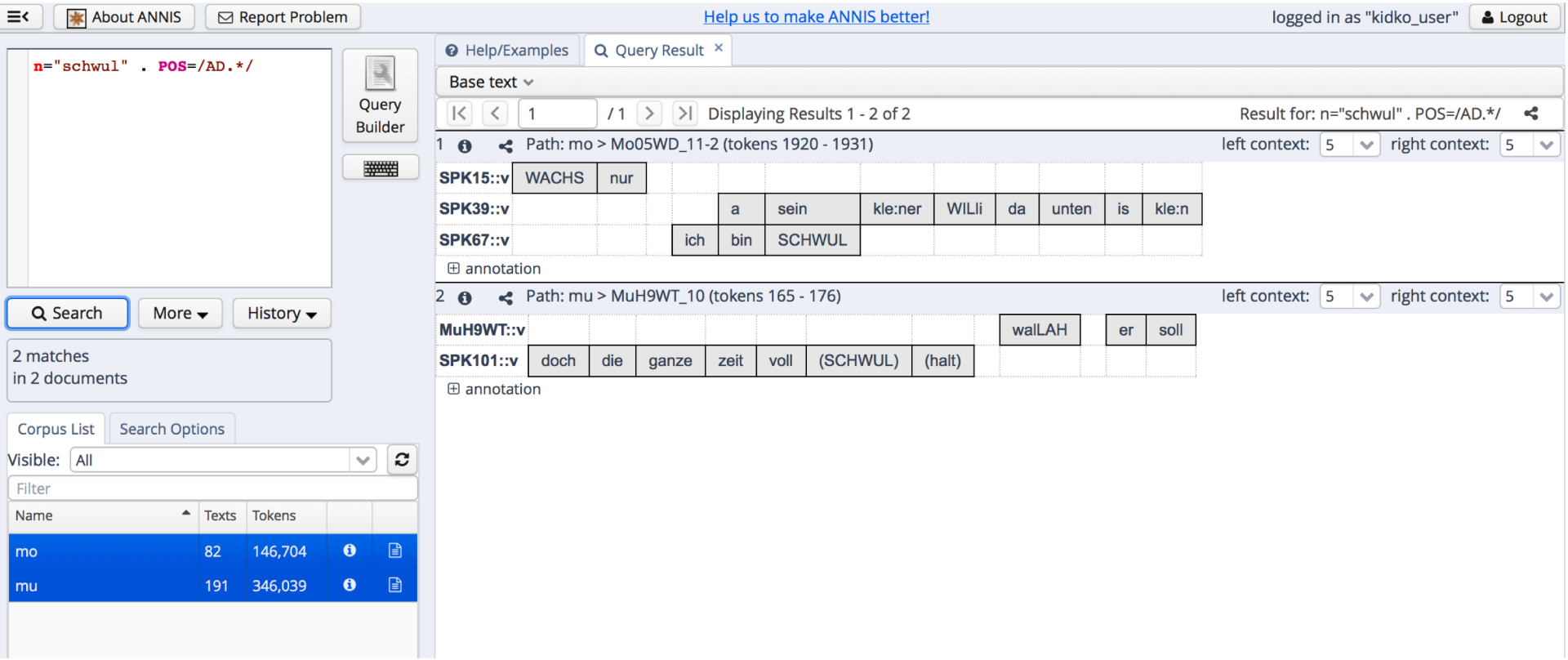
Kürzel	Ebene
nv	nonverbale Ebene
v	Transkriptionsebene
n	Normalisierungsebene
POS	Wortartenebene (part-of-speech)
tr	türkische Transkriptionsebene
trnorm	türkische Normalisierung
trdtwwue	deutsche Übersetzung Wort für Wort
trdtue	freie deutsche Übersetzung

Einzelne Lemmas suchen

- i.d.R. wollen wir die normalisierte Ebene durchsuchen (n)
- um ein Lemma zu suchen, geben wir daher ein: n="

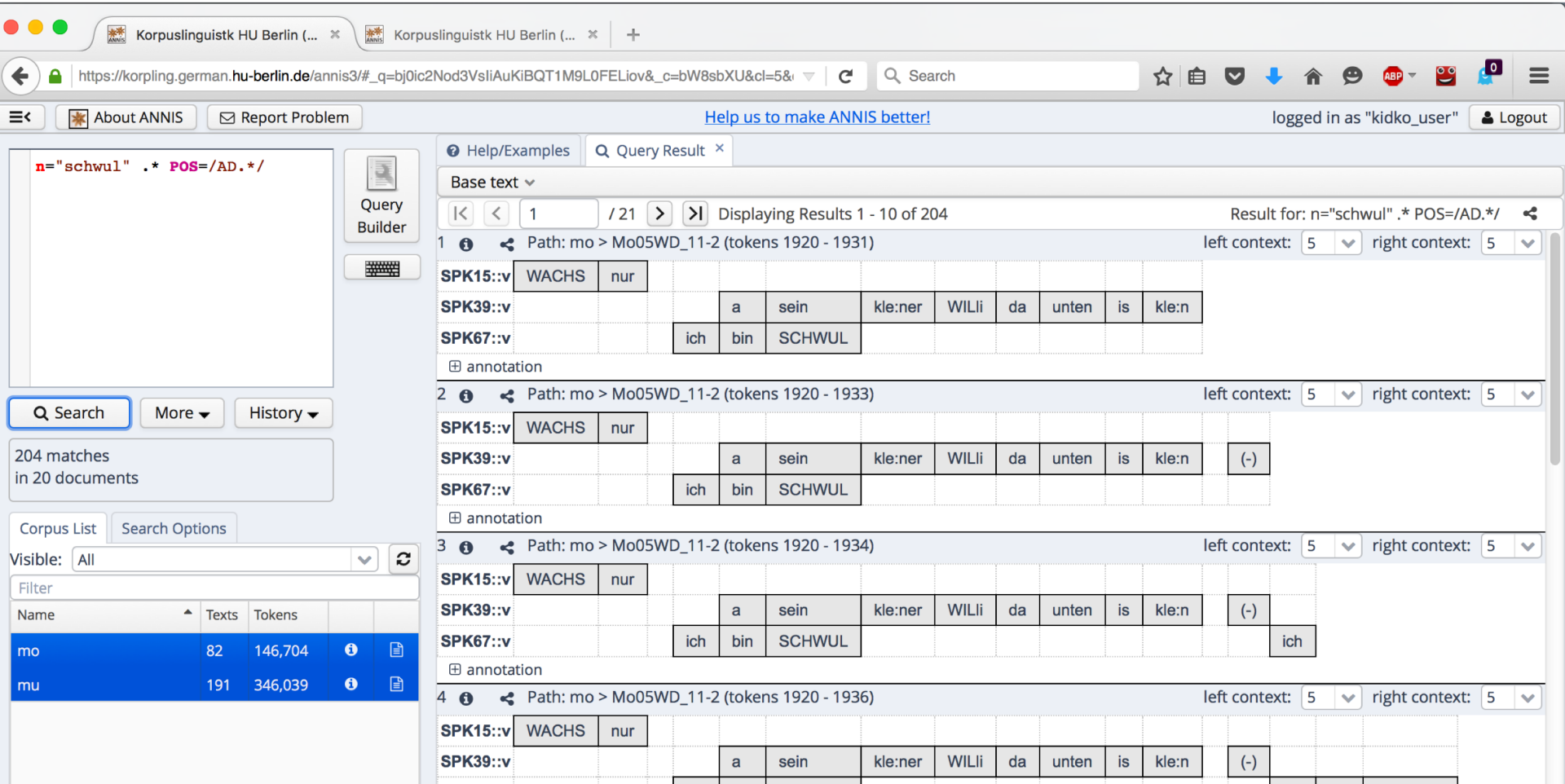
Wortfolgen suchen

. = geht direkt voran



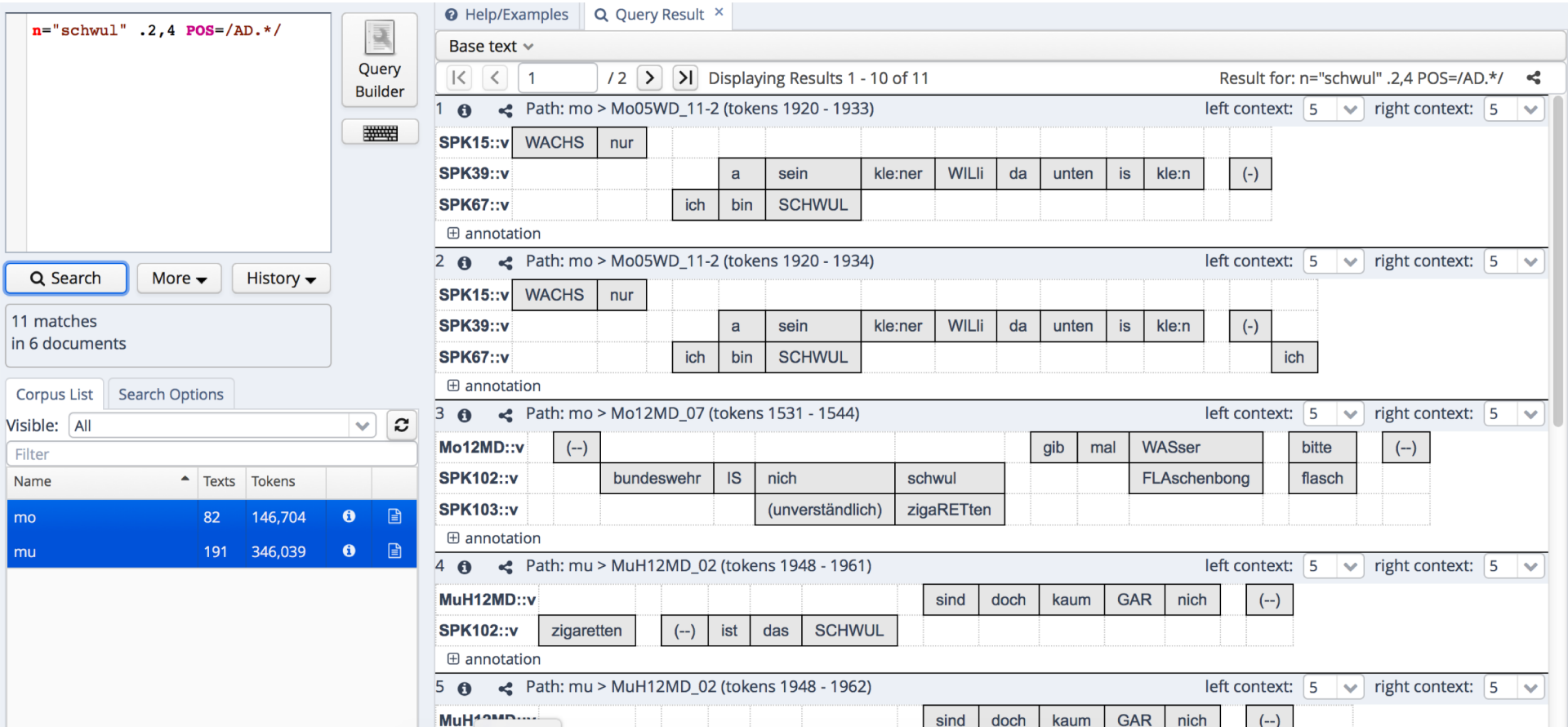
Wortfolgen suchen

.* = geht indirekt voran



Wortfolgen suchen

.x,y = geht im Abstand von min. x, max. y Wörtern voran



Exakte Sequenzen vs. Suche mit Wildcards

- Sucht man bei ANNIS nach exakten Sequenzen, benutzt man **Anführungszeichen:**

tok="Jugendliche"

pos="NN"

- Sucht man hingegen mit **Wildcards** und regulären Ausdrücken, benutzt man Slashes:

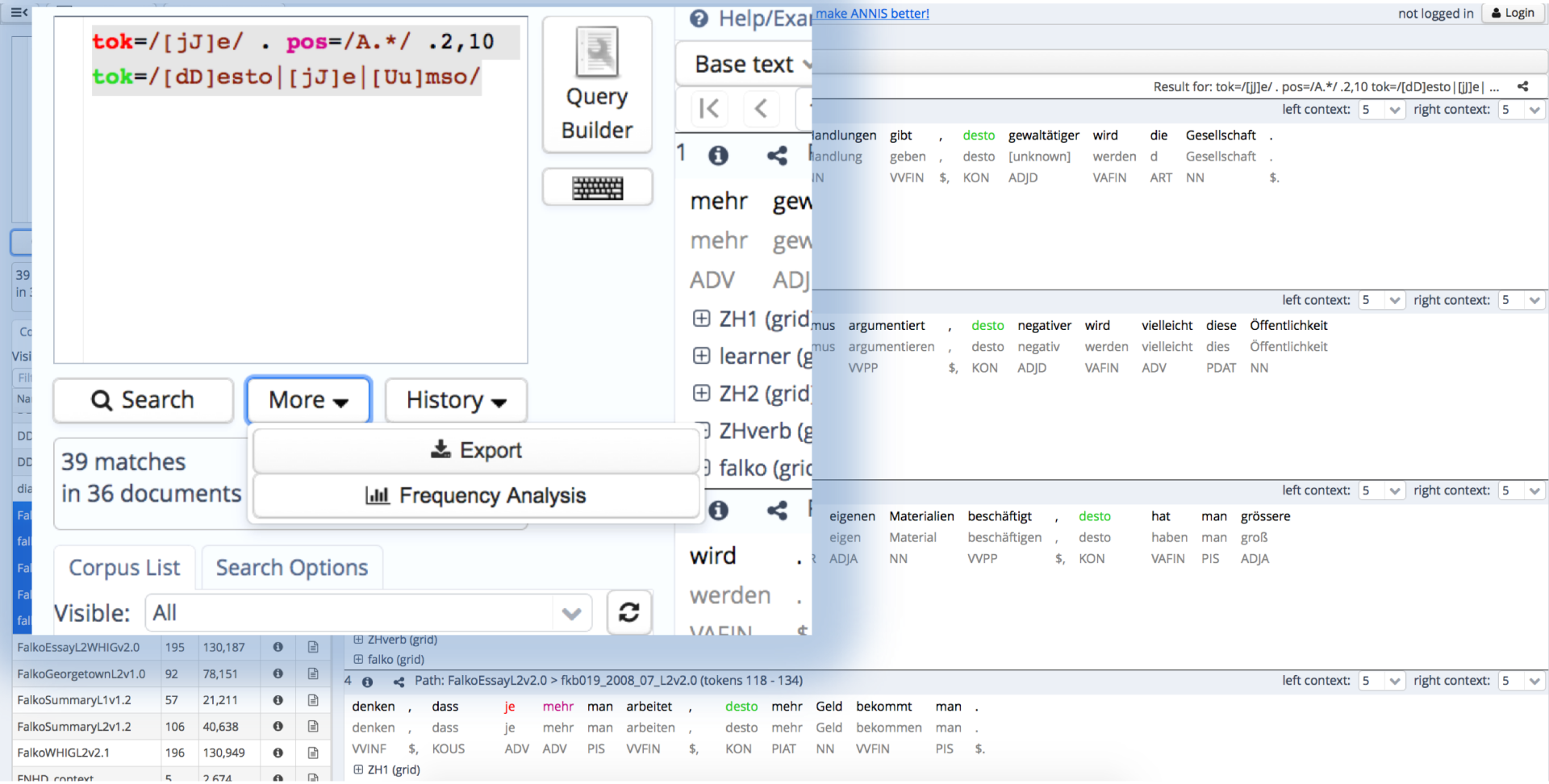
tok=/Jugendliche.*/

pos=/N.*/

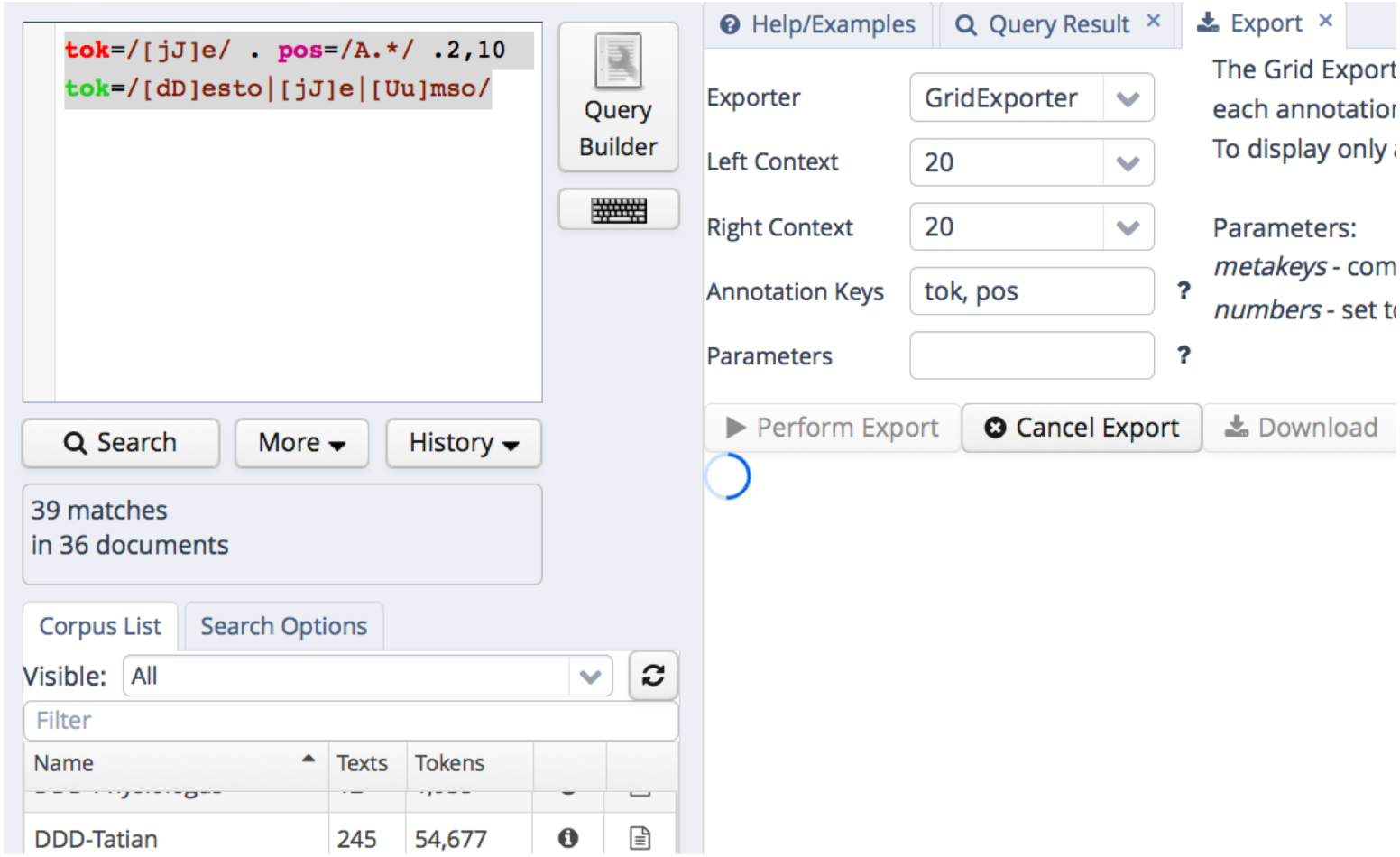
Beispielanfrage

- Wir suchen nach *je* X-*er desto/umso* Y-*er* im Falko-Essay-Korpus
- Worauf müssen wir achten, um alle Varianten zu finden?

Vom Resultat zum Export

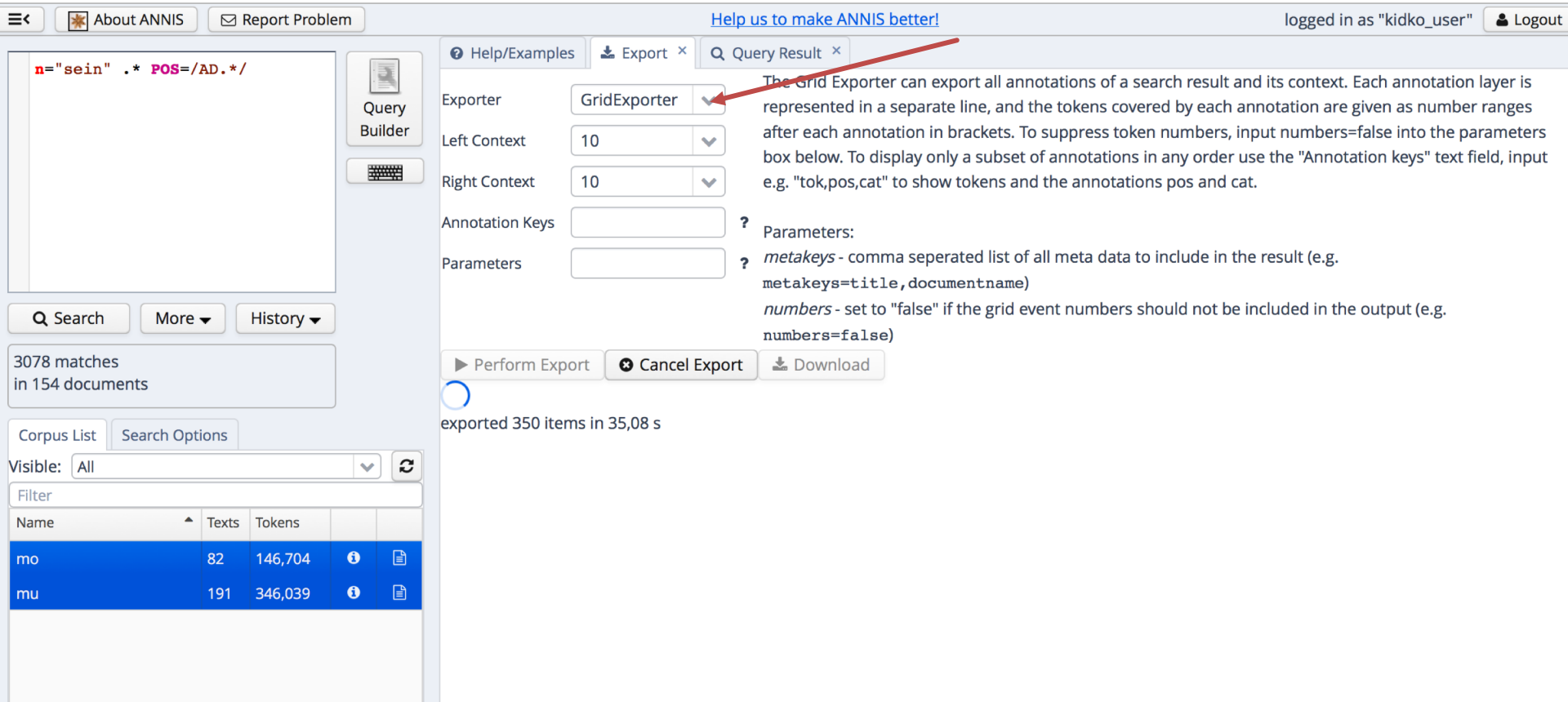


Vom Resultat zum Export

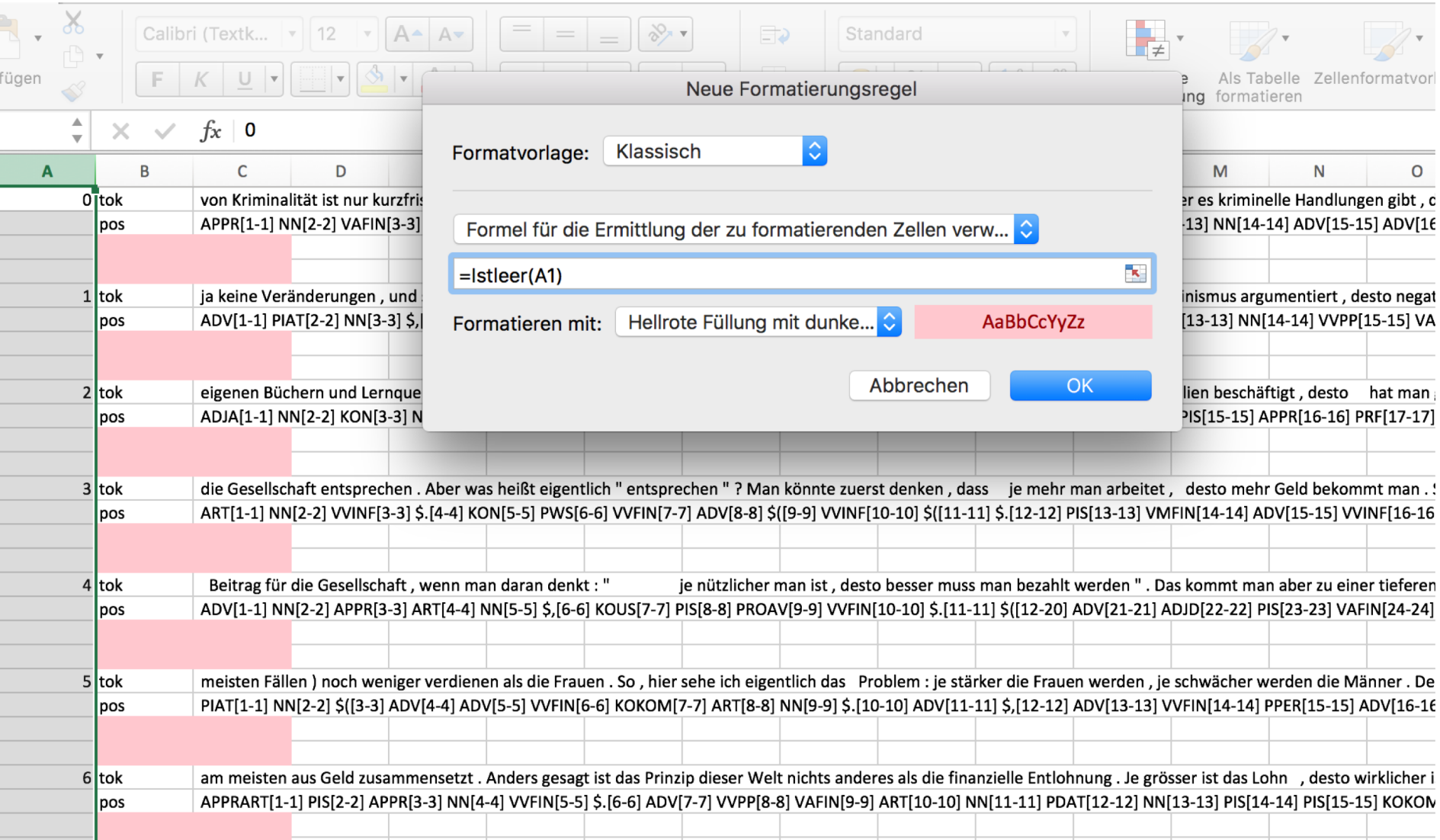


Export

- derzeit scheint fürs Kiezdeutschkorpus nur der **GridExporter** zu funktionieren



Bedingte Formatierung in Excel



Baumbanken in ANNIS

Baumbanken

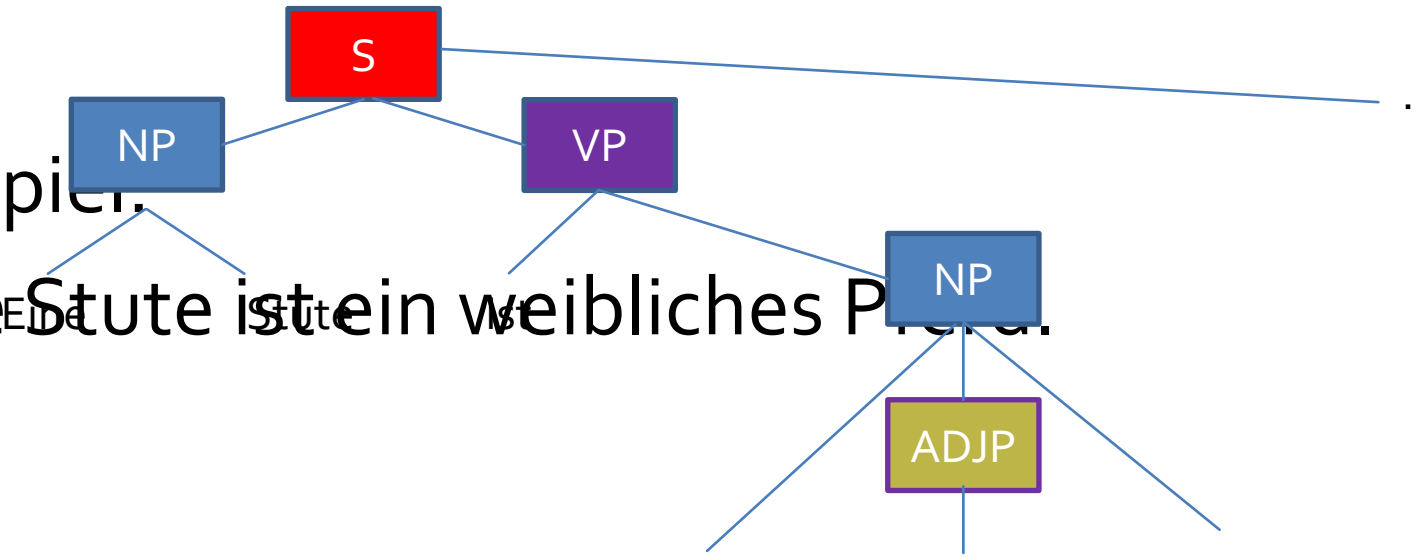
- syntaktisch geparste Korpora
- Syntax wird über sog. Strukturbäume dargestellt
- ein Strukturbaum besteht aus **Kanten** (*edges*) und **Knoten**
- wohl verbreitestes Tagset fürs Deutsche: Tagset der TIGER-Baumbank (Albert et al. 2004)

Konstituentenstruktur und Dependenzstruktur

- zwei zentrale syntaktische Formalismen
- Konstituentenstruktur: Wörter werden in **Phrasen** gruppiert, wobei in jeder Phrase ein Wort als **Kopf** fungiert

Beispiel:

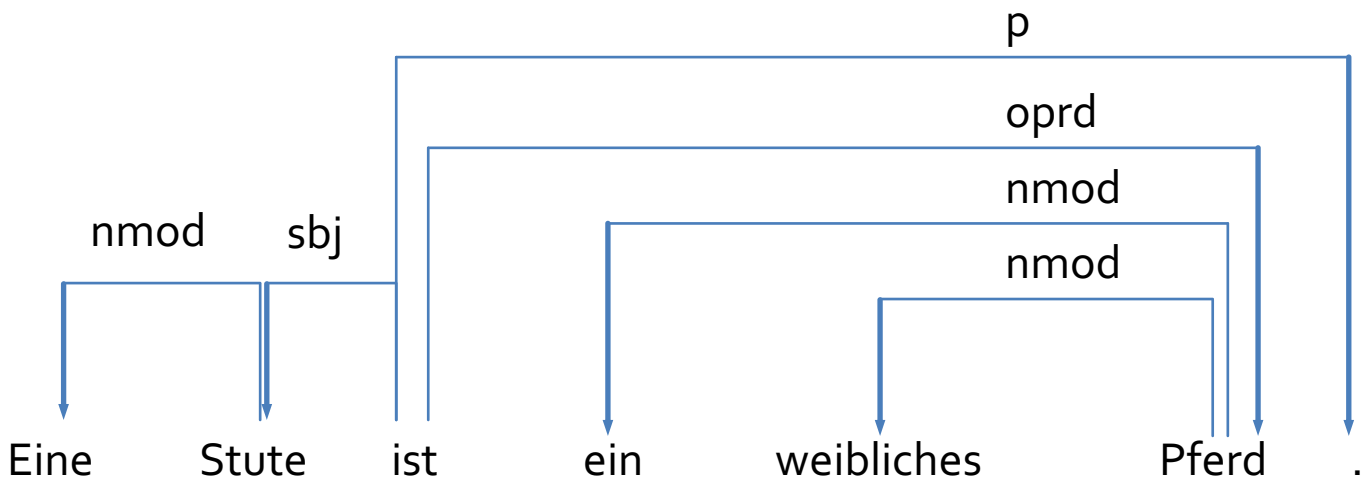
Eine Stute ist ein weibliches Pferd.



(nach Kübler & Zinsmeister 2015: 57)

Konstituentenstruktur und Dependenzstruktur

- Dependenzstruktur: Verhältnis zwischen Wortpaaren (**Kopf** und **Dependent**)



(nach Kübler & Zinsmeister 2015: 58)

Konstituentenstruktur und Dependenzstruktur

- beide Analysen gehen von hierarchischer Strukturierung von Sätzen aus
- aber: in der Konstituentenstrukturanalyse werden **Konstituenten** (also abstrakte Einheiten) hierarchisch geordnet
- die Dependenzstrukturanalyse hingegen beschränkt sich auf die **Wörter** selbst.
- Konstituentenstruktur bildet syntaktische **Kategorien** ab, Dependenzstruktur bezieht auch syntaktische **Funktionen** mit ein.

(s. Lemnitzer & Zinsmeister 2015)

Hybride Modelle

- in vielen Projekten entscheidet man sich für eine Mischung aus beiden Varianten
- Vorteil: Man kann syntaktische Kategorien **und** syntaktische Funktionen mit einbeziehen.

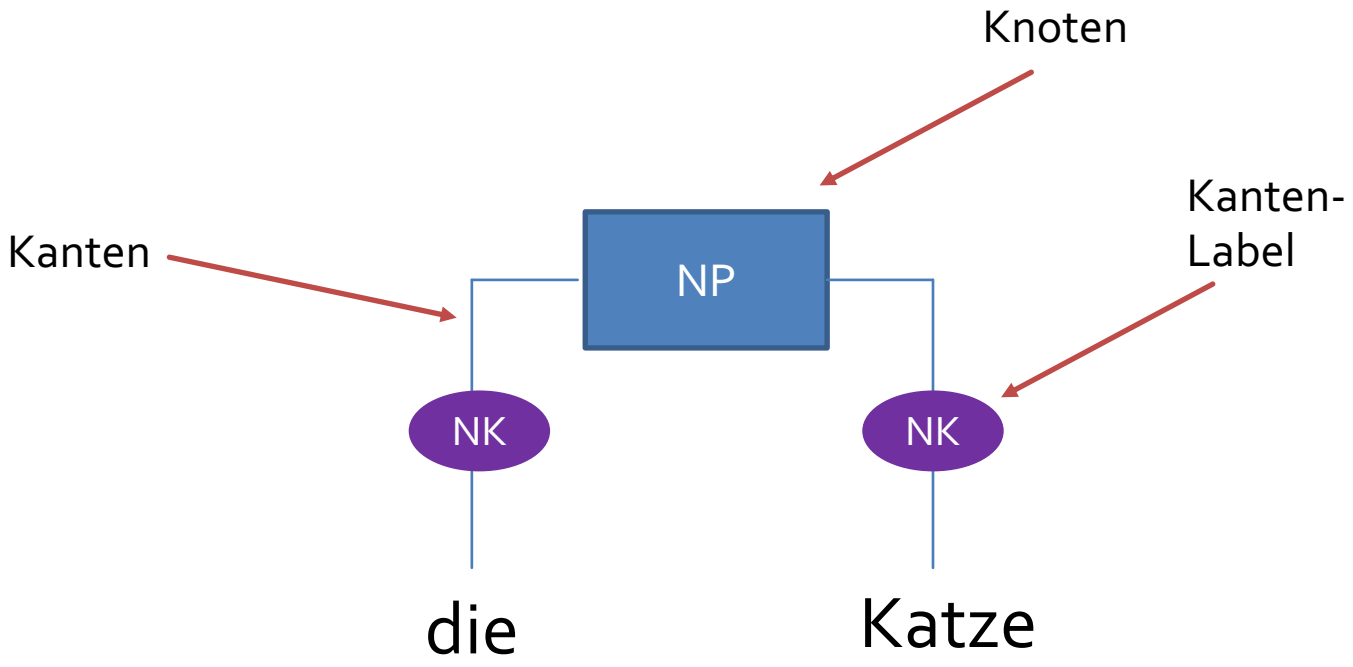
Terminologie

- Konstituente ist kategorieneutrale Beschreibung einer Phrase
- Eine Phrase ist immer einer bestimmten Kategorie zugeordnet, z.B. VP, NP
- Außer mit Phrasen arbeitet man in der syntaktischen Anlayse häufig auch mit **Chunks**
- Chunks sind prosodische Sprecheinheiten (in gesprochener Sprache an kleinen Sprechpausen erkennbar) / zusammengehörige Mehrworteinheiten

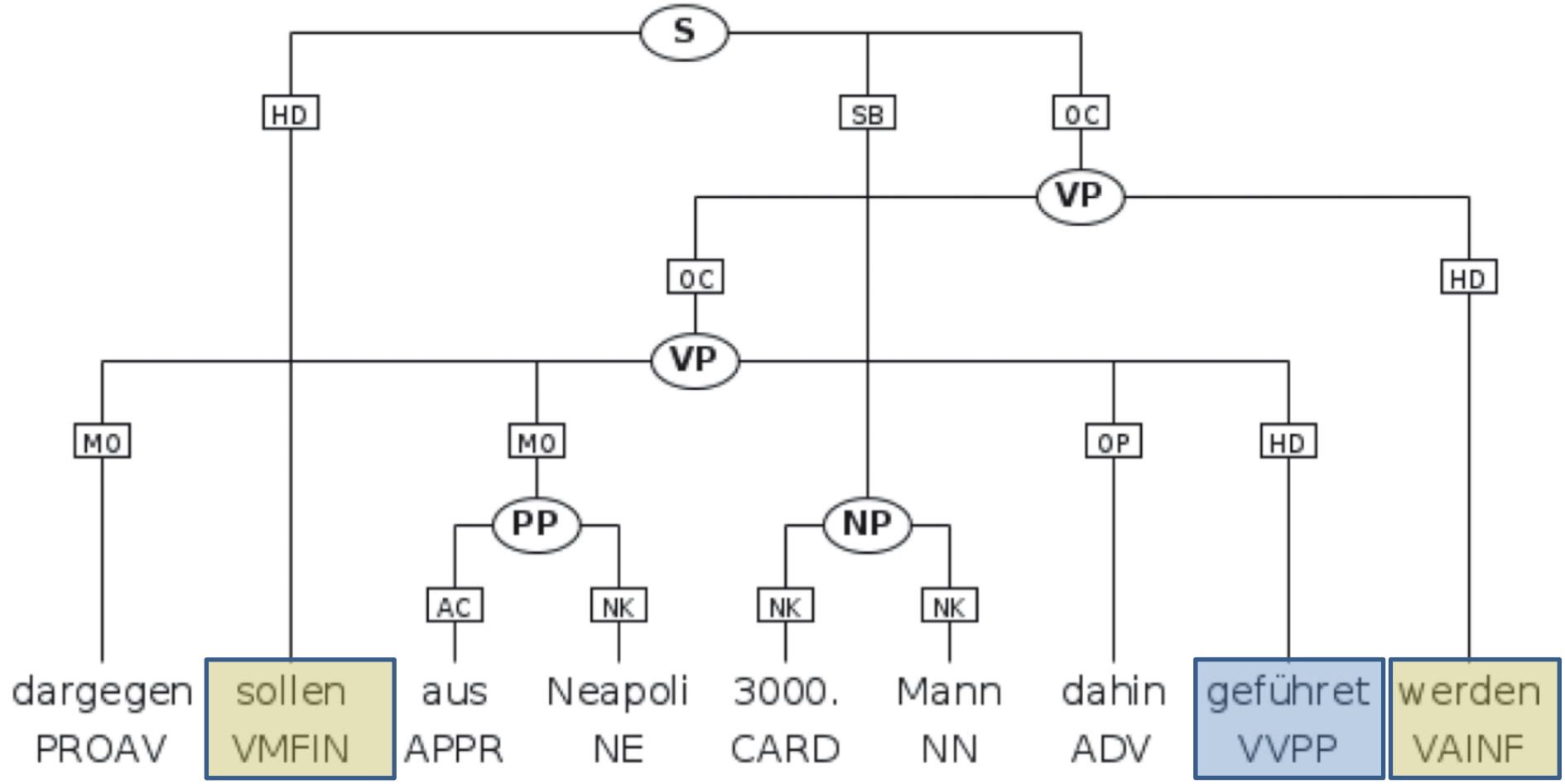
Historische Baumbanken des Deutschen

- Mercurius-Korpus
  - Zeitungstexte aus zwei Jahrgängen: *Annus Christi* 1597, *Mercurius* 1667
- Deutsche Diachrone Baumbank
  - je ~2500 Tokens aus je zwei Texten fürs Ahd., Mhd., Fnhd.

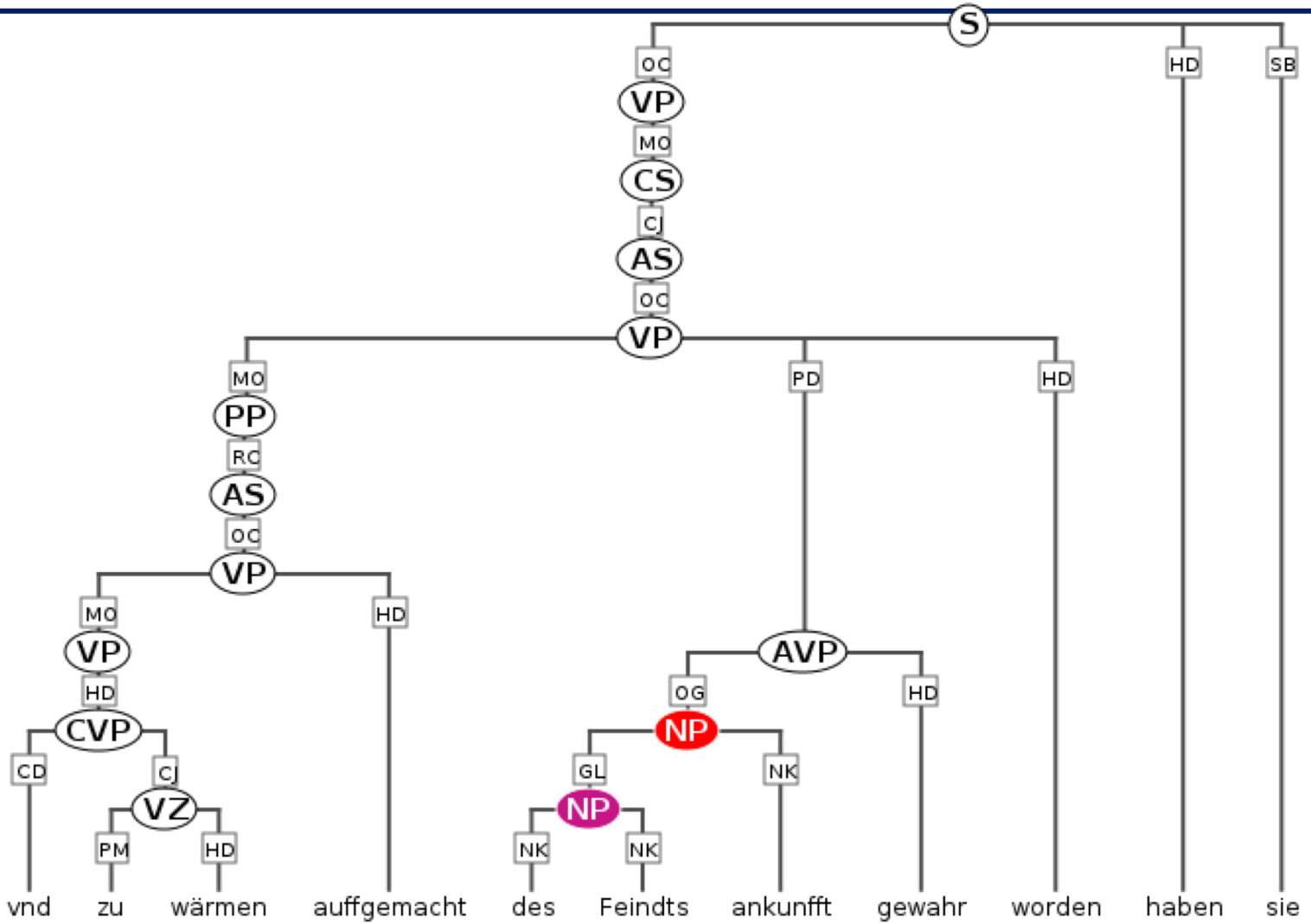
Strukturbaum: Beispiel



Kreuzende Kanten



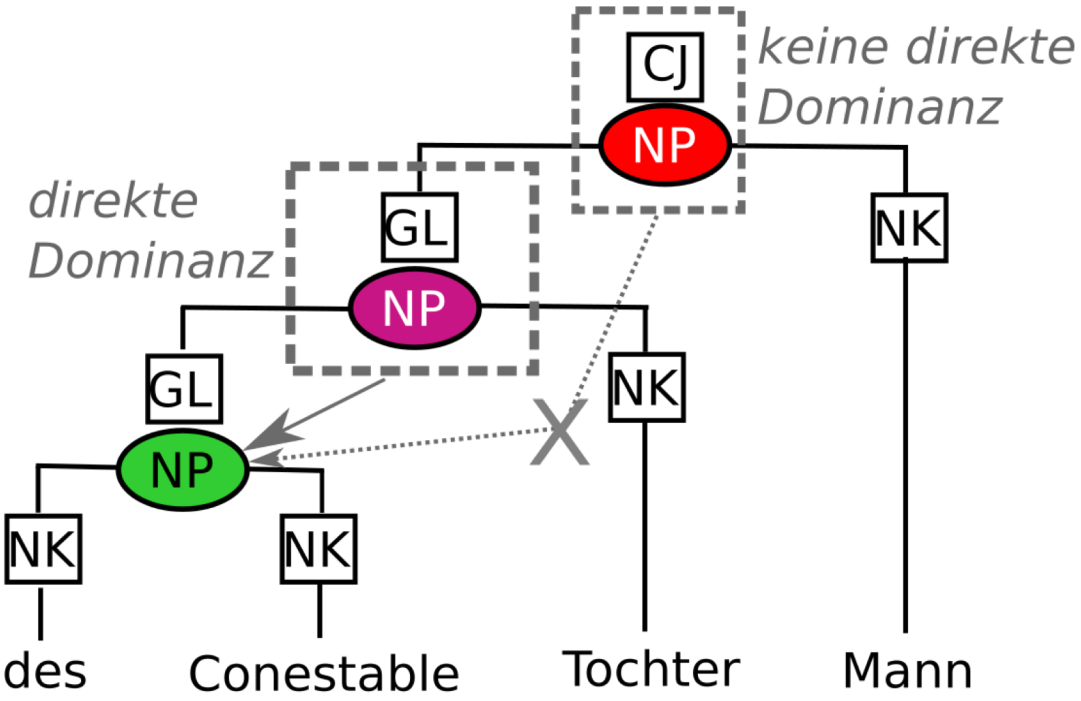
Bsp.: Mercurius-Baumbank



Bsp: Mercurius-Baumbank

- node > [label="GL"] cat="NP"

beliebiger Knoten | direkte Dominanz | Label der Kante | Nominalphrase



Bsp.: Mercurius-Baumbank

Was findet man mit den folgenden Anfragen?

- node > cat="NP",
- node >[label="GL"] cat="NP" & meta::doc="Mercurius-1667"
- node >* cat="NP"