

# 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

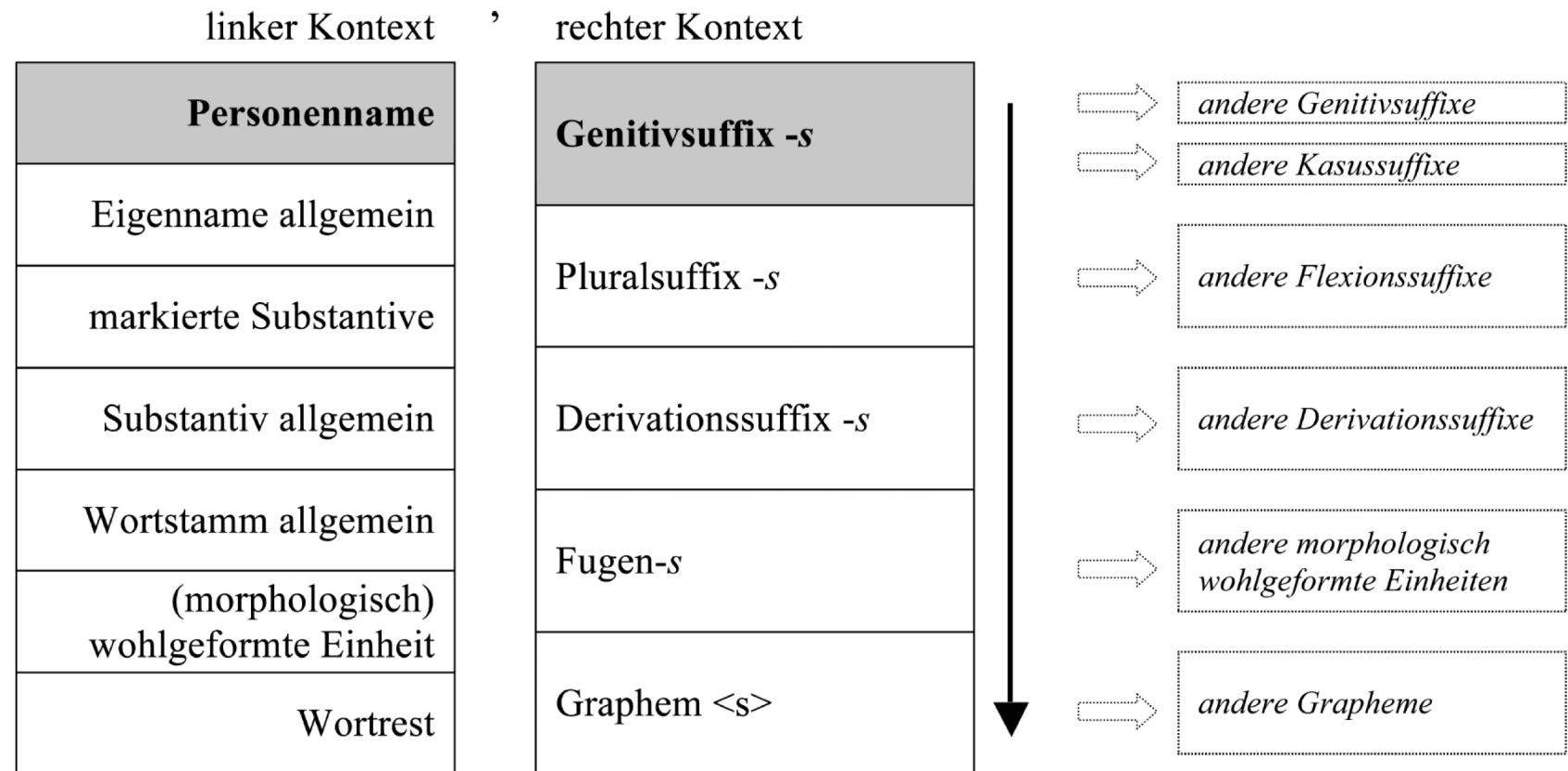
---

- 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  
**Wortbildungsbeschränkungen**, 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

Merkmale	Einheiten	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: *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 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

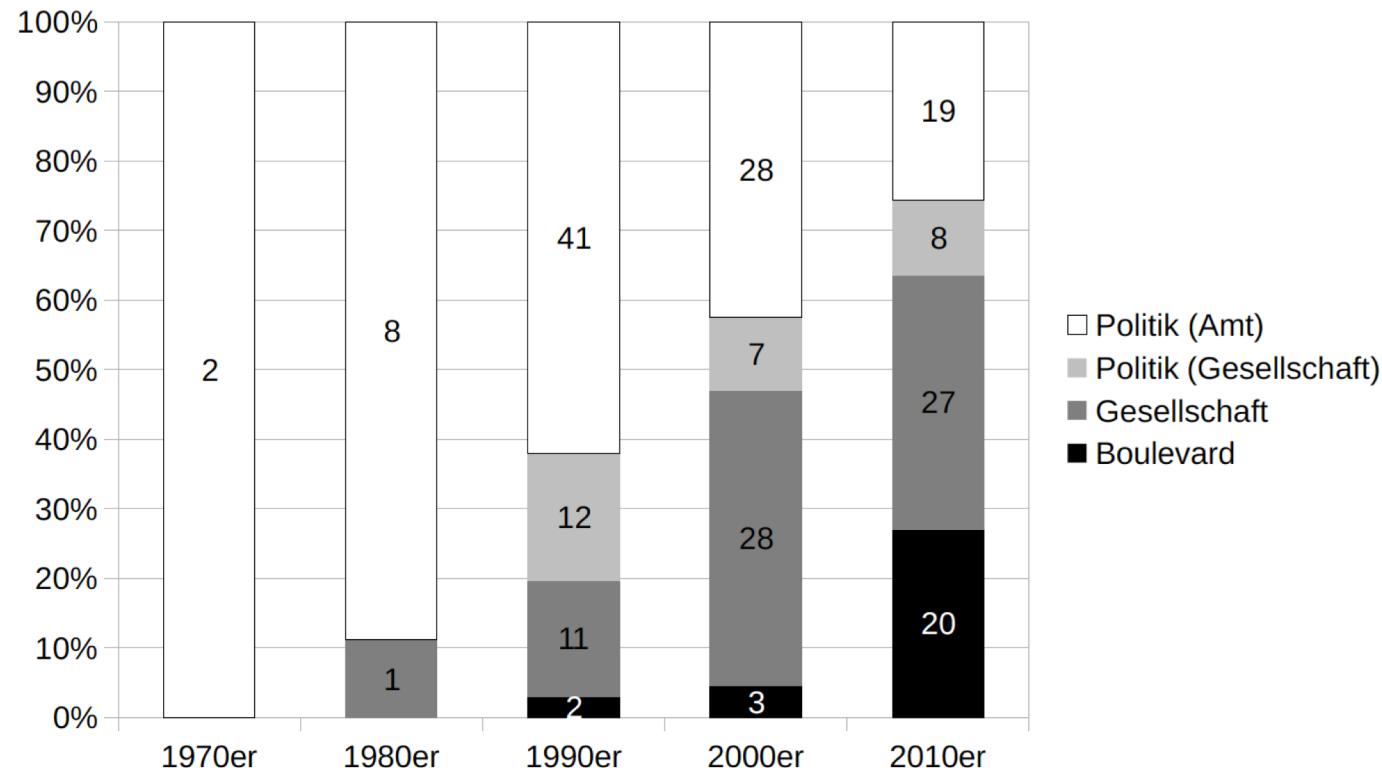


Abb. 5a. Deutsche Erstbelege (Entlehnungen und dt. Bildungen) in DeReKo/ZEIT nach Skandalfeld  
(n = 217).

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](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-kind-a-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 / **Murmeling** noch Ungedult  
würde bey ihnen gespurt (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

# Potenzielle Produktivität

---

- Einfache Formel:

Anzahl der Hapaxe, die zum  
Wortbildungsmuster gehören

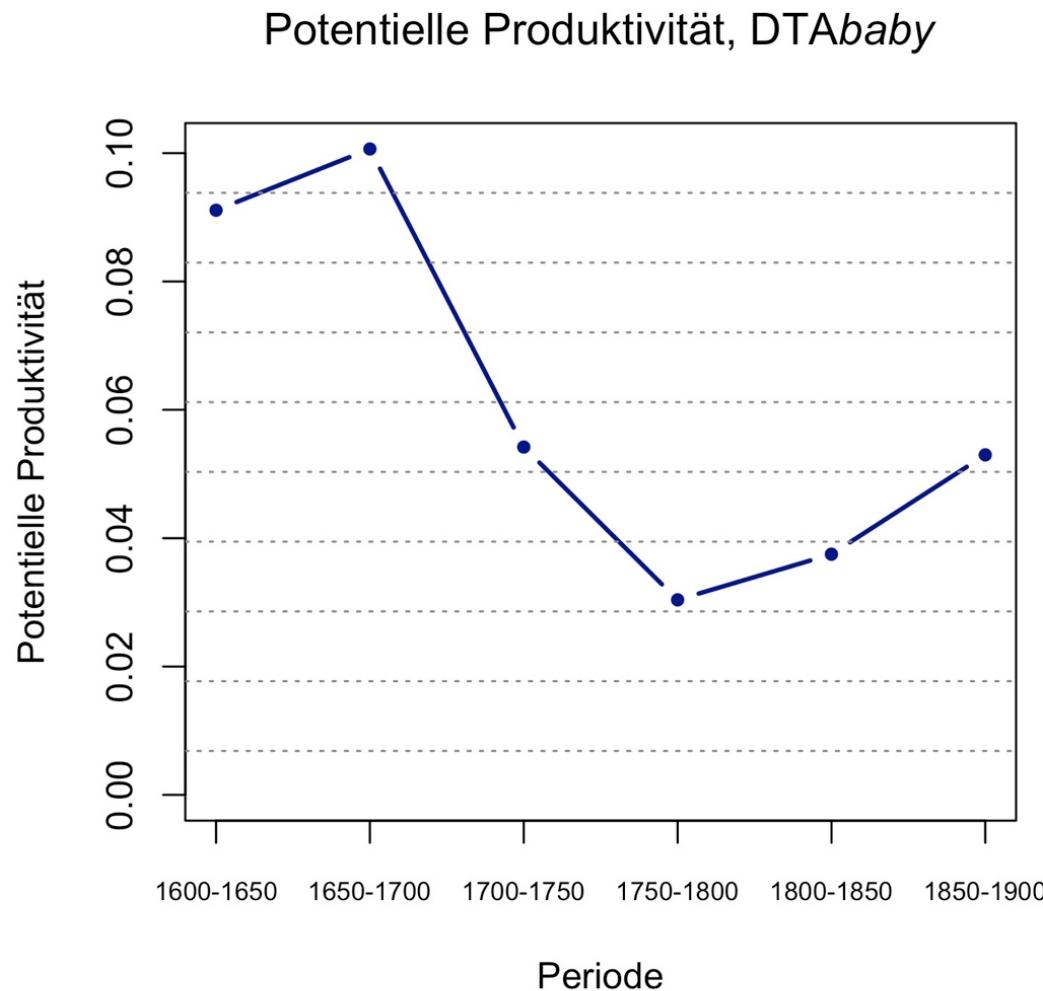
(z.B.: alle Hapaxe auf -heit/-keit)

---

Gesamtzahl der Belege, die zum  
Wortbildungsmuster gehören

(z.B.: alle Belege auf -heit/-keit)

# Beispiel *ung*-Nominalisierung



# Beispiel Fugenelemente

---

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

# Beispiel Fugenelemente

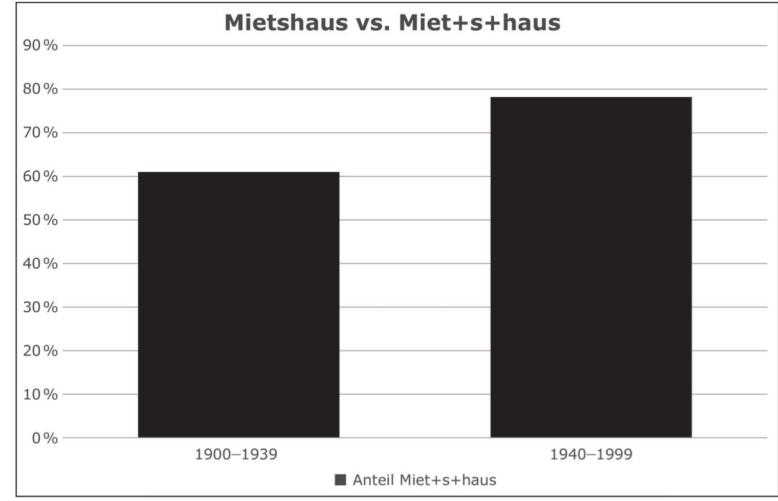
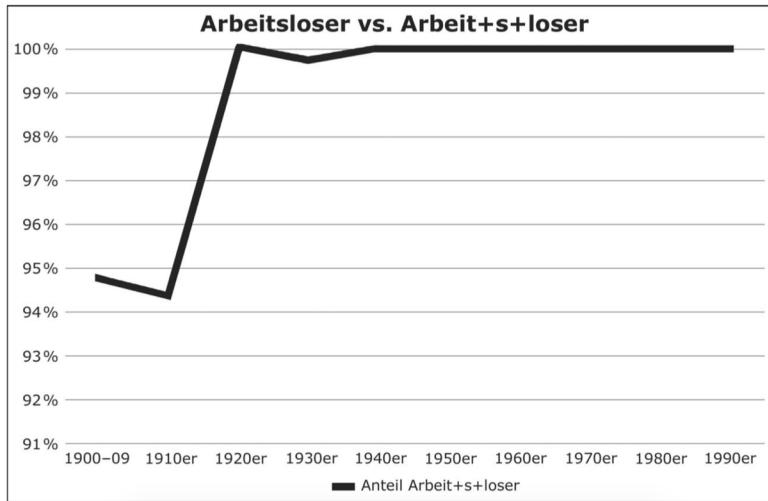


Tabelle 6: *Die Häufigkeit des Fugen-s nach präfigierten Erstgliedern.*

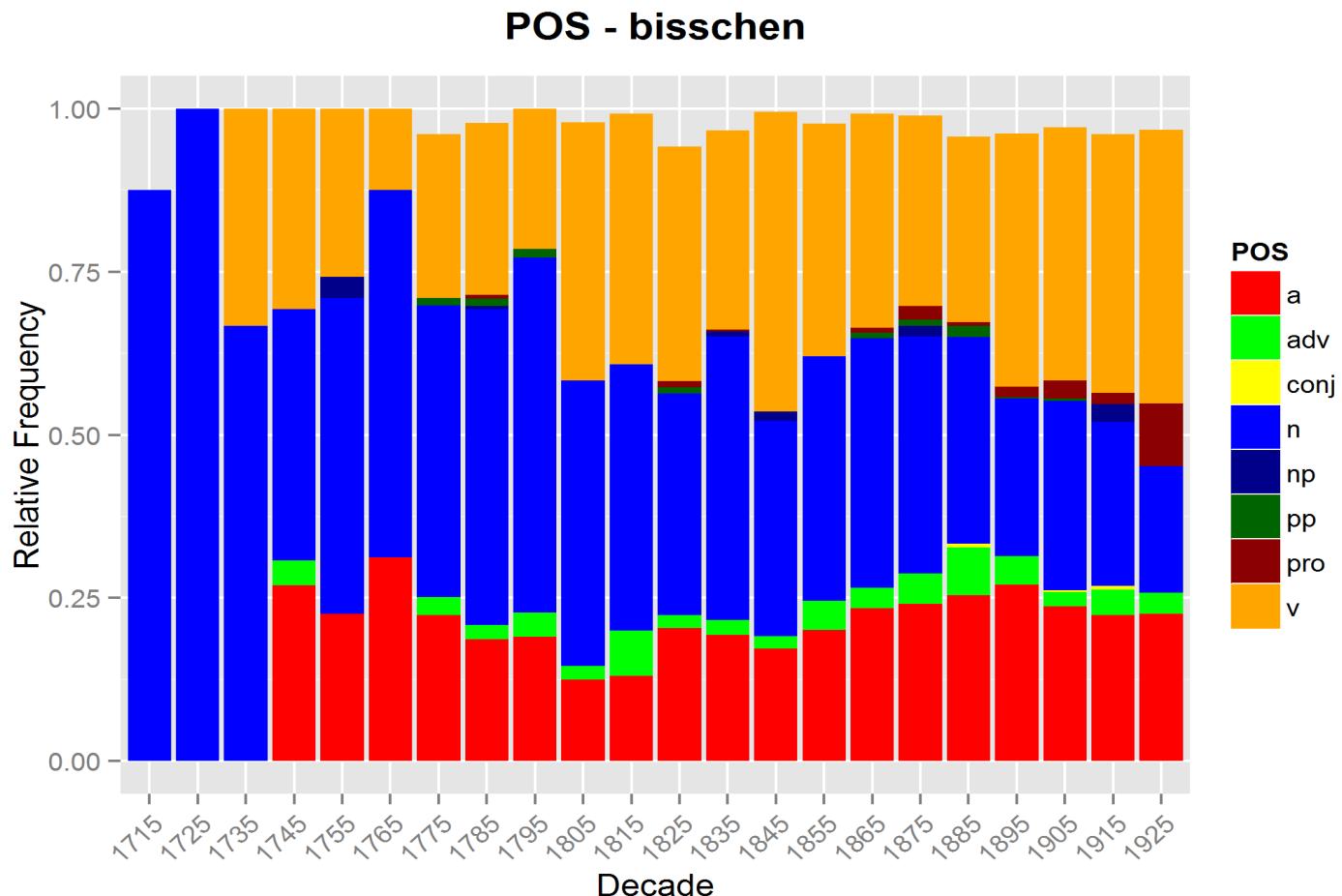
	Erstglied enthält:	
	unbetontes Präfix (Typ <i>Beruf-</i> )	betontes Präfix (Typ <i>Anruf-</i> )
Tokens:	85 % (von 495.887 Komposita)	36 % (von 324.503 Komposita)
Types:	82 % (von 17.999 Komposita)	37 % (von 11.325 Komposita)

# Beispiel Gradmodifikatoren (Neels & Hartmann 2017)

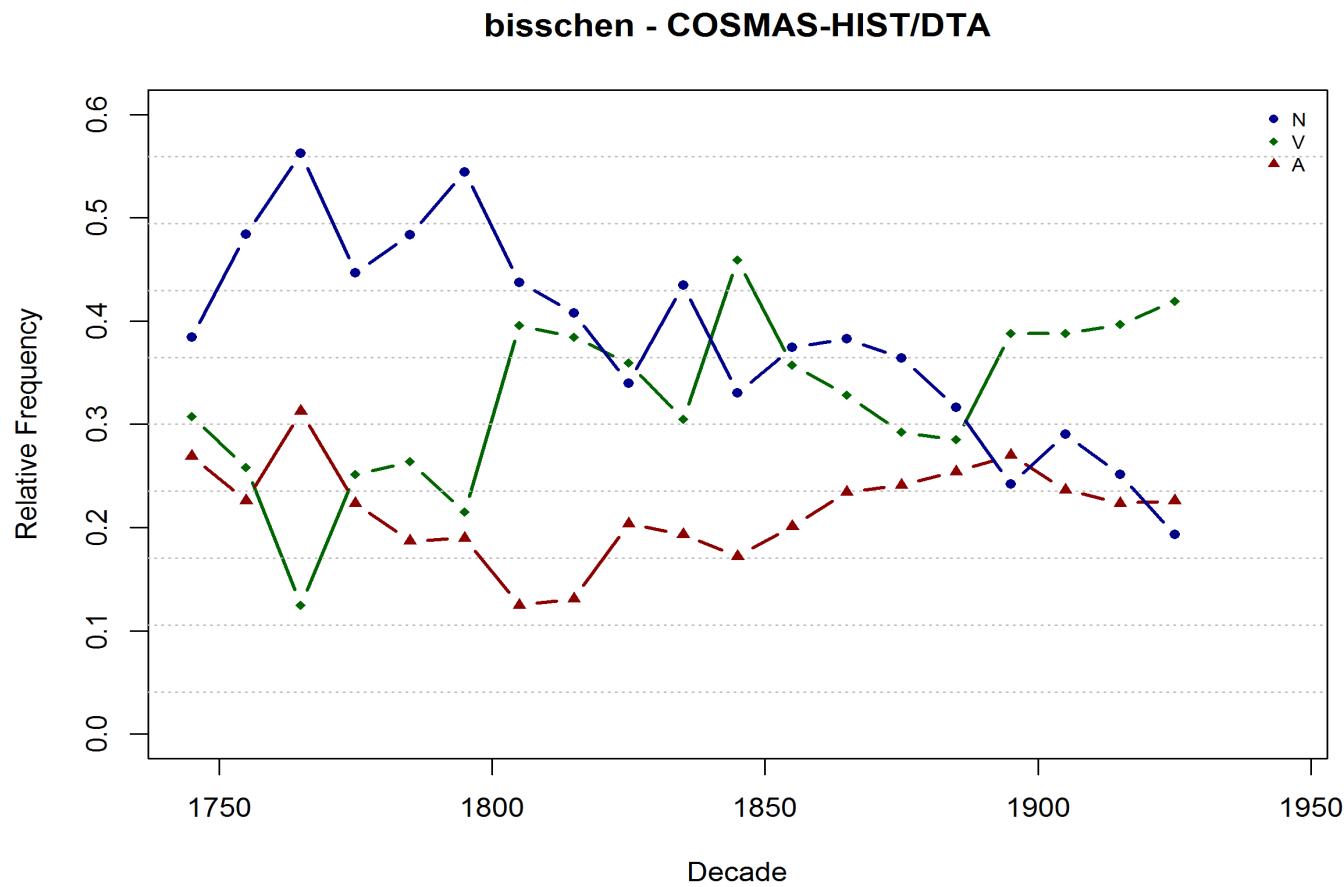
---

- 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

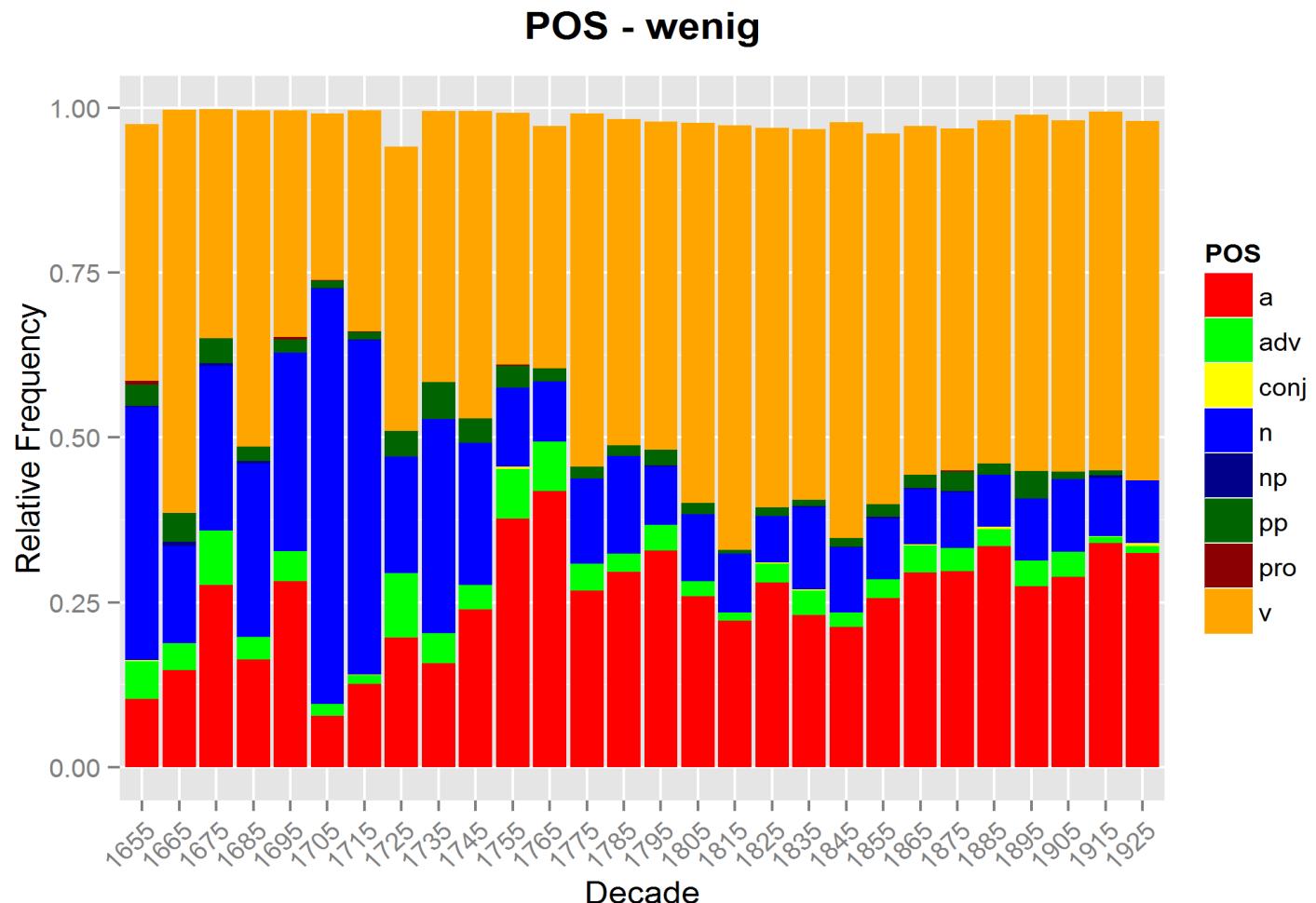
# Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



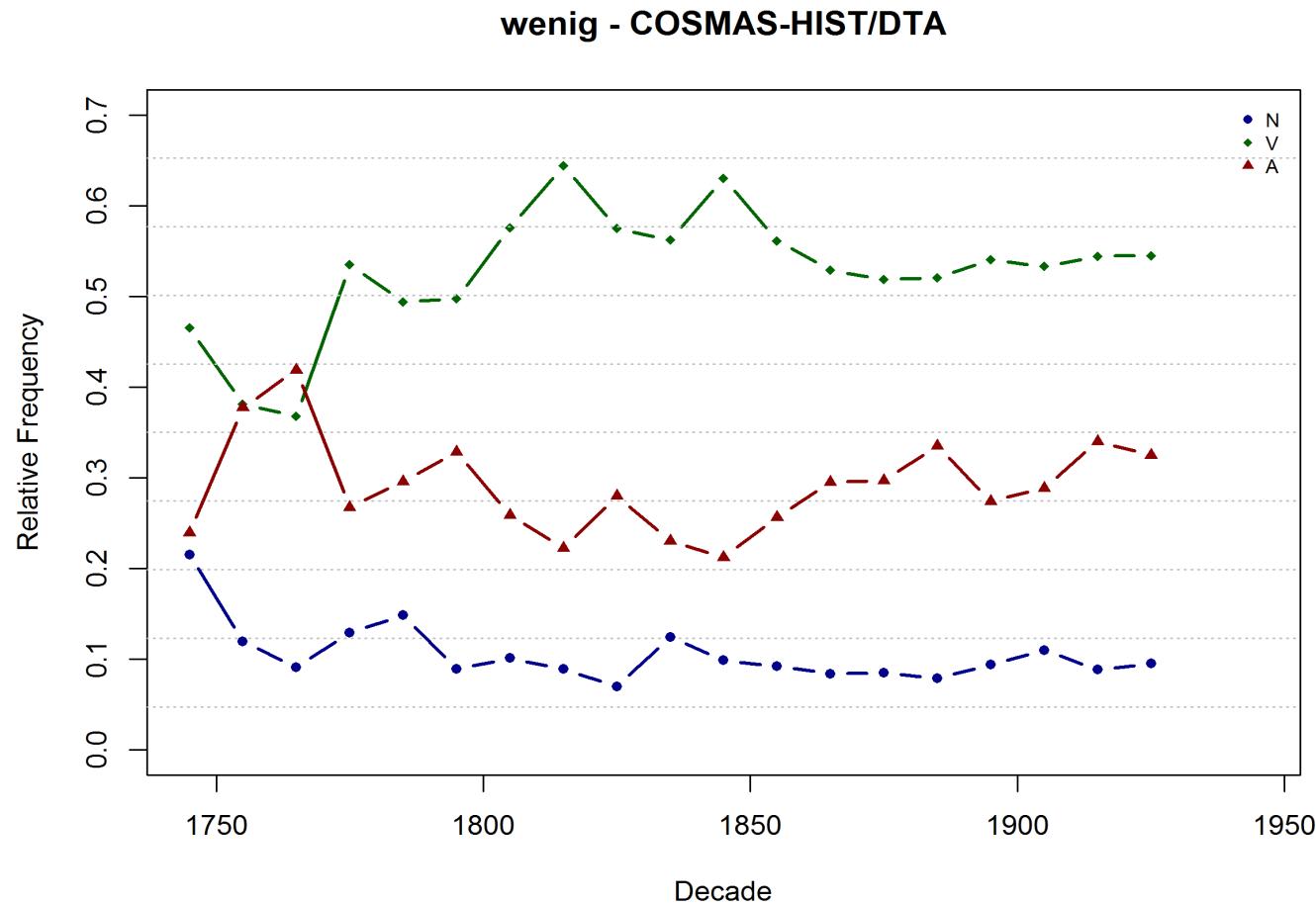
# Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



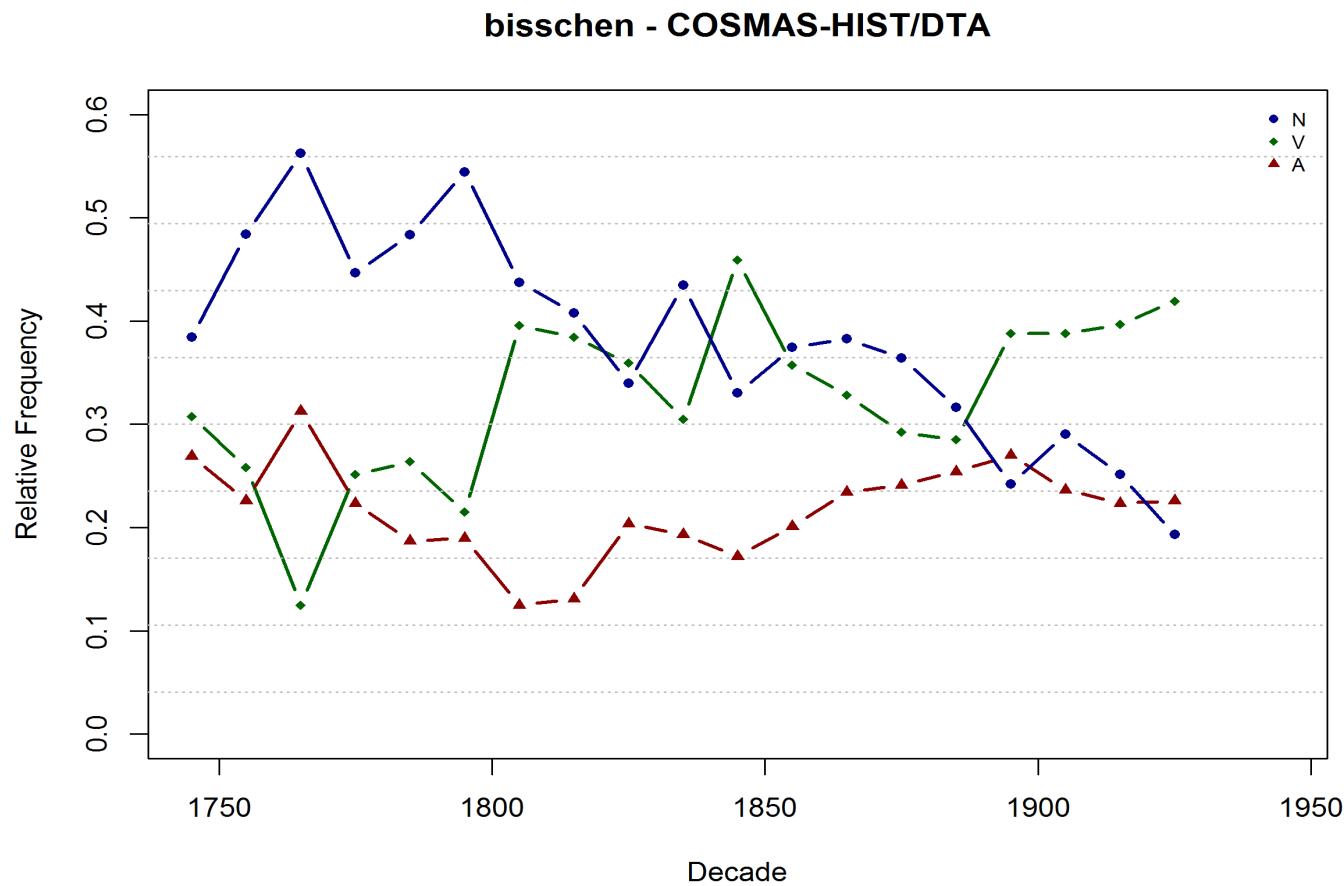
# Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



# 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

The screenshot shows the DTAQ (Deutsches Textarchiv) search interface. At the top, there is a dark brown header bar with the text "Anmelden (DTAQ)" on the left and social media icons for Twitter and Google+ on the right. Below the header is a logo consisting of three squares containing the letters D, T, and A. To the right of the logo is a search input field with a "suchen" button. Below the search field are three radio buttons labeled "in den Titeldaten", "im Korpus", and "in der Dokumentation", with the third one being selected. To the right of these buttons are links for "Hilfe" and "Beispielanfragen". The example query shown is: \$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.deutsches-textarchiv.de](http://www.deutsches-textarchiv.de)
- [www.dwds.de](http://www.dwds.de)

# Deutsches Textarchiv

- DTA ist getaggt und lemmatisiert

```
<tokens>
<token ID="w1">D.</token>
<token ID="w2">Henrici</token>
<token ID="w3">Cai#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="w10"></token>
<token ID="w11">welcher</token>
<token ID="w12"></token>
<token ID="w13"></token>
<token ID="w14"></token>
<token ID="w15"></token>
<token ID="w16"></token>
<token ID="w17"></token>
<token ID="w18"></token>
<token ID="w19"></token>
<token ID="w20"></token>
<token ID="w21"></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="wieyu">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">zur</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 gegenwärtssprachlicher 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

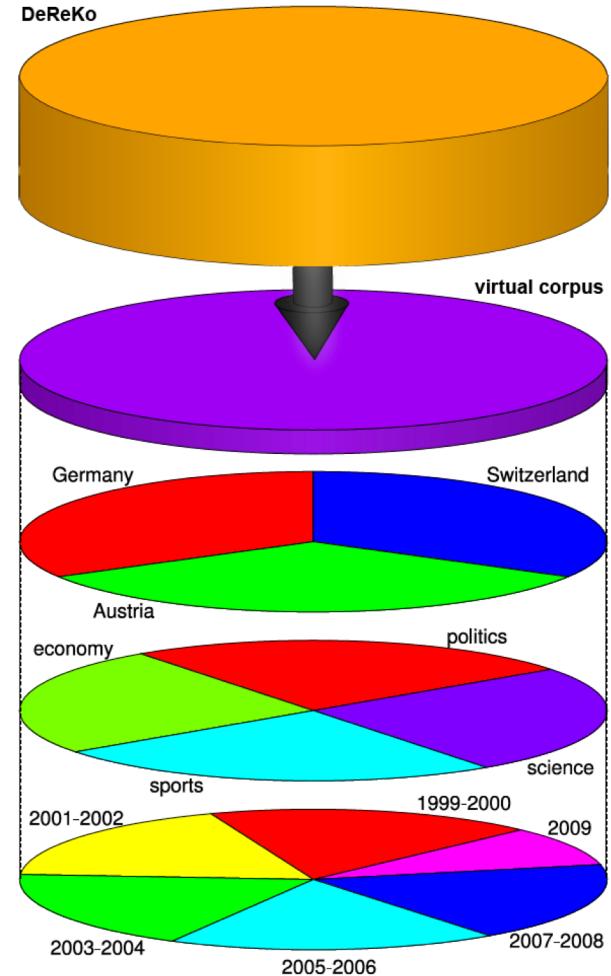


Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

(aus Kupietz et al. 2010)

# Virtuelle Korpora erstellen

W - Archiv der geschriebenen Sprache

Aktuelles Korpus: W-öffentliche - alle öffentlichen Korpora des Archivs W (mit Neuakquisitionen) [1]

Korpusdefinition

Gesamtkorpus

- 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

CorpDefID:

Suchmuster

2015  Hilfe

linksbündig  rechtsbündig

Groß-/ Kleinschreibung beachten

<<  >>

Suchen

Mark. löschen

Korpus:

- 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

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 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

Help us to make ANNIS better!

not logged in | Log in

cat="NP"

Query Builder

Query result

Search More ▾ History ▾

41 matches in 2 documents

Corpus List Search Options

Visible: All

Filter

Name	Texts	Tokens
pcc2	2	399
RIDGES_Herbology_1	22	122.698
RIDGES_Herbology_1	29	154.266
Ridges_Herbology_Ve	13	60.811

Help/Examples Query Result

Base text Token Annotations

1 / 5 Displaying Results 1 - 10 of 41 Result for: cat="NP"

Path: pcc2 > 11299 (tokens 1 - 10)	left context: 5	right context: 5
Feigenblatt Die Jugendlichen in Zossen wollen ein		
Feigenblatt der jugendliche in Zoszen wollen ein		
Nom.Sg.Neut Nom.PI.* Nom.PI.* – Dat.Sg.Neut 3.PI.Pres.Ind Acc.Sg.Neut		
NN ART NN APPR NE VMFIN ART		

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

Visualizations

```
graph TD; S((S)) --- SB((SB)); S --- HD((HD)); S --- OA((OA)); SB --- NK1[NK]; SB --- NK2[NK]; HD --- NK3[NK]; OA --- NK4[NK]; OA --- NK5[NK]; NK1 --- Die[Die]; NK2 --- jugendlichen[Jugendlichen]; NK3 --- in[in]; NK4 --- Zoszen[Zoszen]; NK5 --- wollen[wollen]; NK4 --- ein[ein];
```

# ANNIS

---

zu den über ANNIS verfügbaren Korpora gehören z.B.

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

# ANNIS Query Language (AQL)

---

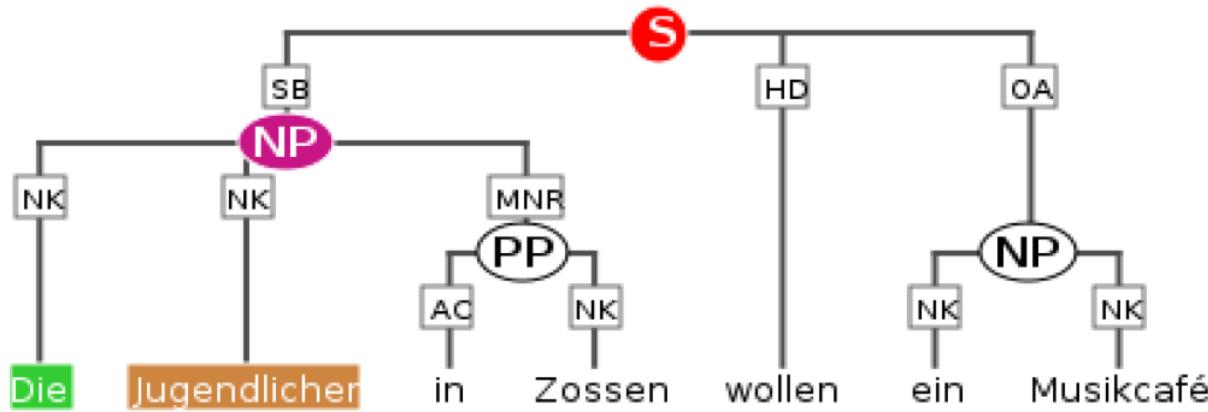
- Ähnlich logisch wie CQL...
- ...aber (m.E.) etwas komplizierter.
- Vorteil: extrem detaillierte syntaktische Suchanfragen möglich.
- vgl. "Cheat Sheet für ANNIS"



# 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

# ANNIS Query Language

The screenshot shows the ANNIS3 web interface. The top section, labeled "Eingabefeld" (Input Field), contains a text area for entering AQL queries, a "Query Builder" button, and a "Help/Examples" dropdown menu with "Tutorial" selected. The bottom section, labeled "Korpora" (Corpora), displays a "Corpus List" table with two entries: "mo" and "mu".

**Eingabefeld**

Please enter AQL query

Query Builder

Help/Examples

Tutorial

Example Queries

Example Query Description

Q " " search for the word " "

Q " " search for the word " "

open corpus browser

mo

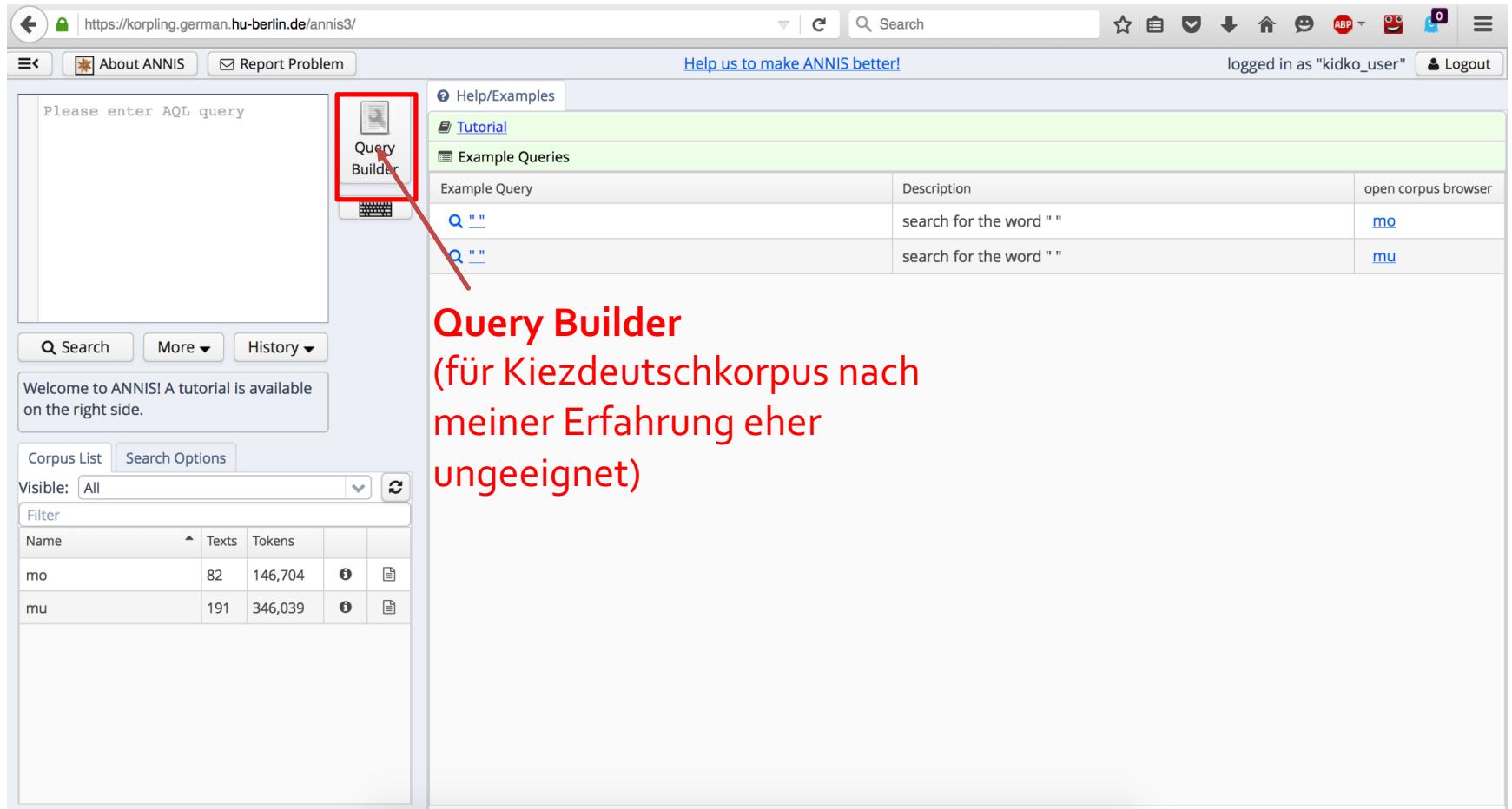
mu

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

**Korpora**

Name	Texts	Tokens	Actions
mo	82	146,704	Info Edit
mu	191	346,039	Info Edit

# ANNIS Query Language



The screenshot shows the ANNIS web interface at <https://korpling.german.hu-berlin.de/annis3/>. The left sidebar contains a text input field for "Please enter AQL query", search, more, and history buttons, and a "Corpus List" section with a table showing two entries: "mo" and "mu". The main area has a "Help us to make ANNIS better!" banner and a "logged in as 'kidko\_user'" message. On the right, there's a "Help/Examples" sidebar with "Tutorial" and "Example Queries" sections. The "Query Builder" button in the sidebar is highlighted with a red box and a red arrow pointing to it. The "Example Queries" table lists three entries:

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

**Query Builder**  
(für Kiezdeutschkorpus nach  
meiner Erfahrung eher  
ungeeignet)

# REM: Ein Korpus wählen

Corpus List    Search Options

Visible:  ▼ ↻

Filter	Name	Count	Size	Info	File
All	MiGraKo	5	956	<span>i</span>	<span>file</span>
	eReM	7	44,986	<span>i</span>	<span>file</span>
	11-12_1- <del>obd</del> -PV-X	4	36,281	<span>i</span>	<span>file</span>
	11_2-12_1-obd-PV-G	5	33,992	<span>i</span>	<span>file</span>
	12-13_1-mdnd-PV-X	4	30,309	<span>i</span>	<span>file</span>
	12_2-alem-PV-G	20,457			

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

# Einzelne Tokens suchen

---

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

"so"

"mann"

"vrouw" – **keine Treffer**

# Lemmasuche

---

- Lemmas in REM sind mhd. Lemmas (wie im Mhd. Wörterbuch)
- im Zweifelsfall unter [mhdwb.de](http://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 hellevarte . ern vverde ir geverte . des

annotations

	6a,8	6a,9				6a,10				
reference	.	der	vbel	hellevar	.	ern	vverde	ir	geverte	.
tok_dipl	.	der	vbel	hellevar	.	er	n	vverde	ir	g
tok_anno	.	der	vbel	hellevar	.	er	ne	vverde	ir	g
norm		der	übel	hellewarte		MS1	MS2			
tokenization										
pos	\$_	DDART	ADJA	NA	\$_	PPER	PTKNEG	VAFIN	DPOSA	N
posLemma	\$_	DD	ADJ	NA	\$_	PPER	PTK	VA	DPOS	N
lemma	dér	übel	helle-warte			ér	ne	wérden	ir(e)	g
lemmald	<a href="#">29817000</a>	<a href="#">174642000</a>	<a href="#">71433000</a>			<a href="#">40380000</a>	<a href="#">119841000</a>	<a href="#">225138000</a>	<a href="#">83982000</a>	<a href="#">5</a>
lemmaLemma	dér	übel	helle-warte			ér	ne	wérden	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 with the following details:

- Query:** `n="schwul" . POS=/AD.*"`
- Help/Examples:** Help us to make ANNIS better!
- Base text:** Displaying Results 1 - 2 of 2
- Result for: n="schwul". POS=/AD.\*/**
- Left context:** 5 | Right context: 5
- Result 1 (Path: mo > Mo05WD\_11-2 (tokens 1920 - 1931)):**
  - SPK15::v WACHS nur
  - SPK39::v a sein kle:ner WIlli da unten is kle:n
  - SPK67::v ich bin SCHWUL
- Result 2 (Path: mu > MuH9WT\_10 (tokens 165 - 176)):**
  - MuH9WT::v wallAH er soll
  - SPK101::v doch die ganze zeit voll (SCHWUL) (halt)
- Corpus List:** Visible: All
- Filter:** Name, Texts, Tokens
- Documents:** mo (82, 146,704), mu (191, 346,039)

# Wortfolgen suchen

.\* = geht indirekt voran

The screenshot shows the ANNIS web interface with the following details:

- Query:** n="schwul" .\* POS=/AD.\*
- Results:** 204 matches in 20 documents.
- Annotations:** The interface displays three examples of matches with annotations:

  - Result 1:** Path: mo > Mo05WD\_11-2 (tokens 1920 - 1931). Left context: 5, right context: 5. Annotations: WACHS, nur, a, sein, kle:ner, WILli, da, unten, is, kle:n.
  - Result 2:** Path: mo > Mo05WD\_11-2 (tokens 1920 - 1933). Left context: 5, right context: 5. Annotations: WACHS, nur, a, sein, kle:ner, WILli, da, unten, is, kle:n, (-).
  - Result 3:** Path: mo > Mo05WD\_11-2 (tokens 1920 - 1934). Left context: 5, right context: 5. Annotations: WACHS, nur, a, sein, kle:ner, WILli, da, unten, is, kle:n, (-), ich.

# Wortfolgen suchen

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

The screenshot shows a corpus search interface with the following details:

- Query:** `n="schwul" .2,4 POS=/AD.*/`
- Results:** 11 matches in 6 documents.
- Annotations:** The results are presented in five rows, each corresponding to a different document path and token range. Each row shows the context around the word "schwul".
  - Row 1:** Path: mo > Mo05WD\_11-2 (tokens 1920 - 1933). Context: WACHS nur a sein kle:ner WILli da unten is kle:n (-)
  - Row 2:** Path: mo > Mo05WD\_11-2 (tokens 1920 - 1934). Context: WACHS nur a sein kle:ner WILli da unten is kle:n (-) ich
  - Row 3:** Path: mo > Mo12MD\_07 (tokens 1531 - 1544). Context: gib mal WASser bitte fiasch (-)
  - Row 4:** Path: mu > MuH12MD\_02 (tokens 1948 - 1961). Context: sind doch kaum GAR nich (-)
  - Row 5:** Path: mu > MuH12MD\_02 (tokens 1948 - 1962). Context: sind doch kaum GAR nich (-)
- Corpus List:** Shows two entries: "mo" and "mu".
- Search Options:** Includes "Visible: All" and a "Filter" section.

# 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 shows the ANNIS interface with the following elements:

- Search Bar:** Contains the query `tok=/[jJ]e/ . pos=/A.*/ .2,10 tok=/[dD]esto|[jJ]e|[Uu]mso/`.
- Results Summary:** Shows 39 matches in 36 documents.
- Buttons:** Search, More, History, Export, Frequency Analysis.
- Visible:** All
- Corpus List:** FalkoEssayL2WHIGv2.0, FalkoGeorgetownL2v1.0, FalkoSummaryL1v1.2, FalkoSummaryL2v1.2, FalkoWHIGL2v2.1.
- Help/Example:** make ANNIS better!
- Base text:** A grid view of the search results with columns for tokens and their parts of speech.
- Context:** Left context: 5, right context: 5.
- Annotations:** ZH1 (grid), learner (grid), ZHverb (grid), falko (grid).
- Footnote:** FNHD context: 5, 2674.

# Vom Resultat zum Export

The screenshot shows a software interface for corpus search and export. On the left, a search bar contains the query `tok=/[jJ]e/ . pos=/A.*/ .2,10`. Below it, a message indicates `39 matches in 36 documents`. A 'Query Builder' button is available. At the bottom, there's a 'Corpus List' tab and a 'Search Options' tab.

On the right, the 'Query Result' panel displays search parameters:

- Exporter: GridExporter
- Left Context: 20
- Right Context: 20
- Annotation Keys: tok, pos
- Parameters: (empty)

Buttons at the bottom include 'Perform Export' (disabled), 'Cancel Export', and 'Download'. A status message on the right says: "The Grid Export each annotation To display only". A help section defines *metakeys* and *numbers*.

Name	Texts	Tokens		
DDD-Tatian	245	54,677		

# Export

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

The screenshot shows the ANNIS interface with the following details:

- Query Builder:** A search query `n="sein" .* POS=/AD.*/` is entered.
- Results:** 3078 matches in 154 documents.
- Export Dialog:** The "Export" tab is selected. The "Exporter" dropdown is set to "GridExporter".
  - Left Context:** Set to 10.
  - Right Context:** Set to 10.
  - Annotation Keys:** Empty.
  - Parameters:** Empty.
- Description:** A tooltip explains the Grid Exporter: "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."
- Buttons:** Perform Export, Cancel Export, Download.
- Status:** "exported 350 items in 35,08 s"

# Bedingte Formatierung in Excel

Screenshot of an Excel spreadsheet showing conditional formatting rules.

The spreadsheet contains columns A, B, C, and D. Column A contains row numbers (0, 1, 2, 3, 4, 5, 6) and tokens (tok). Column B contains the tokens, and column D contains their POS tags. Rows 0, 1, 2, 3, 4, and 6 have pink background color applied via conditional formatting.

A screenshot of the "New Conditional Formatting Rule" dialog box is overlaid on the spreadsheet:

- Formatvorlage:** Klassisch
- Formel für die Ermittlung der zu formatierenden Zellen verw...**: =Istleer(A1)
- Formatieren mit:** Hellrote Füllung mit dunkler...

The dialog box has "OK" and "Abbrechen" buttons at the bottom.

# Baumbanken in ANNIS

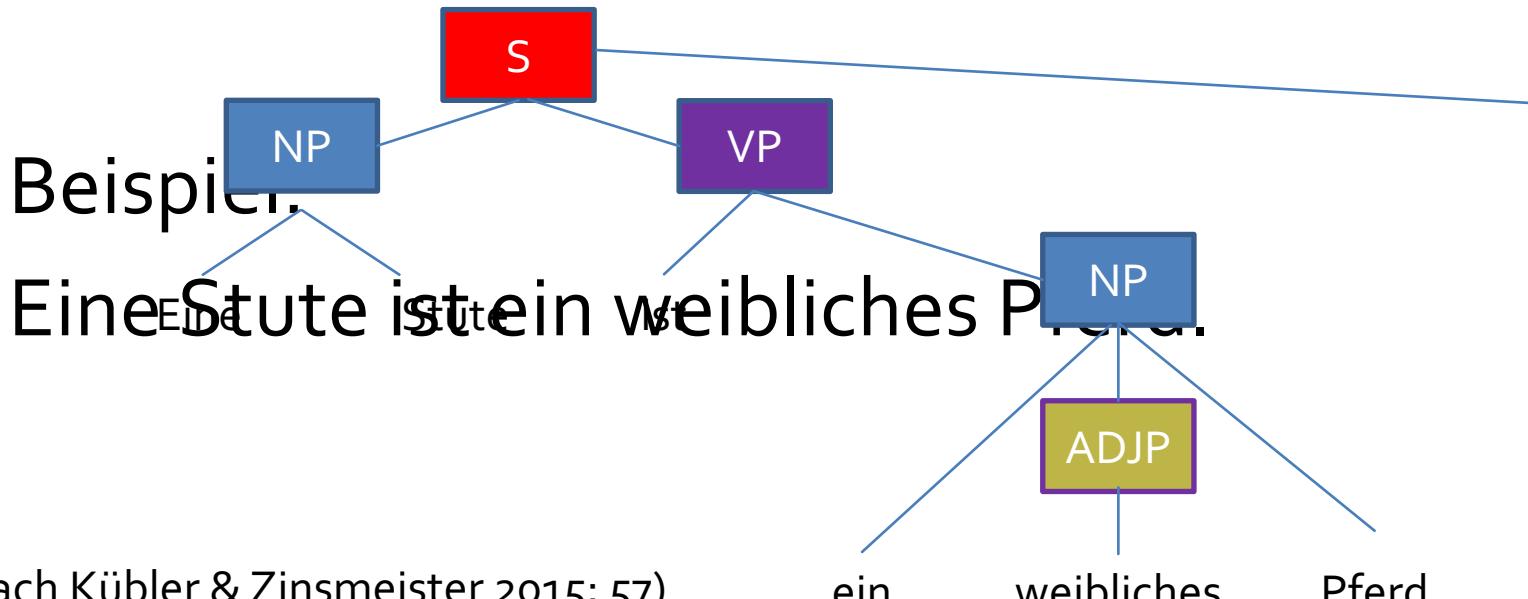
# Baumbanken

---

- 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)

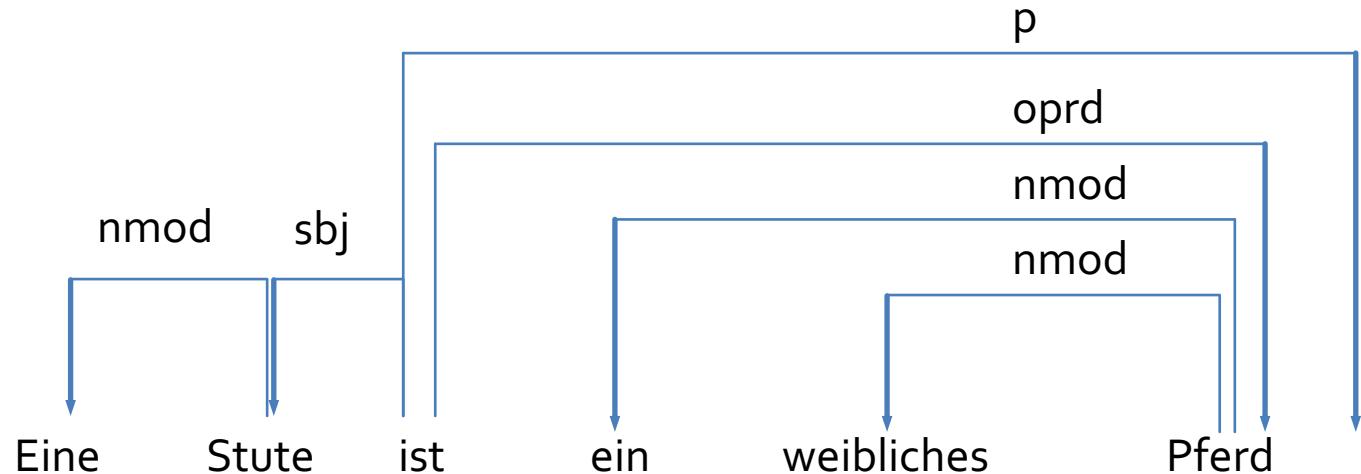
# Konstituentenstruktur und Dependenzstruktur

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



# Konstituentenstruktur und Dependenzstruktur

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



# 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.

# 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

---

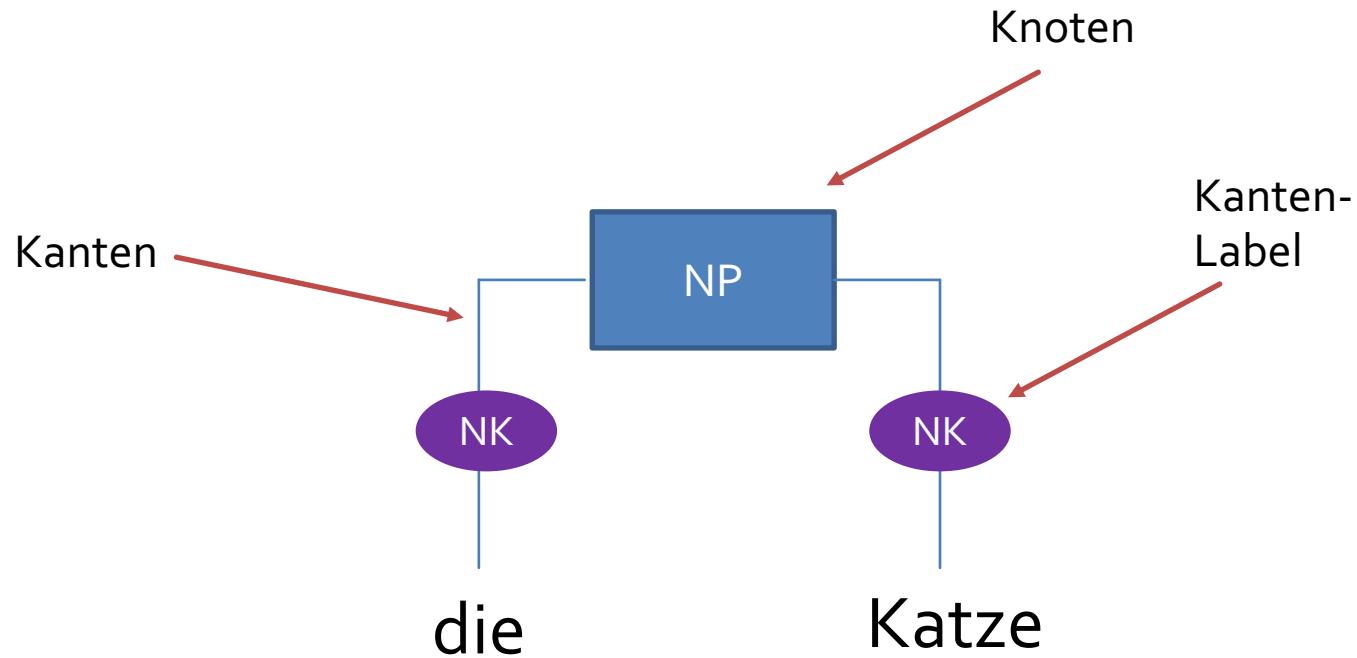
- Konstituente ist kategorieneutrale Beschreibung einer Phrase
- Eine Phrase ist immer einer bestimmten Kategorie zugeordnet, z.B. VP, NP
- Außer mit Phrasen arbeitet man in der syntaktischen Analyse häufig auch mit **Chunks**
- Chunks sind prosodische Sprecheinheiten (in gesprochener Sprache an kleinen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

# Historische Baumbanken des Deutschen

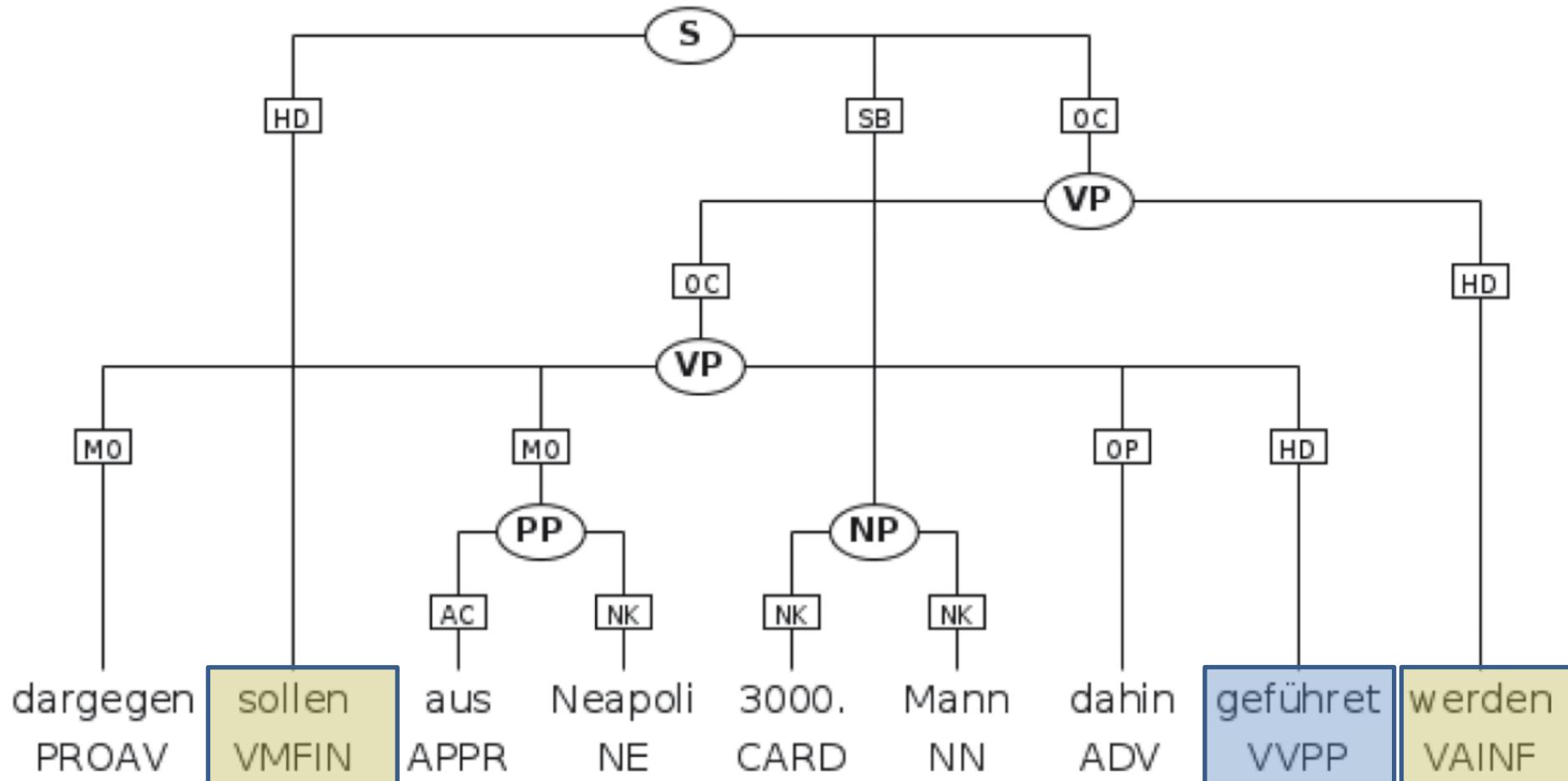
---

- 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.

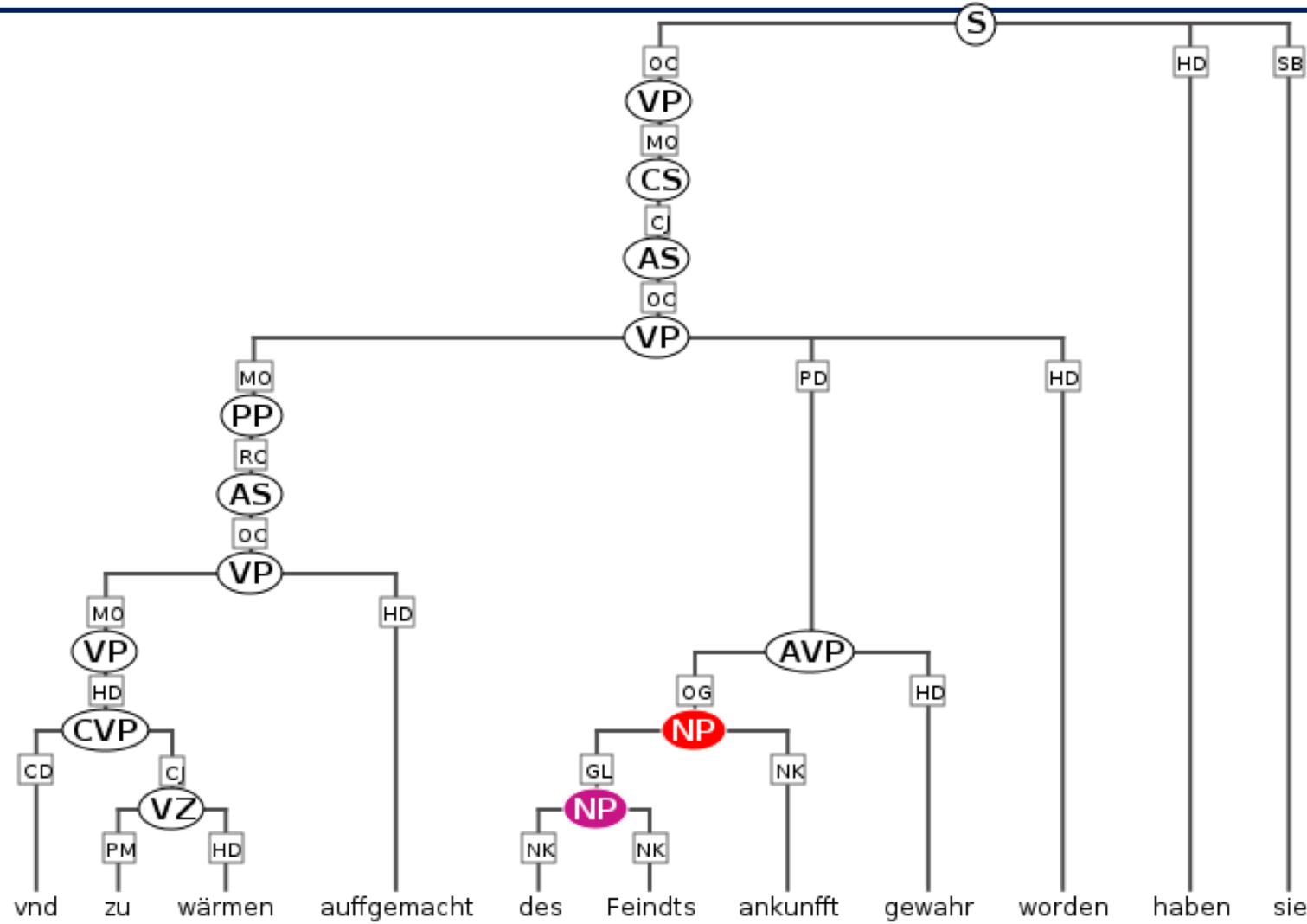
# Strukturbau: Beispiel



# Kreuzende Kanten



# Bsp.: Mercurius-Baumbank



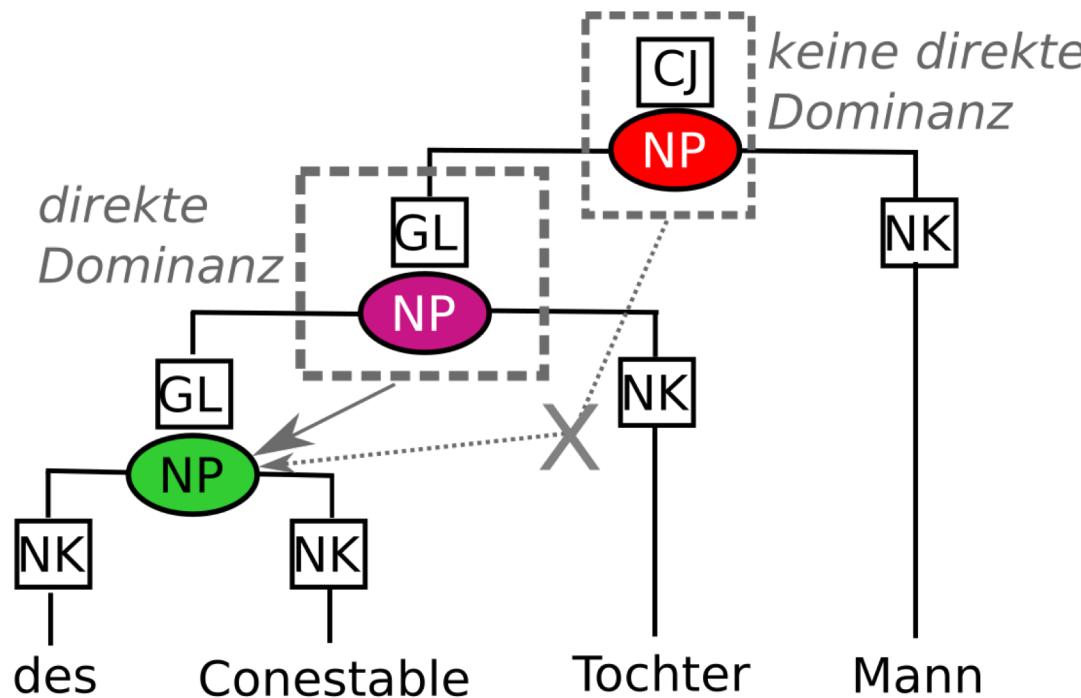
# Bsp: Mercurius-Baumbank

- 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"