

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

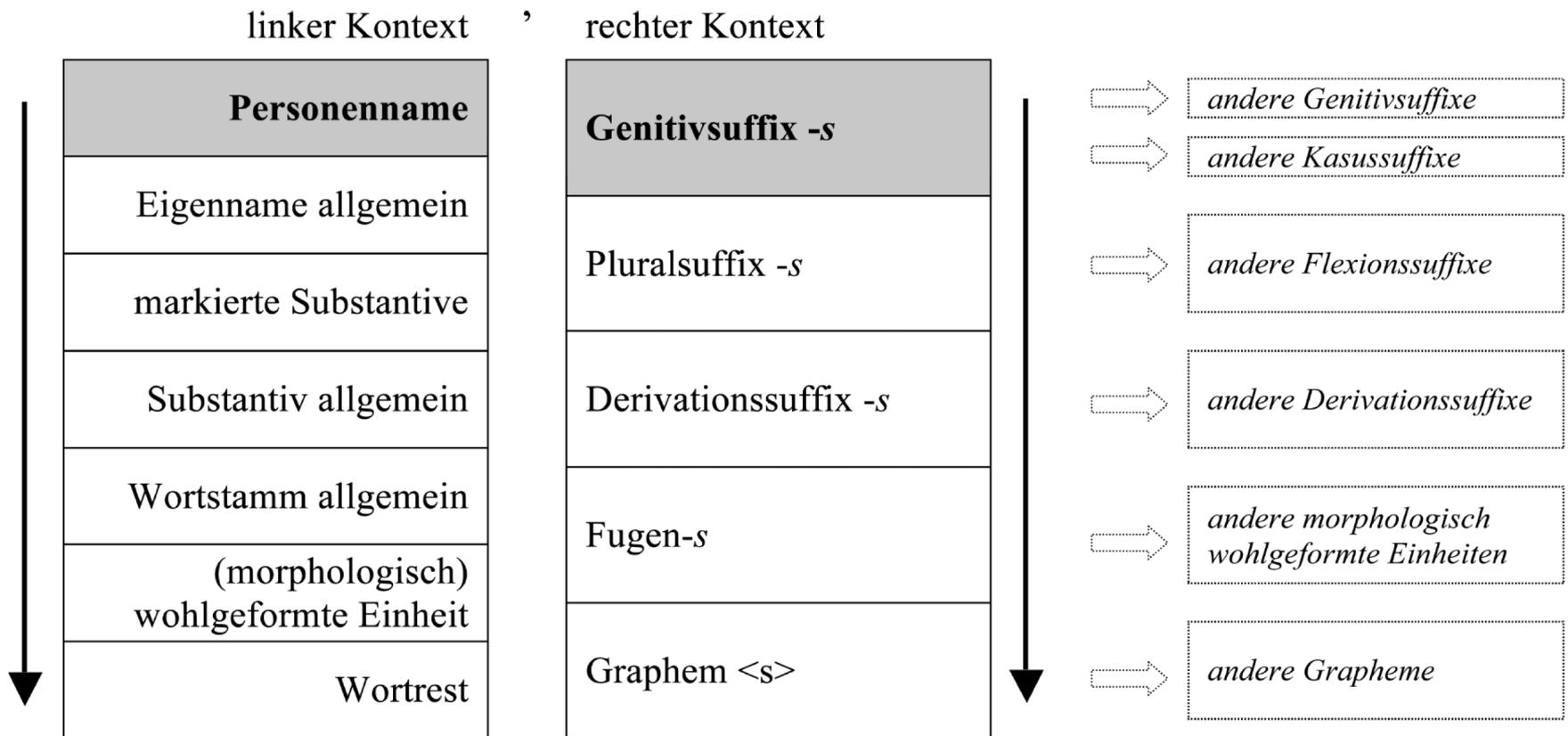
Ergebnis

- 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	Wortstamm	Konfix	Affix
Merkmale				
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: **Riesenkrach** (nicht *'Krach eines Riesen'), **Laubwerk** (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*

Beispiele (aus Wortwarte)

- "Schnell war von " **Guacamole-Gate** " die Rede . Die Debatte nahm beinahe Lorient'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

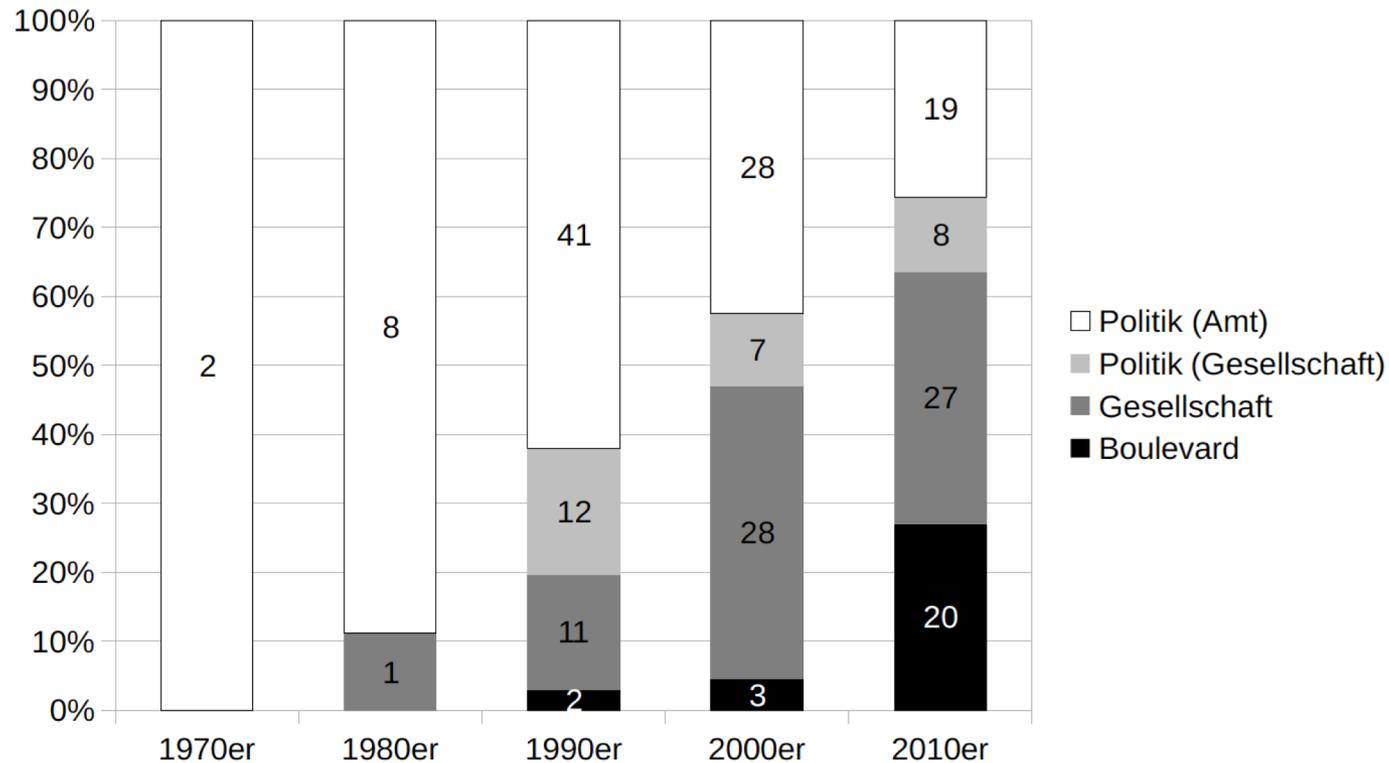


Abb. 5a. Deutsche Erstbelege (Entlehnungen und dt. Bildungen) in DeReKo/ZEIT nach Skandalfeld (n = 217).

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. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Traugott, Eliabeth 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/wisner-a-moderate-kind-a-sorta-and-thats-why-dems-should-vote-low>)
- **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
würde bey ihnen gespürt (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

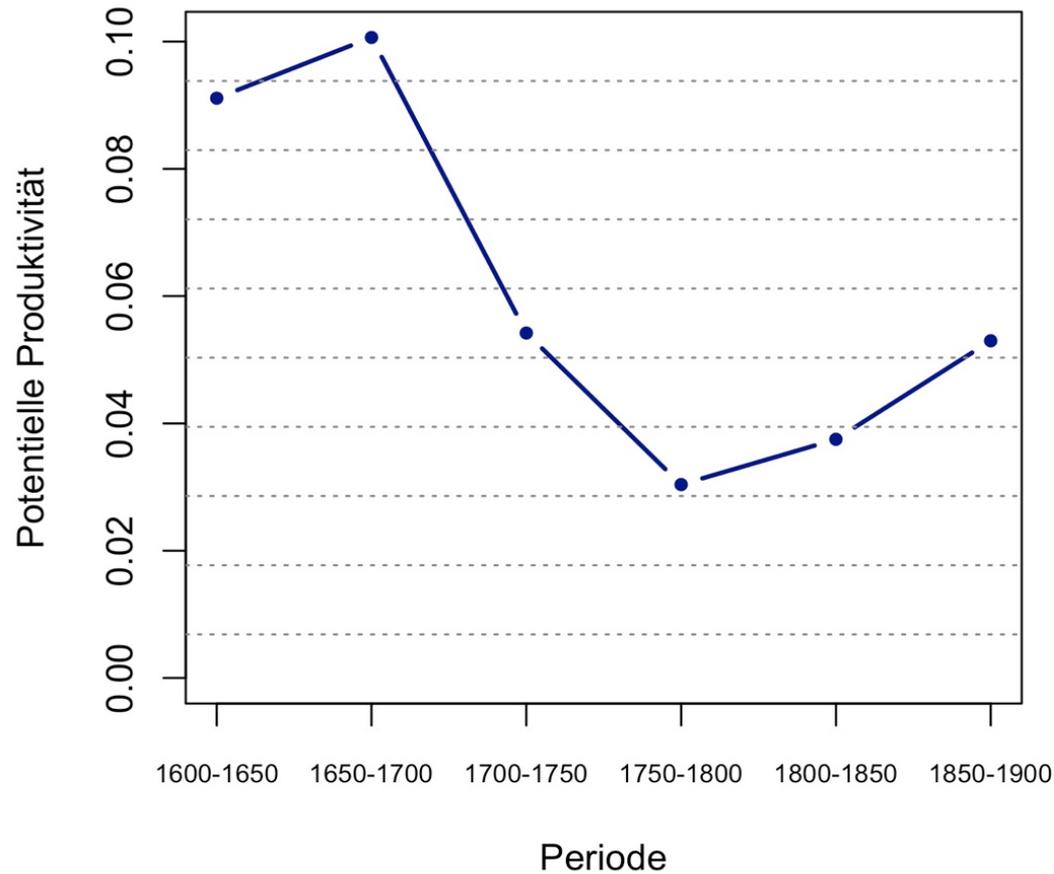
- Einfache Formel:

Anzahl der Hapaxe, die zum
Wortbildungsmuster gehören
(z.B.: alle Hapaxe auf -heit/-keit)

Gesamtzahl der Belege, die zum
Wortbildungsmuster gehören
(z.B.: alle Belege auf -heit/-keit)

Beispiel *ung*-Nominalisierung

Potentielle Produktivität, DTAbaby



Beispiel Fugenelemente

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

Beispiel Fugenelemente

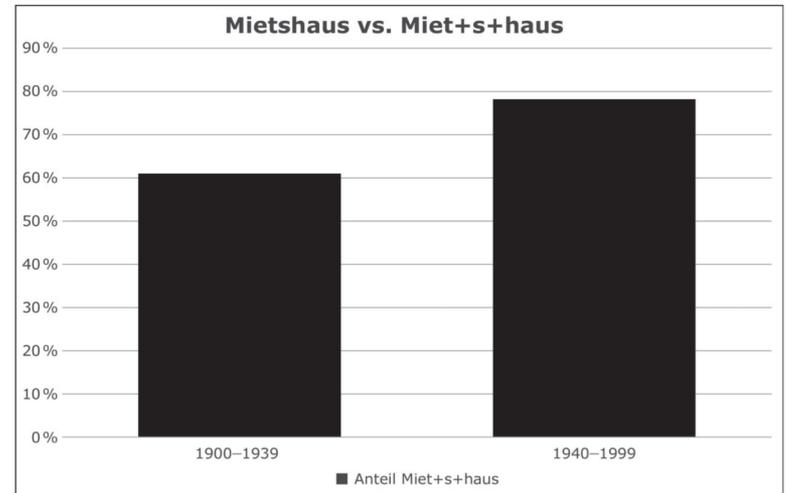
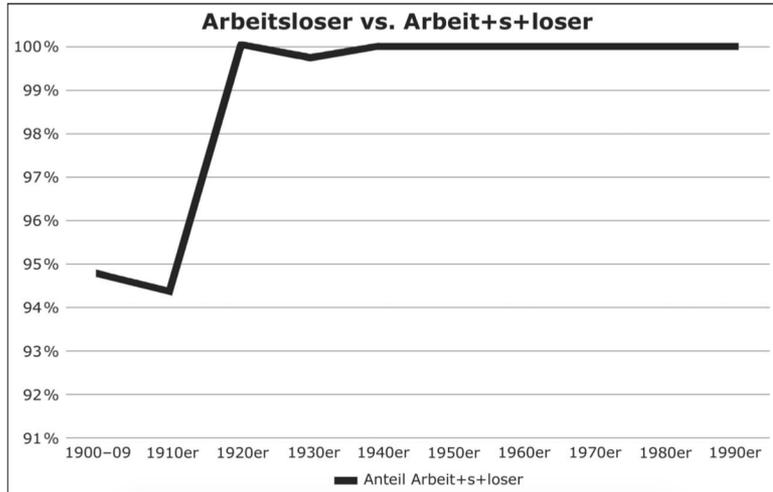


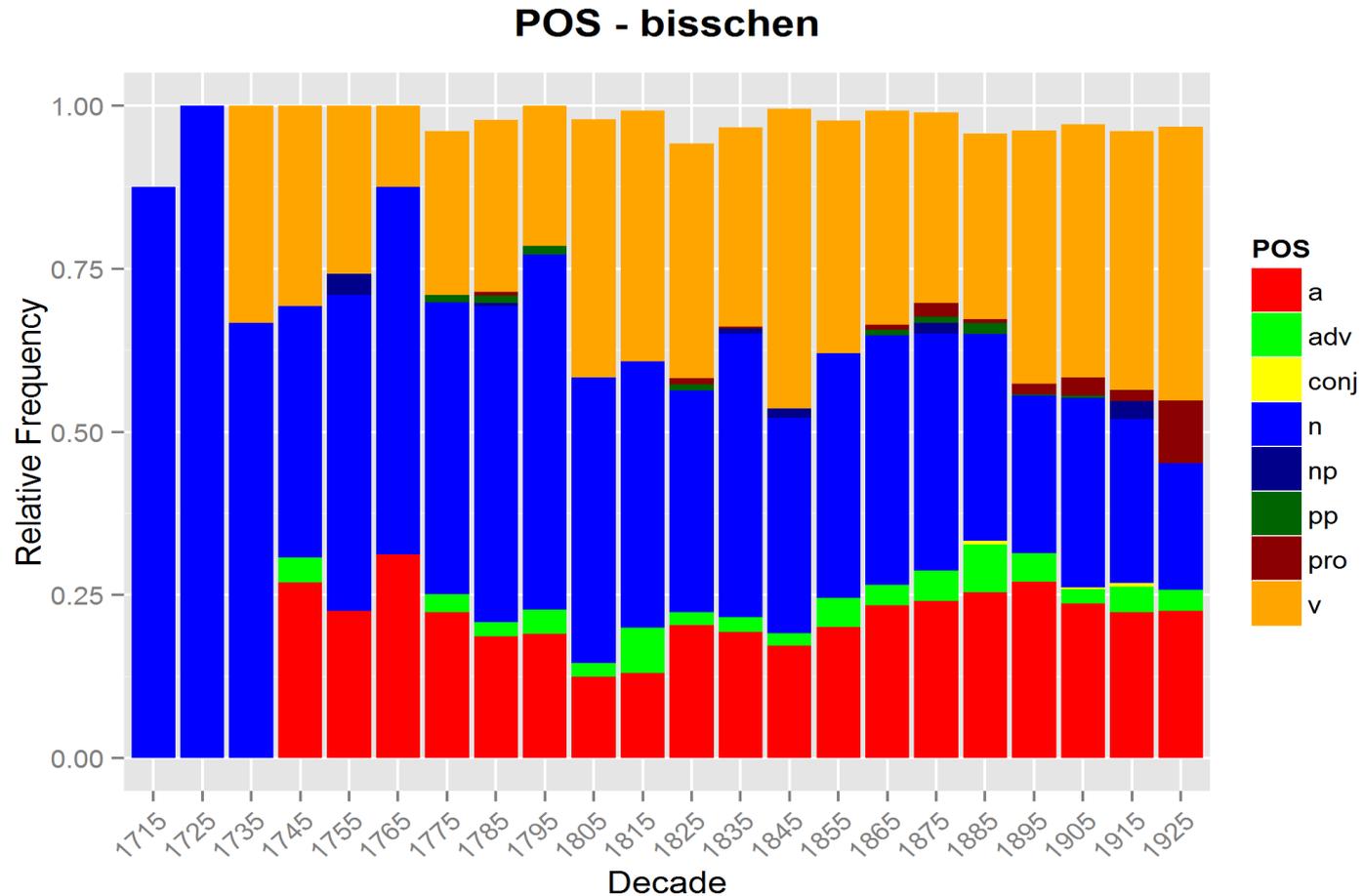
Tabelle 6: Die Häufigkeit des Fugen-s nach präfigierten Erstgliedern.

	Erstglied enthält:	
	unbetontes Präfix (Typ <i>Beruf-</i>)	betontes Präfix (Typ <i>Anruf-</i>)
Tokens:	85 % (von 495.887 Komposita)	36 % (von 324.503 Komposita)
Types:	82 % (von 17.999 Komposita)	37 % (von 11.325 Komposita)

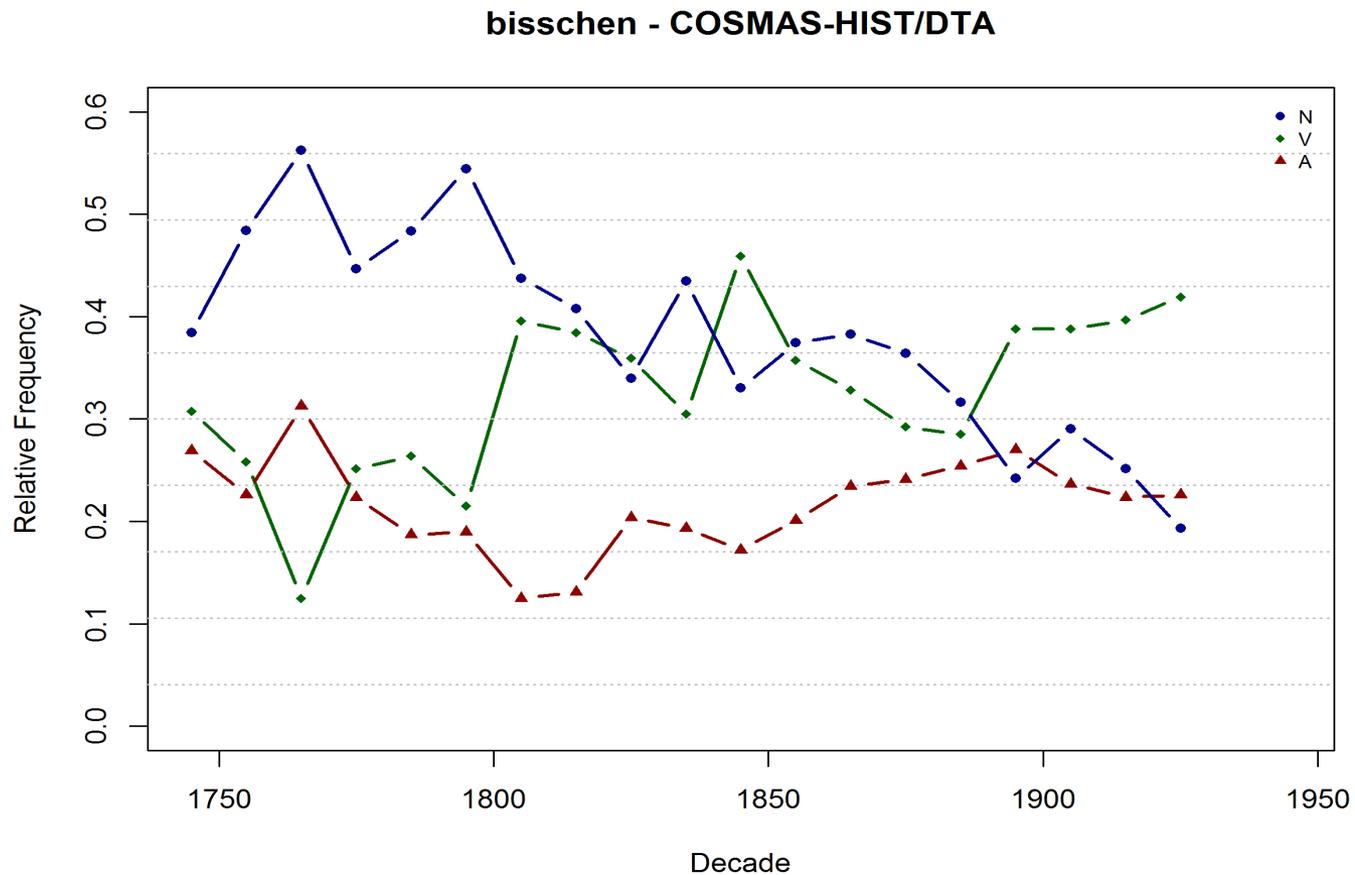
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)

- 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

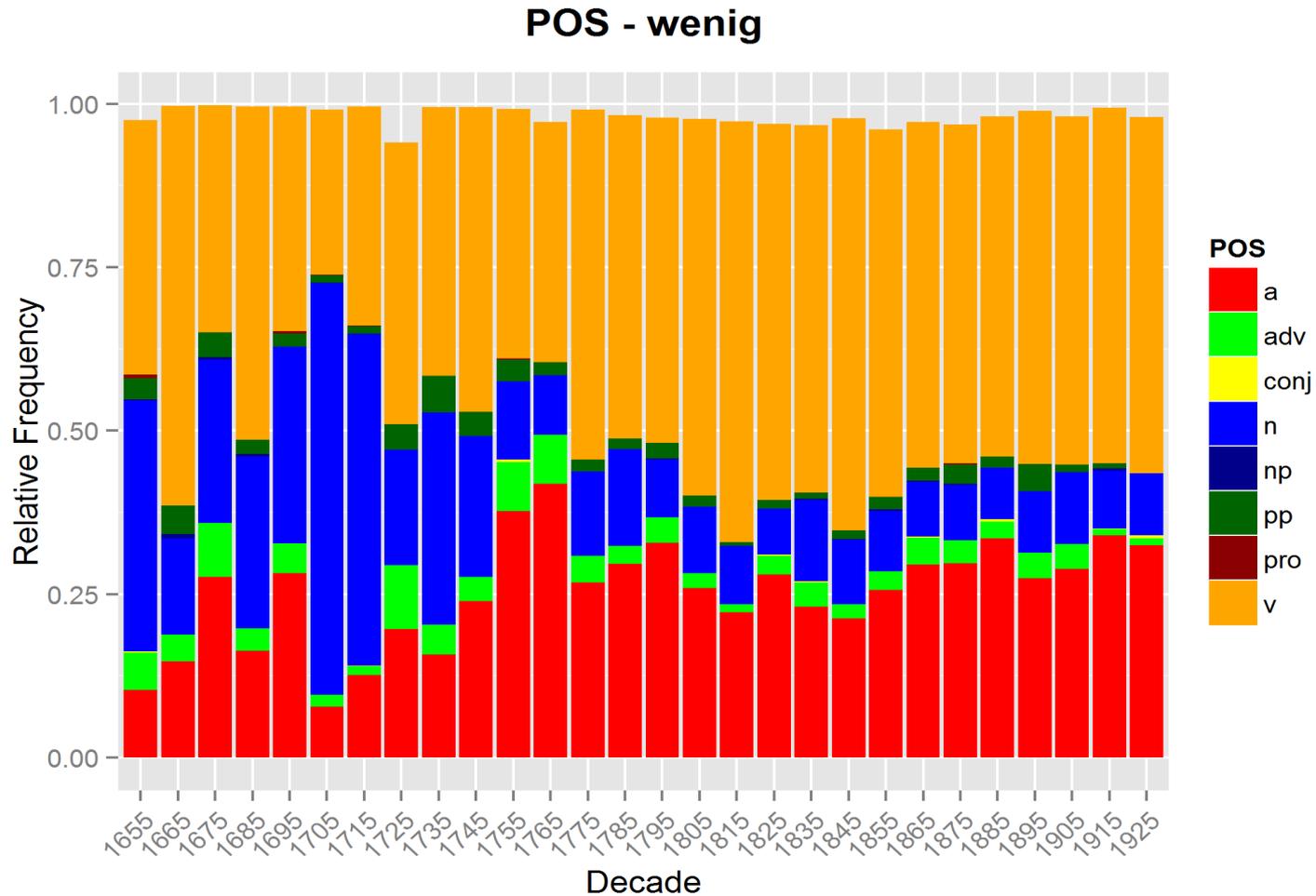
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



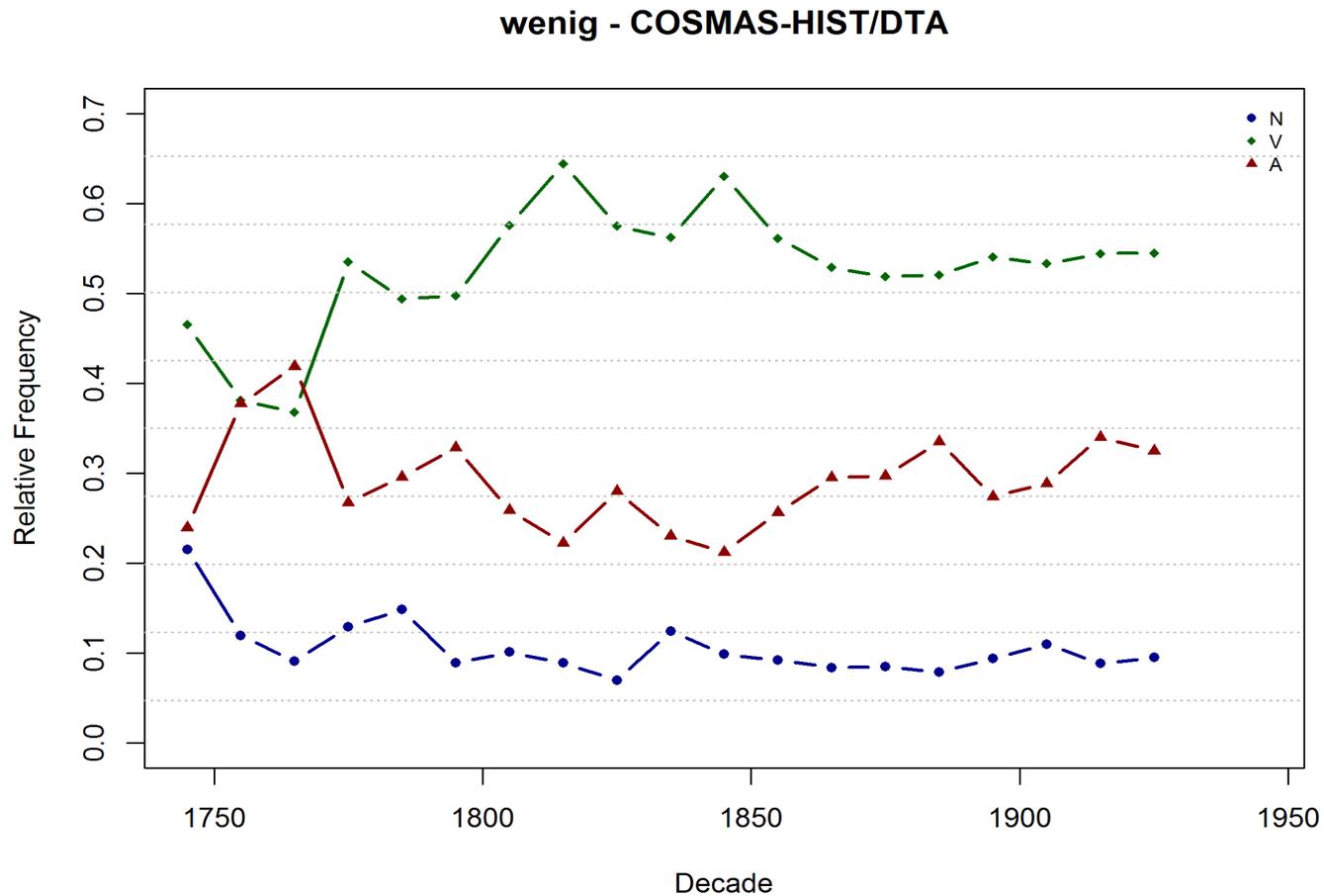
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



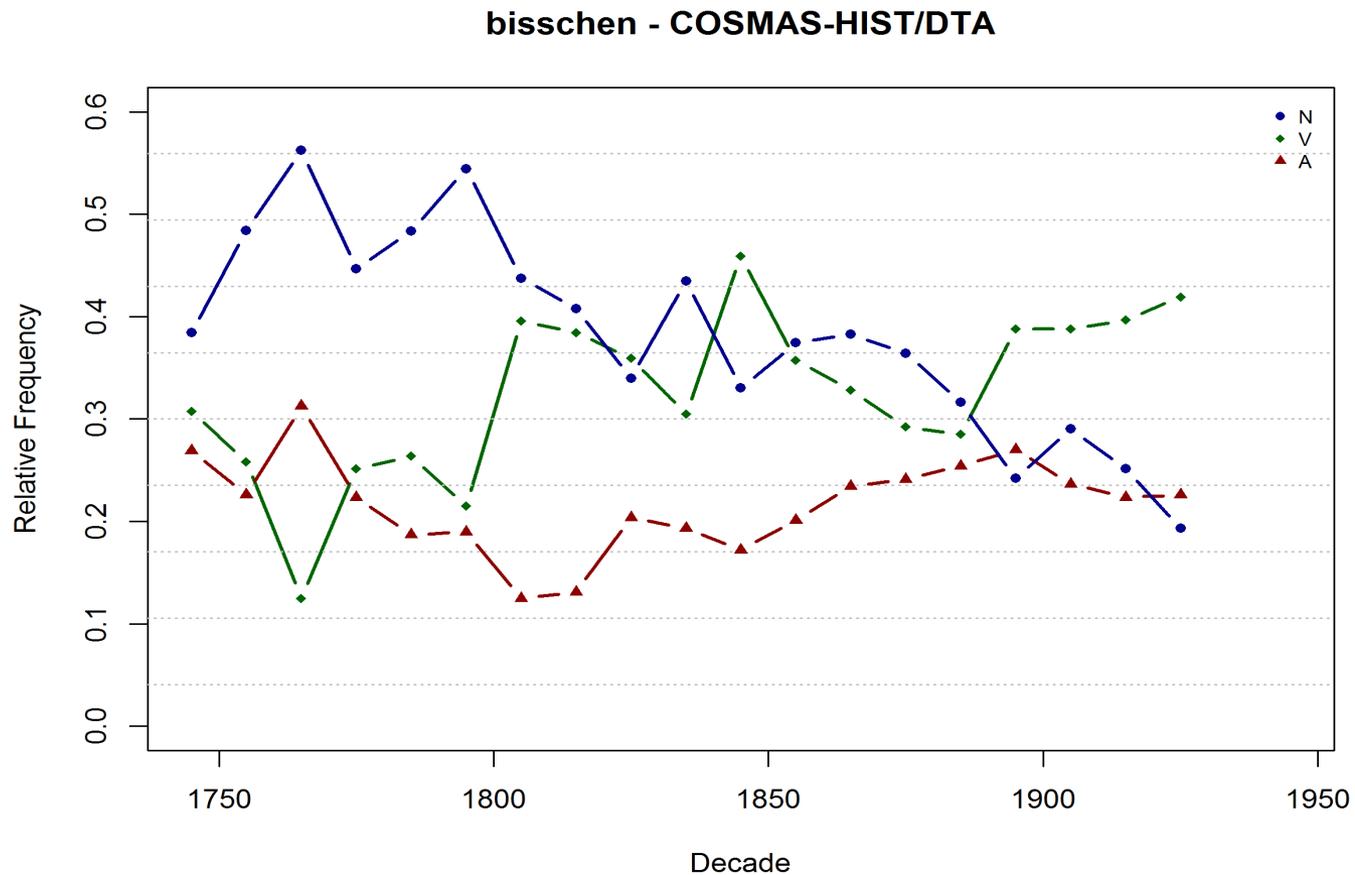
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



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

Anmelden (DTAQ)

D T A

in den Titeldaten im Korpus in der Dokumentation [Hilfe](#)

Beispielanfragen: `$con=/cit/ #has[dtadir,'mendelssohn_jerusalem_1783'] ehelichen with $p=VVINF $l=Erkenntnis`

Deutsches Textarchiv

GRUNDLAGE FÜR EIN REFERENZKORPUS DER NEUHOCHDEUTSCHEN SPRACHE

Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern erstellten und ausführlich kommentierten, umfangreichen Bibliographie. In Ergänzung wurden einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Aus der Gesamtliste der auf diesem Wege ermittelten Titel wurde von der DTA-Projektgruppe ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus zusammengestellt (weitere Informationen zur Textauswahl).

Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die

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">Caï&#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">Regulin</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="wo05a"></lemma>
<lemma tokenIDs="wo05b">jedoch</lemma>
<lemma tokenIDs="wo05c">aber</lemma>
<lemma tokenIDs="wo05d">d</lemma>
<lemma tokenIDs="wo05e">beide</lemma>
<lemma tokenIDs="wo05f"></lemma>
<lemma tokenIDs="wo060">d</lemma>
<lemma tokenIDs="wo061">äußerlich</lemma>
<lemma tokenIDs="wo062">und</lemma>
<lemma tokenIDs="wo063">verbergen</lemma>
<lemma tokenIDs="wo064">Verstand</lemma>
<lemma tokenIDs="wo065">sich</lemma>
<lemma tokenIDs="wo066">in</lemma>
<lemma tokenIDs="wo067">d</lemma>
<lemma tokenIDs="wo068">Kontext</lemma>
<lemma tokenIDs="wo069">geschichte</lemma>
<lemma tokenIDs="wo06a">erweisen</lemma>
<lemma tokenIDs="wo06b">mögen</lemma>
<lemma tokenIDs="wo06c"></lemma>
<lemma tokenIDs="wo06d">damit</lemma>
<lemma tokenIDs="wo06e">beide</lemma>
<lemma tokenIDs="wo06f">d</lemma>
<lemma tokenIDs="wo070">Geheimnis</lemma>
<lemma tokenIDs="wo071">nicht</lemma>
<lemma tokenIDs="wo072">merken</lemma>
<lemma tokenIDs="wo073"></lemma>
<lemma tokenIDs="wo074">und</lemma>
<lemma tokenIDs="wo075">doch</lemma>
<lemma tokenIDs="wo076">auch</lemma>
<lemma tokenIDs="wo077">verstehen</lemma>
<lemma tokenIDs="wo078">werden</lemma>
<lemma tokenIDs="wo079">.</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) Wortform *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)

- alternatives 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

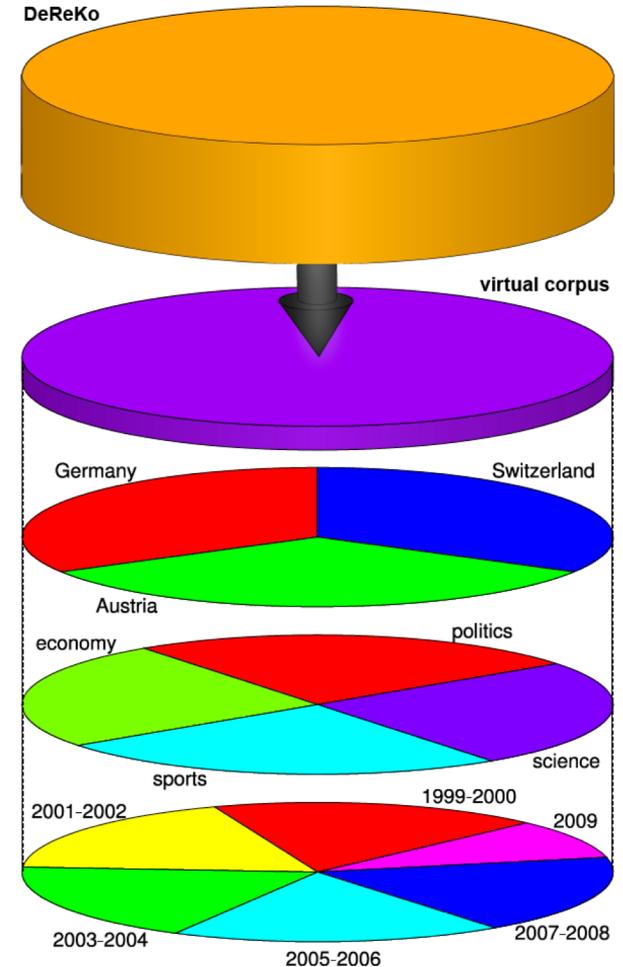


Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

(aus Kupietz et al. 2010)

Virtuelle Korpora erstellen

W - Archiv der geschriebenen Sprache Aktuelles Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W (mit Neuakquisitionen) [1]

Korpusdefinition

Korpus editieren

Gesamtkorpus	CorpDefID:	Korpus:
<p>A97/APR St. Galler Tagblatt, April 1997 A97/MAI St. Galler Tagblatt, Mai 1997 A97/JUN St. Galler Tagblatt, Juni 1997 A97/JUL St. Galler Tagblatt, Juli 1997 A97/AUG St. Galler Tagblatt, August 1997 A97/SEP St. Galler Tagblatt, September 1997 A97/OKT St. Galler Tagblatt, Oktober 1997 A97/NOV St. Galler Tagblatt, November 1997 A97/DEZ St. Galler Tagblatt, Dezember 1997 A98/JAN St. Galler Tagblatt, Januar 1998 A98/FEB St. Galler Tagblatt, Februar 1998 A98/MAR St. Galler Tagblatt, März 1998 A98/APR St. Galler Tagblatt, April 1998 A98/MAI St. Galler Tagblatt, Mai 1998 A98/JUN St. Galler Tagblatt, Juni 1998 A98/JUL St. Galler Tagblatt, Juli 1998 A98/AUG St. Galler Tagblatt, August 1998 A98/SEP St. Galler Tagblatt, September 1998 A98/OKT St. Galler Tagblatt, Oktober 1998 A98/NOV St. Galler Tagblatt, November 1998 A98/DEZ St. Galler Tagblatt, Dezember 1998 A99/JAN St. Galler Tagblatt, Januar 1999 A99/FEB St. Galler Tagblatt, Februar 1999</p>	<p style="text-align: center;">Suchmuster</p> <p>2015 Hilfe</p> <p><input type="checkbox"/> linksbündig <input type="checkbox"/> rechtsbündig</p> <p><input type="checkbox"/> Groß-/ Kleinschreibung beachten</p> <p style="text-align: center;"><input checked="" type="radio"/> << <input type="radio"/> >></p> <p style="text-align: center;"><input type="button" value="Suchen"/></p> <p style="text-align: center;"><input type="button" value="Mark. löschen"/></p> <p style="text-align: center;">→</p> <p style="text-align: center;">←</p>	<p>VDI15/JUN VDI nachrichten, Juni 2015 VDI15/JUL VDI nachrichten, Juli 2015 WWO15/JAN Weltwoche, Januar 2015 WWO15/FEB Weltwoche, Februar 2015 WWO15/MAR Weltwoche, März 2015 WWO15/APR Weltwoche, April 2015 WWO15/MAI Weltwoche, Mai 2015 WWO15/JUN Weltwoche, Juni 2015 WWO15/JUL Weltwoche, Juli 2015 Z15/JAN Die ZEIT, Januar 2015 Z15/FEB Die ZEIT, Februar 2015 Z15/MAR Die ZEIT, März 2015 Z15/APR Die ZEIT, April 2015 Z15/MAI Die ZEIT, Mai 2015 Z15/JUN Die ZEIT, Juni 2015 Z15/JUL Die ZEIT, Juli 2015 ZCA15/FEB Zeit Campus, Februar 2015 ZCA15/MAR Zeit Campus, März 2015 ZCA15/APR Zeit Campus, April 2015 ZGE15/FEB Zeit Geschichte, Februar 2015 ZGE15/MAI Zeit Geschichte, Mai 2015 ZWI15/FEB Zeit Wissen, Februar 2015 ZWI15/APR Zeit Wissen, April 2015 ZWI15/JUN Zeit Wissen, Juni 2015</p>
10555 Dokumente, 0 selektiert		164 Dokumente, 0 selektiert

Als neue Korpusdefinition übernehmen

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.

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 getaggtten 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

ANNIS interface showing search results and linguistic visualizations.

Query result (red box) points to the search input field containing "cat='NP'".

Visualizations (red box) points to the visualization options: dependencies (arches), information structure (grid), discourse referents (grid), and constituents (tree).

The main interface displays search results for "cat='NP'". The search input field contains "cat='NP'". The search results show 41 matches in 2 documents. The corpus list shows the following data:

Name	Texts	Tokens
pcc2	2	399
RIDGES_Herbology_\	22	122.698
RIDGES_Herbology_\	29	154.266
Ridges_Herbology_Ve	13	60.811

The search results table shows the following text and annotations:

Text	Token	Annotation
Feigenblatt	Die	Jugendlichen
Feigenblatt	der	jugendliche
Nom.Sg.Neut	Nom.Pl.*	Nom.Pl.*
NN	ART	NN

The visualization options are:

- dependencies (arches)
- information structure (grid)
- discourse referents (grid)
- constituents (tree)

The constituent tree diagram shows the following structure:

```
graph TD
    S((S)) --- SB[SB]
    S --- HD[HD]
    S --- OA[OA]
    SB --- NK1[NK]
    SB --- NK2[NK]
    SB --- MNR[MNR]
    NK1 --- Die[Die]
    NK2 --- Jugendlichen[Jugendlichen]
    MNR --- PP((PP))
    PP --- AC[AC]
    PP --- NK3[NK]
    AC --- in[in]
    NK3 --- Zossen[Zossen]
    HD --- wollen[wollen]
    OA --- NP1[NP]
    NP1 --- NK4[NK]
    NP1 --- NK5[NK]
    NK4 --- ein[ein]
    NK5 --- Musikcafe[Musikcafé]
    OA --- S2((S))
    S2 --- OA2[OA]
    OA2 --- Das[Das]
```

ANNIS

zu den über ANNIS verfügbaren Korpora gehören z.B.

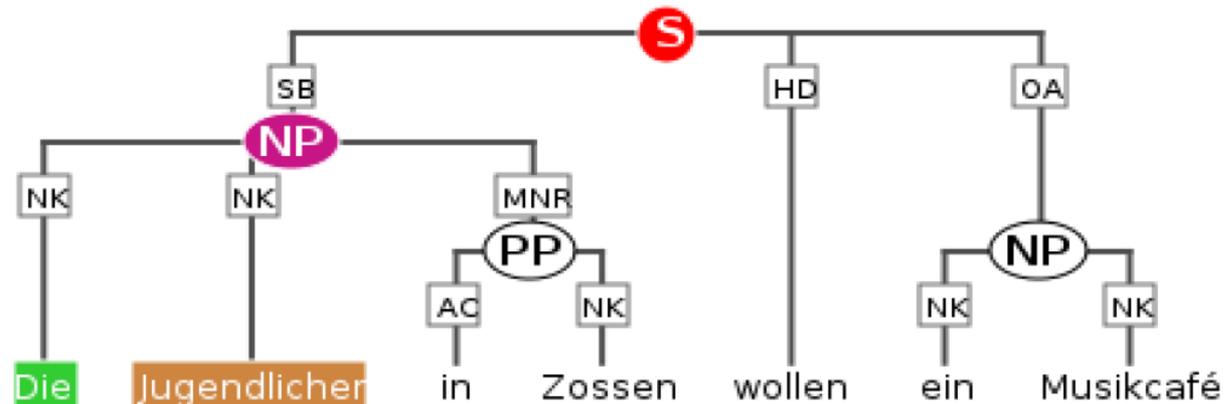
- Deutsch Diachron Digital (im Entstehen)
- RIDGES (historische Kräutertexte)
- Potsdamer Kommentarkorpus (Zetungskommentare)
- KiezDeutschKorpus
- seit kurzem: Bonner Frühneuhochdeutschkorpus

ANNIS Query Language (AQL)

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

- 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

ANNIS Query Language

Eingabefeld

Please enter AQL query

Query Builder

Q Search More History

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

Hilfe/Tutorial

Help/Examples

Tutorial

Example Queries

Example Query	Description	open corpus browser
Q ""	search for the word ""	mo
Q ""	search for the word ""	mu

Korpora

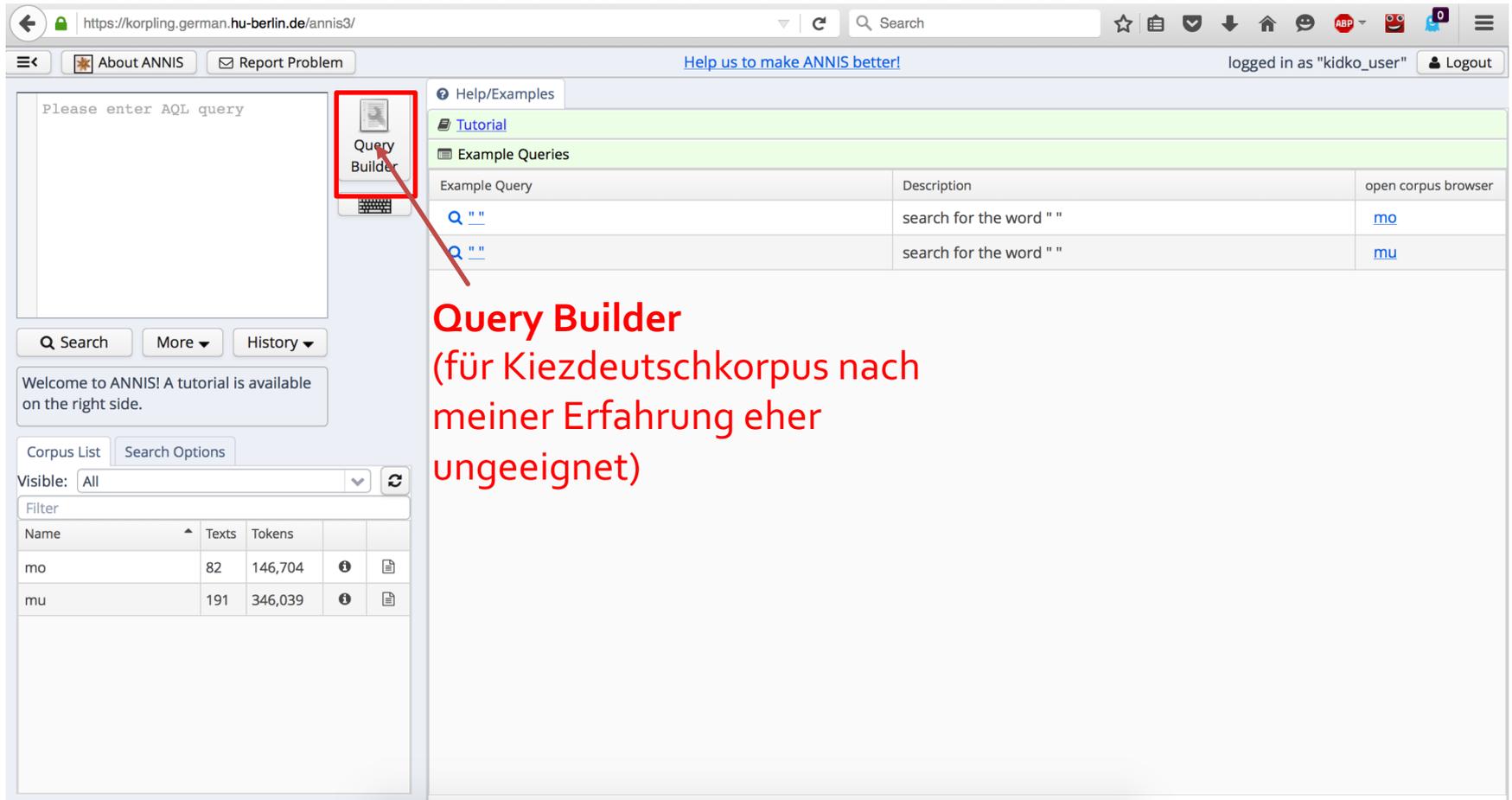
Corpus List Search Options

Visible: All

Filter

Name	Texts	Tokens		
mo	82	146,704	i	d
mu	191	346,039	i	d

ANNIS Query Language



The screenshot shows the ANNIS web interface. The browser address bar displays <https://korpling.german.hu-berlin.de/annis3/>. The page includes navigation links for "About ANNIS" and "Report Problem", a "Help us to make ANNIS better!" link, and a user login status "logged in as 'kidko_user'" with a "Logout" button. The main content area is divided into several sections:

- A "Please enter AQL query" input field with "Q Search", "More", and "History" buttons below it.
- A "Welcome to ANNIS! A tutorial is available on the right side." message.
- A "Corpus List" section with "Search Options" and a "Visible: All" dropdown.
- A table with columns "Name", "Texts", and "Tokens" showing data for corpora "mo" and "mu".
- A "Help/Examples" sidebar containing a "Tutorial" link and an "Example Queries" table.

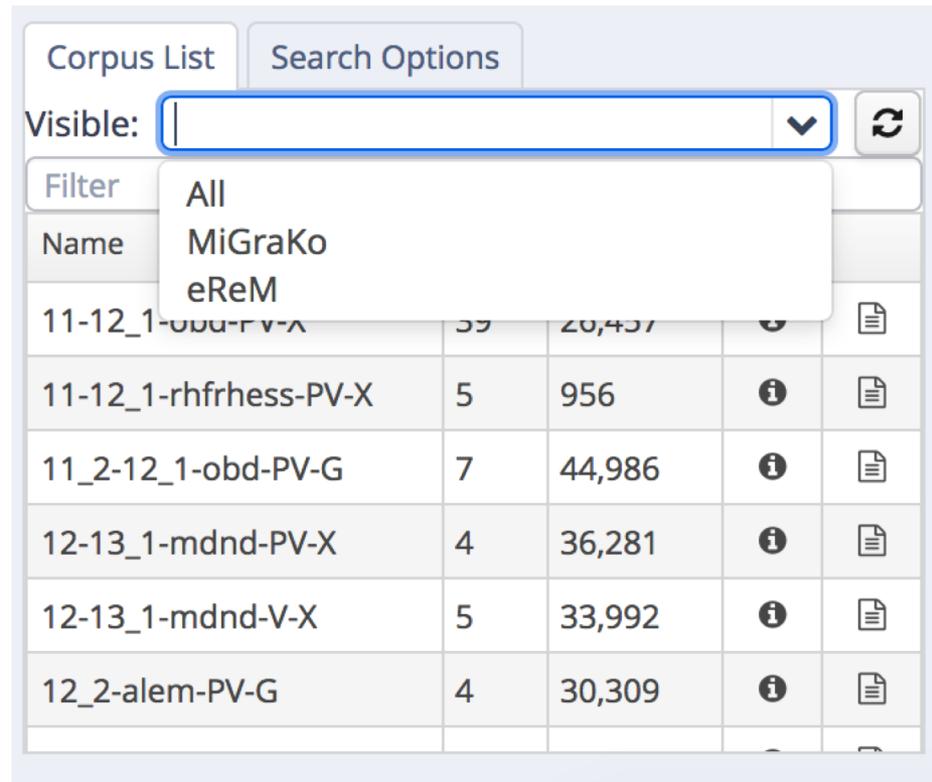
The "Query Builder" button, located in the sidebar, is highlighted with a red box and a red arrow pointing to it. The text "Query Builder" is written in red below the button.

Example Query	Description	open corpus browser
Q ""	search for the word ""	mo
Q ""	search for the word ""	mu

Name	Texts	Tokens		
mo	82	146,704	i	d
mu	191	346,039	i	d

Query Builder
(für Kiezdeutschkorpus nach
meiner Erfahrung eher
ungeeignet)

REM: Ein Korpus wählen



The screenshot shows a software interface with two tabs: "Corpus List" and "Search Options". The "Corpus List" tab is active. At the top, there is a "Visible:" label followed by a dropdown menu and a refresh icon. The dropdown menu is open, showing three options: "All", "MiGraKo", and "eReM". Below the dropdown is a table with columns for "Filter", "Name", and other data. The table contains several rows of corpus information.

Filter	Name			
	11-12_1-obd-PV-X	59	20,457	
	11-12_1-rhfrhess-PV-X	5	956	
	11_2-12_1-obd-PV-G	7	44,986	
	12-13_1-mdnd-PV-X	4	36,281	
	12-13_1-mdnd-V-X	5	33,992	
	12_2-alem-PV-G	4	30,309	

- 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

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	g	
tok_anno	.	der	vbel	hellevvarte	.	er	n	vverde	ir	g
norm		der	übel	hellewarte		er	ne	verde	ire	g
tokenization						MS1	MS2			
pos	\$_	DDART	ADJA	NA	\$_	PPER	PTKNEG	VAFIN	DPOSA	N
posLemma	\$_	DD	ADJ	NA	\$_	PPER	PTK	VA	DPOS	N
lemma		dër	übel	hëlle-warte		ër	ne	wërden	ir(e)	g
lemmald		29817000	174642000	71433000		40380000	119841000	225138000	83982000	5
lemmaLemma		dër	übel	hëlle-warte		ër	ne	wërden	ir(e)	g
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

The screenshot shows the ANNIS search interface. The search query is `n="schwul" . POS=/AD.*/`. The results show two matches in two documents. The first result is from document `mo` (tokens 1920-1931) and the second is from document `mu` (tokens 165-176). The token grids show the following words:

Document	Token	Word
1 (mo)	1920	WACHS
	1921	nur
	1922	
2 (mu)	165	
	166	
	167	
	168	
	169	
	170	
	171	
	172	walLAH
	173	er
	174	soll

The interface also includes a search bar, a corpus list, and a detailed view of search results with token grids and annotations.

Wortfolgen suchen

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

The screenshot displays a search interface with a query builder on the left and search results on the right. The query is `n="schwul" .2,4 POS=/AD.*/`. The results show five matches, each with a path, document name, and a grid of context words. The grid highlights the word 'schwul' in the original context.

Query Builder: `n="schwul" .2,4 POS=/AD.*/`

Search Results:

- 1** Path: mo > Mo05WD_11-2 (tokens 1920 - 1933) left context: 5 right context: 5
SPK15::v WACHS nur
SPK39::v a sein kle:ner WILli da unten is kle:n (-)
SPK67::v ich bin SCHWUL
- 2** Path: mo > Mo05WD_11-2 (tokens 1920 - 1934) left context: 5 right context: 5
SPK15::v WACHS nur
SPK39::v a sein kle:ner WILli da unten is kle:n (-)
SPK67::v ich bin SCHWUL ich
- 3** Path: mo > Mo12MD_07 (tokens 1531 - 1544) left context: 5 right context: 5
Mo12MD::v (-) gib mal WASSer bitte (-)
SPK102::v bundeswehr IS nich schwul FLASchenbong flasch
SPK103::v (unverständlich) zigaRETten
- 4** Path: mu > MuH12MD_02 (tokens 1948 - 1961) left context: 5 right context: 5
MuH12MD::v sind doch kaum GAR nich (-)
SPK102::v zigaretten (-) ist das SCHWUL
- 5** Path: mu > MuH12MD_02 (tokens 1948 - 1962) left context: 5 right context: 5
MuH12MD::v sind doch kaum GAR nich (-)

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

The screenshot displays the ANIS search interface. On the left, the 'Query Builder' shows the search query: `tok=/[jJ]e/ . pos=/A.*/ .2,10` and `tok=/[dD]esto|[jJ]e|[Uu]mso/`. Below the query builder are buttons for 'Search', 'More', and 'History'. A notification indicates '39 matches in 36 documents'. A dropdown menu is open over the 'More' button, showing options for 'Export' and 'Frequency Analysis'. The main search results area shows a list of documents with a table of search results. The table includes columns for document ID, token, part of speech, and the surrounding context. The results are filtered to show matches for the query. The interface also includes a 'Corpus List' and 'Search Options' section at the bottom left, and a 'Visible: All' filter at the bottom center.

Query Builder

```
tok=/[jJ]e/ . pos=/A.*/ .2,10
tok=/[dD]esto|[jJ]e|[Uu]mso/
```

Search Results

Document	Token	POS	Context
FalkoEssayL2WHIGv2.0	gibt	desto	gewaltätiger wird die Gesellschaft .
FalkoEssayL2WHIGv2.0	geben	[unknown]	werden d Gesellschaft .
FalkoEssayL2WHIGv2.0	VVFIN	\$, KON	ADJD VAFIN ART NN \$.
FalkoEssayL2WHIGv2.0	mehr	gew	
FalkoEssayL2WHIGv2.0	mehr	gew	
FalkoEssayL2WHIGv2.0	ADV	ADJD	
FalkoEssayL2WHIGv2.0	argumentiert	desto	negativer wird vielleicht diese Öffentlichkeit
FalkoEssayL2WHIGv2.0	argumentieren	desto	negativ werden vielleicht dies Öffentlichkeit
FalkoEssayL2WHIGv2.0	VVPP	\$, KON	ADJD VAFIN ADV PDAT NN
FalkoEssayL2WHIGv2.0	eigenen	Materialien	beschäftigt , desto hat man größere
FalkoEssayL2WHIGv2.0	eigen	Material	beschäftigen , desto haben man groß
FalkoEssayL2WHIGv2.0	ADJA	NN	VVPP \$, KON VAFIN PIS ADJA
FalkoEssayL2WHIGv2.0	wird		
FalkoEssayL2WHIGv2.0	werden		
FalkoEssayL2WHIGv2.0	VAFIN		
FalkoEssayL2WHIGv2.0	denken	dass	je mehr man arbeitet , desto mehr Geld bekommt man .
FalkoEssayL2WHIGv2.0	denken	dass	je mehr man arbeiten , desto mehr Geld bekommen man .
FalkoEssayL2WHIGv2.0	VVINF	\$, KOUS	ADV ADV PIS VVFIN \$, KON PIAT NN VVFIN PIS \$.
FalkoEssayL2WHIGv2.0	ZH1	(grid)	

Export Options: Export, Frequency Analysis

Corpus List: FalkoEssayL2WHIGv2.0 (195, 130,187), FalkoGeorgetownL2v1.0 (92, 78,151), FalkoSummaryL1v1.2 (57, 21,211), FalkoSummaryL2v1.2 (106, 40,638), FalkoWHIGL2v2.1 (196, 130,949), FNHD_context (5, 2,674).

Search Options: Visible: All

Vom Resultat zum Export

The screenshot shows a search interface with a query builder on the left and export settings on the right. The query builder contains two lines of text: `tok=/[jJ]e/ . pos=/A.*/ .2,10` and `tok=/[dD]esto|[jJ]e|[Uu]mso/`. Below the query builder are buttons for 'Search', 'More', and 'History'. A summary box indicates '39 matches in 36 documents'. At the bottom, there are tabs for 'Corpus List' and 'Search Options', a 'Visible:' dropdown set to 'All', and a table with columns 'Name', 'Texts', and 'Tokens'. The table lists 'DDD-Tatian' with 245 texts and 54,677 tokens. On the right, the 'Export' panel shows settings for 'GridExporter', 'Left Context' (20), 'Right Context' (20), and 'Annotation Keys' (tok, pos). It also includes a 'Parameters' field and buttons for 'Perform Export', 'Cancel Export', and 'Download'.

tok=/[jJ]e/ . pos=/A.*/ .2,10
tok=/[dD]esto|[jJ]e|[Uu]mso/

Query Builder

Search More History

39 matches in 36 documents

Corpus List Search Options

Visible: All

Name	Texts	Tokens
DDD-Tatian	245	54,677

Help/Examples Query Result x Export x

Exporter GridExporter

Left Context 20

Right Context 20

Annotation Keys tok, pos

Parameters

The Grid Export each annotation To display only : Parameters: *metakeys* - com *numbers* - set to

Perform Export Cancel Export Download

Export

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

The screenshot shows the ANNIS interface with the following elements:

- Header:** "Help us to make ANNIS better!" and "logged in as 'kidko_user' Logout".
- Search Bar:** Contains the query `n="sein" .* POS=/AD.*/`.
- Search Results:** "3078 matches in 154 documents".
- Export Panel:** Shows the "GridExporter" selected. It includes fields for "Left Context" (10) and "Right Context" (10). A red arrow points to the "GridExporter" dropdown.
- Help Text:** "The Grid Exporter can export all annotations of a search result and its context. Each annotation layer is represented in a separate line, and the tokens covered by each annotation are given as number ranges after each annotation in brackets. To suppress token numbers, input numbers=false into the parameters box below. To display only a subset of annotations in any order use the 'Annotation keys' text field, input e.g. 'tok,pos,cat' to show tokens and the annotations pos and cat."
- Parameters:** `metakeys` - comma separated list of all meta data to include in the result (e.g. `metakeys=title,documentname`); `numbers` - set to "false" if the grid event numbers should not be included in the output (e.g. `numbers=false`).
- Buttons:** "Perform Export", "Cancel Export", and "Download".
- Status:** "exported 350 items in 35,08 s".
- Table:** A table with columns "Name", "Texts", and "Tokens".

Name	Texts	Tokens		
mo	82	146,704	?	📄
mu	191	346,039	?	📄

Bedingte Formatierung in Excel

Neue Formatierungsregel

Formatvorlage: Klassisch

Formel für die Ermittlung der zu formatierenden Zellen verw...
=Istleer(A1)

Formatieren mit: Hellrote Füllung mit dunkelroter Schrift AaBbCcYyZz

Abbrechen OK

	A	B	C	D	M	N	O
0	tok	von Kriminalität ist nur kurzfr					
	pos	APPR[1-1] NN[2-2] VAFIN[3-3]					
1	tok	ja keine Veränderungen , und					
	pos	ADV[1-1] PIAT[2-2] NN[3-3] \$,					
2	tok	eigenen Büchern und Lernque					
	pos	ADJA[1-1] NN[2-2] KON[3-3] N					
3	tok	die Gesellschaft entsprechen . Aber was heißt eigentlich " entsprechen " ? Man könnte zuerst denken , dass je mehr man arbeitet , desto mehr Geld bekommt man . !					
	pos	ART[1-1] NN[2-2] VVIN[3-3] \$.[4-4] KON[5-5] PWS[6-6] VVFIN[7-7] ADV[8-8] \$([9-9] VVIN[10-10] \$([11-11] \$.[12-12] PIS[13-13] VMFIN[14-14] ADV[15-15] VVIN[16-16]					
4	tok	Beitrag für die Gesellschaft , wenn man daran denkt : " je nützlicher man ist , desto besser muss man bezahlt werden " . Das kommt man aber zu einer tieferen					
	pos	ADV[1-1] NN[2-2] APPR[3-3] ART[4-4] NN[5-5] \$.[6-6] KOUS[7-7] PIS[8-8] PROAV[9-9] VVFIN[10-10] \$.[11-11] \$([12-20] ADV[21-21] ADJD[22-22] PIS[23-23] VAFIN[24-24]					
5	tok	meisten Fällen) noch weniger verdienen als die Frauen . So , hier sehe ich eigentlich das Problem : je stärker die Frauen werden , je schwächer werden die Männer . De					
	pos	PIAT[1-1] NN[2-2] \$([3-3] ADV[4-4] ADV[5-5] VVFIN[6-6] KOKOM[7-7] ART[8-8] NN[9-9] \$.[10-10] ADV[11-11] \$.[12-12] ADV[13-13] VVFIN[14-14] PPER[15-15] ADV[16-16]					
6	tok	am meisten aus Geld zusammensetzt . Anders gesagt ist das Prinzip dieser Welt nichts anderes als die finanzielle Entlohnung . Je grösser ist das Lohn , desto wirklicher i					
	pos	APPRART[1-1] PIS[2-2] APPR[3-3] NN[4-4] VVFIN[5-5] \$.[6-6] ADV[7-7] VVPP[8-8] VAFIN[9-9] ART[10-10] NN[11-11] PDAT[12-12] NN[13-13] PIS[14-14] PIS[15-15] KOKOM					

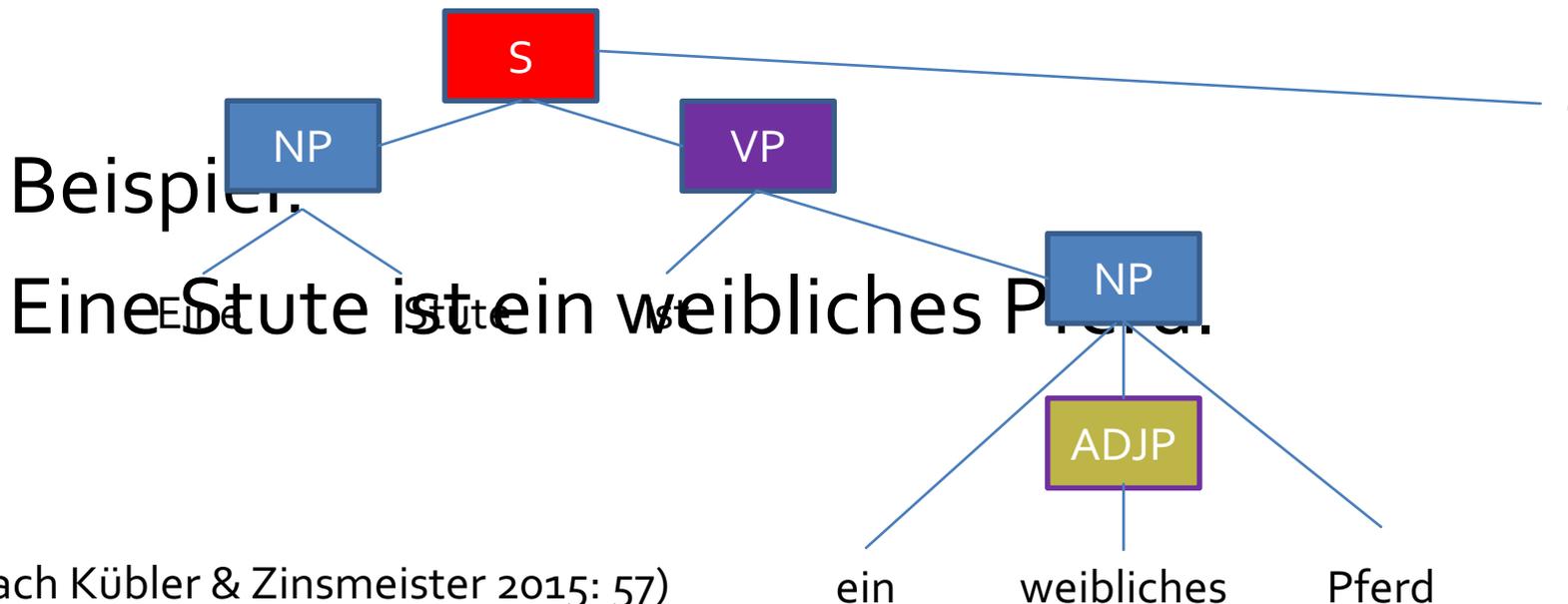
Baumbanken in ANNIS

Baumbanken

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

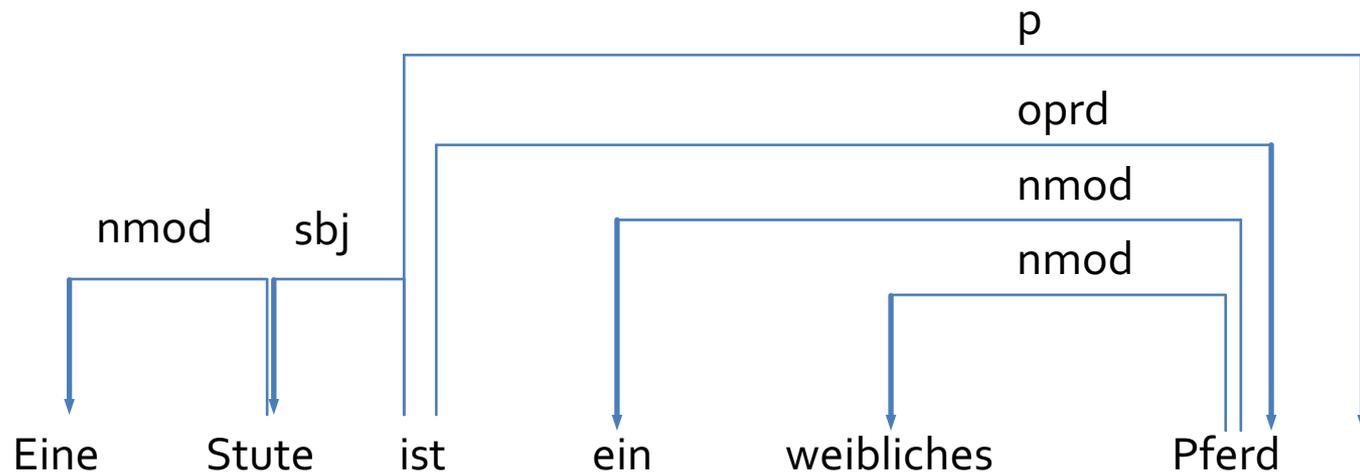
Konstituentenstruktur und Dependenzstruktur

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



Konstituentenstruktur und Abhängigkeitsstruktur

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



Konstituentenstruktur und Abhängigkeitsstruktur

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

Hybride Modelle

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

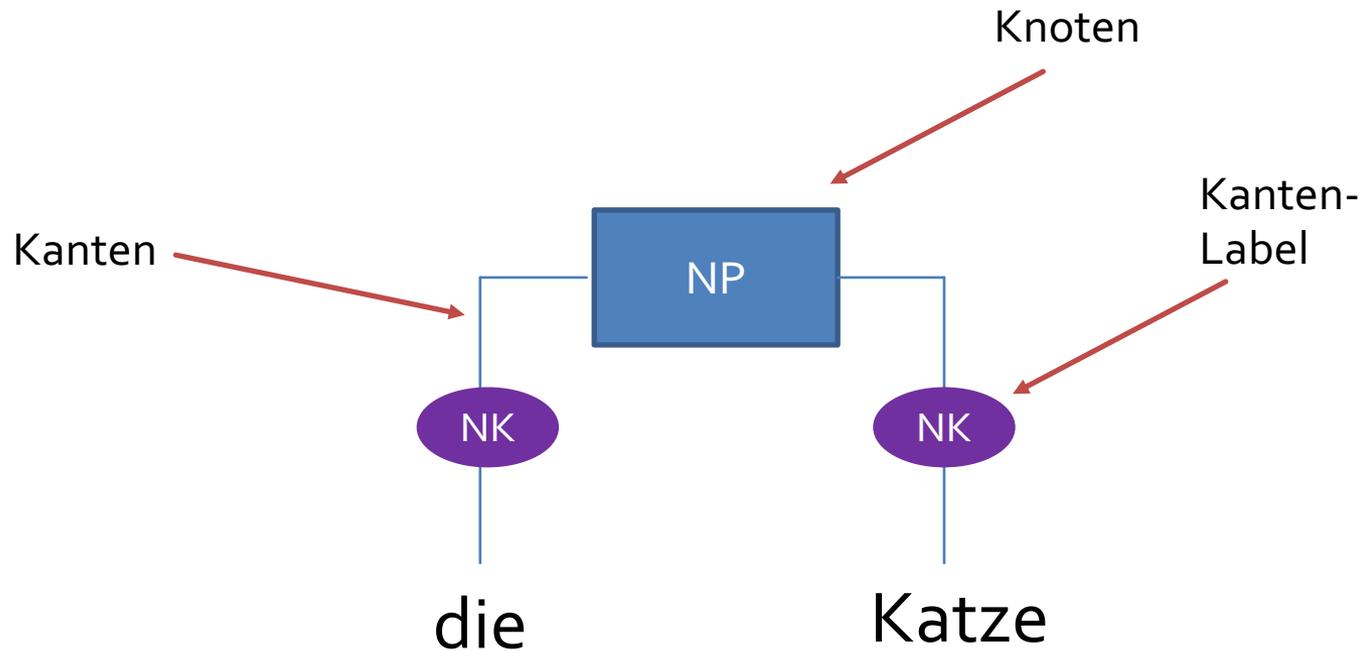
Terminologie

- 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 Analyse häufig auch mit **Chunks**
- Chunks sind prosodische Sprechereinheiten (in gesprochener Sprache an kleinen Sprechpausen erkennbar) / zusammengehörige Mehrworteinheiten

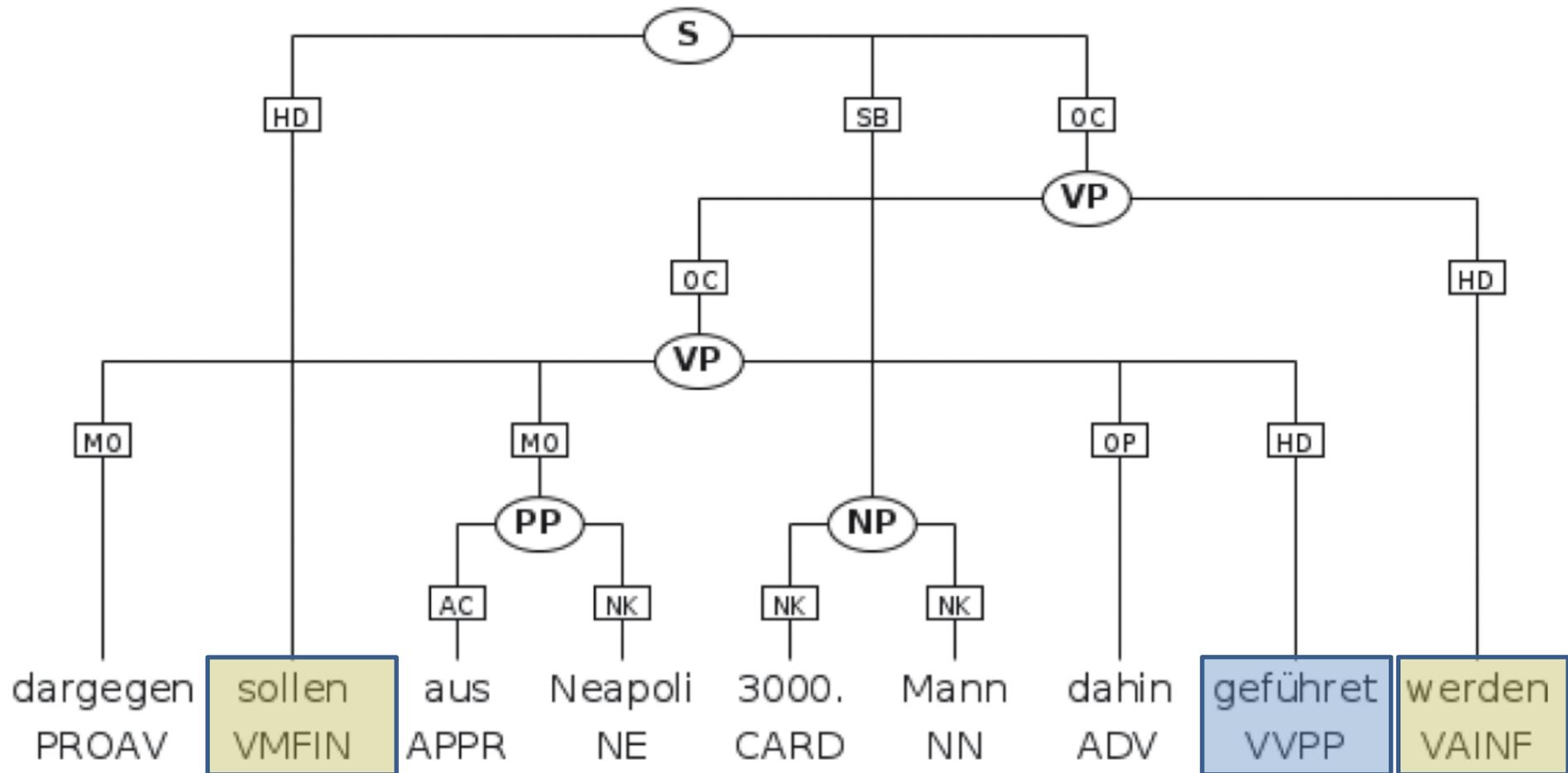
Historische Baumbanken des Deutschen

- 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.

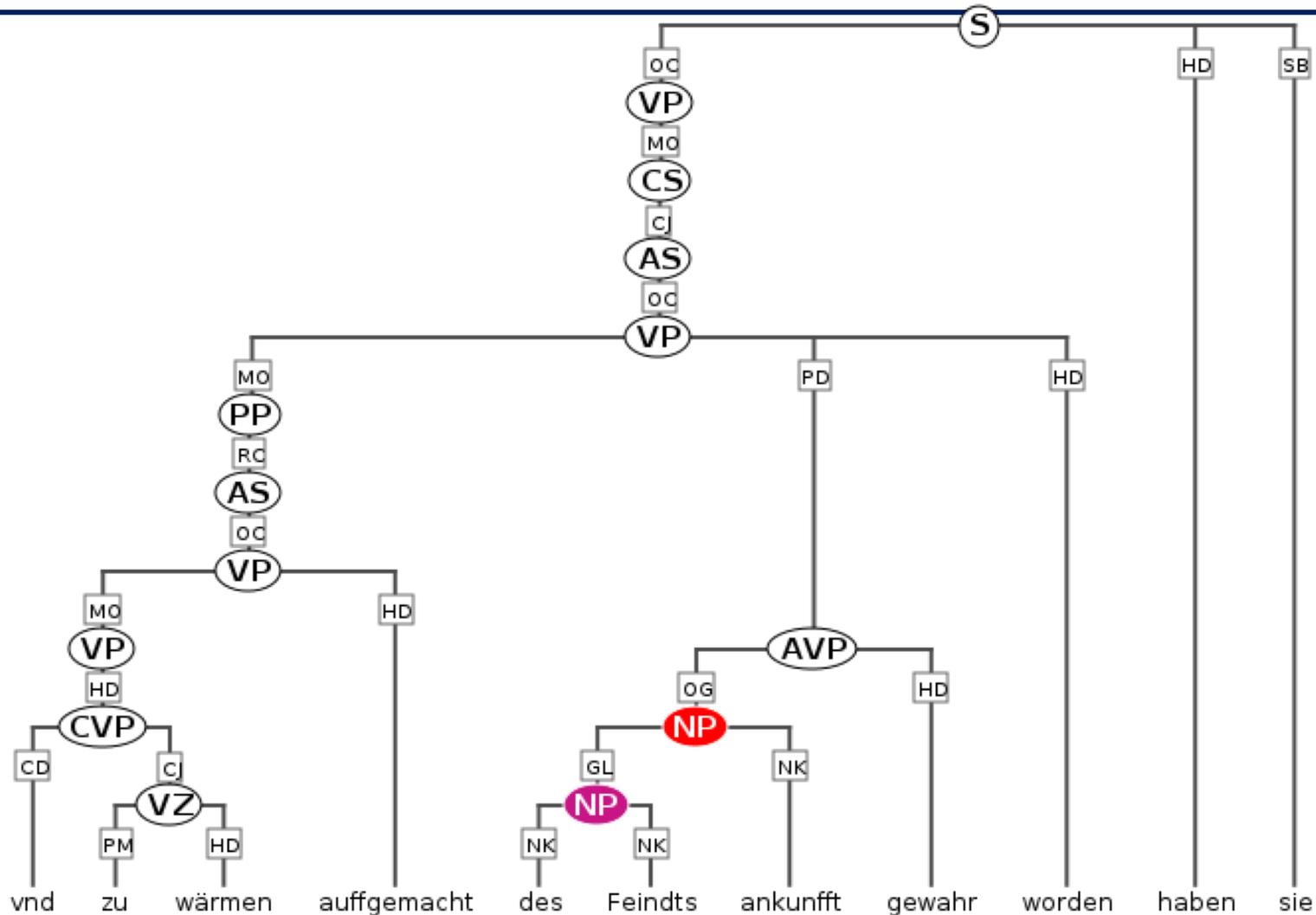
Strukturbaum: Beispiel



Kreuzende Kanten



Bsp.: Mercurius-Baumbank



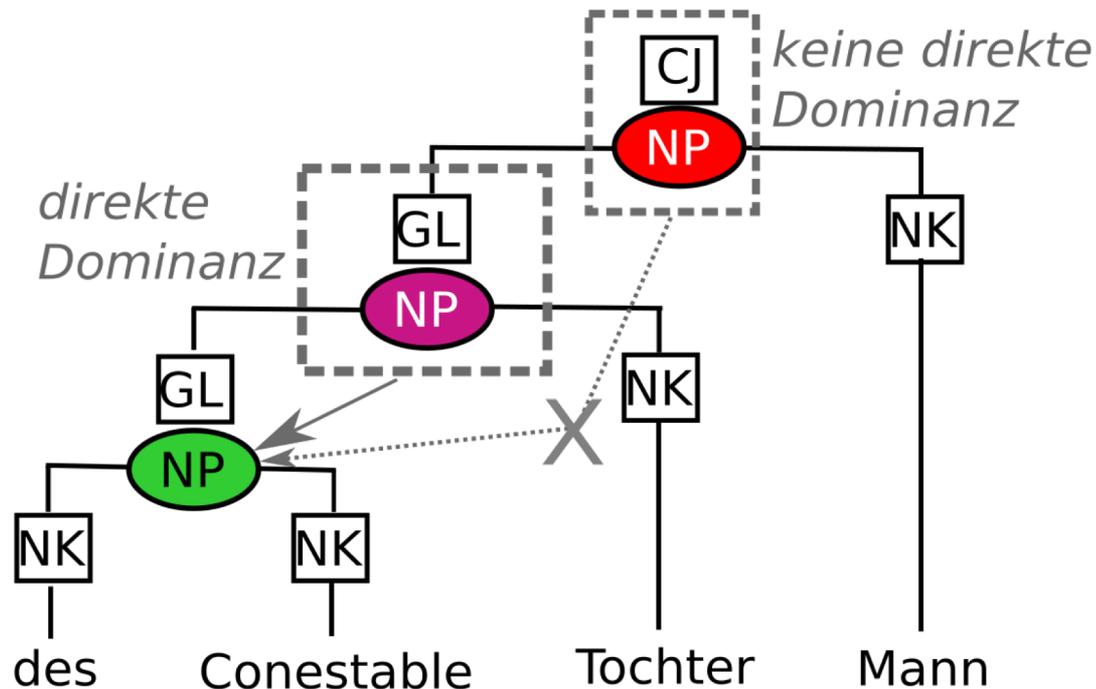
Bsp: Mercurius-Baumbank

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

beliebiger direkte
Knoten 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"`