

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

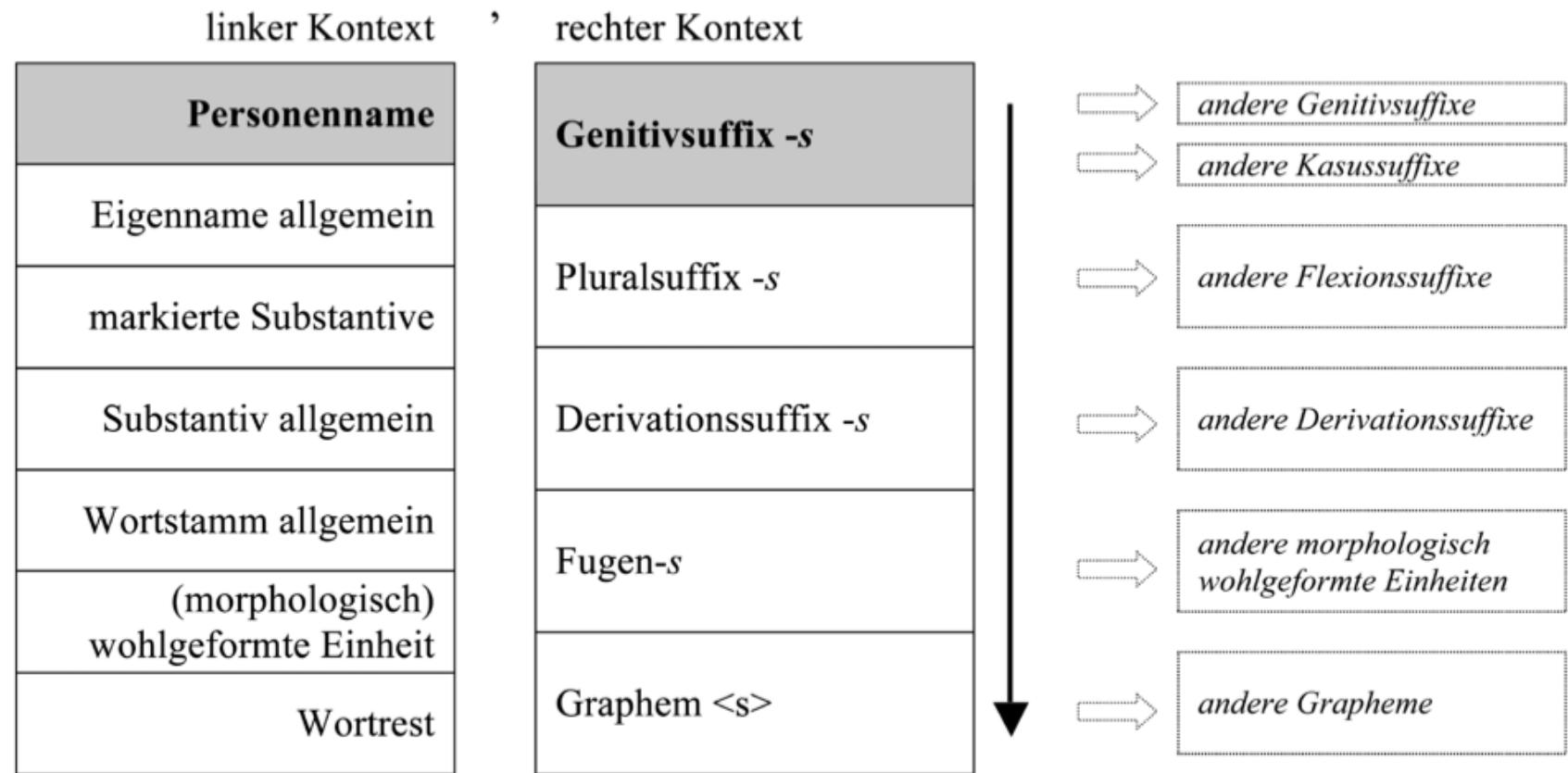
Ergebnis

- 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

-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

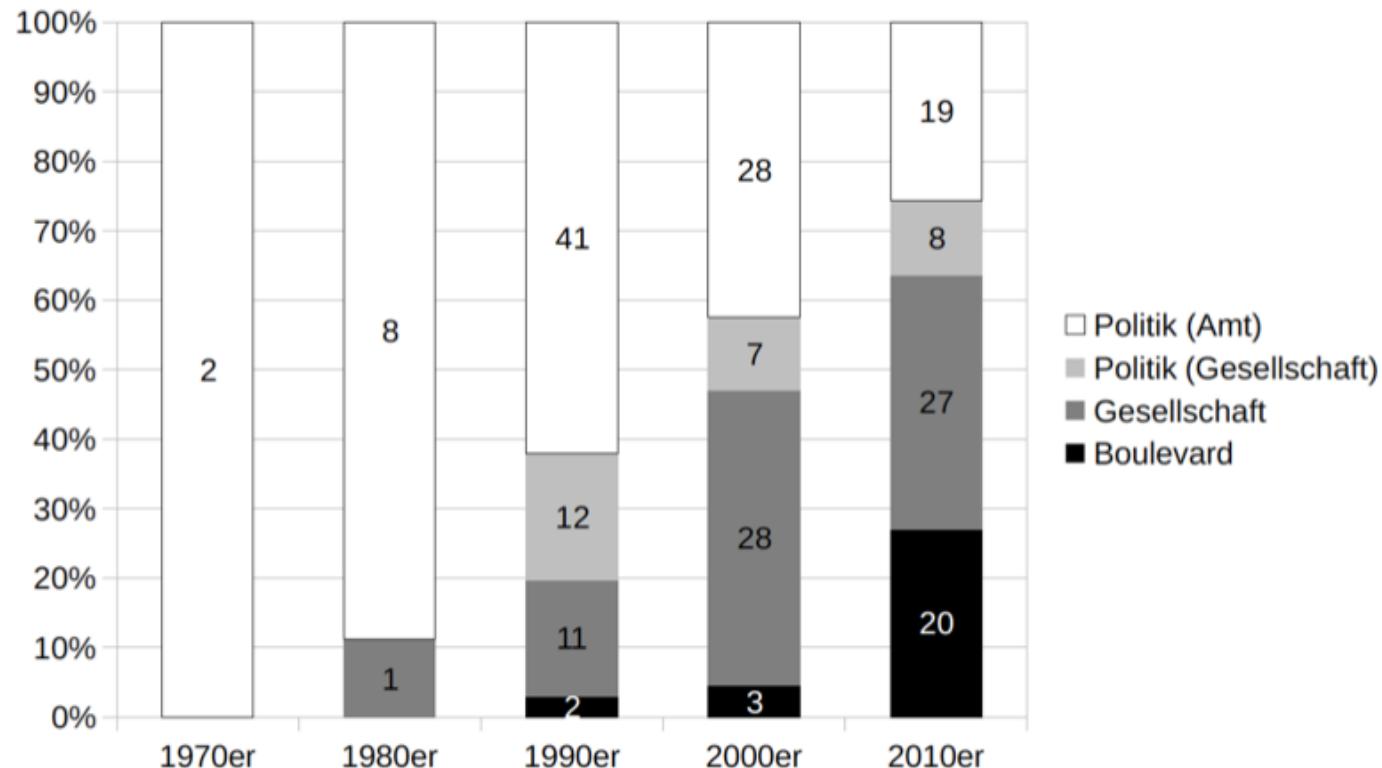


Abb. 5a. Deutsche Erstbelege (Entlehnungen und dt. Bildungen) in DeReKo/ZEIT nach Skandalfeld
(n = 217).

aus Flach, Kopf & Stefanowitsch (im Ersch.)

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:

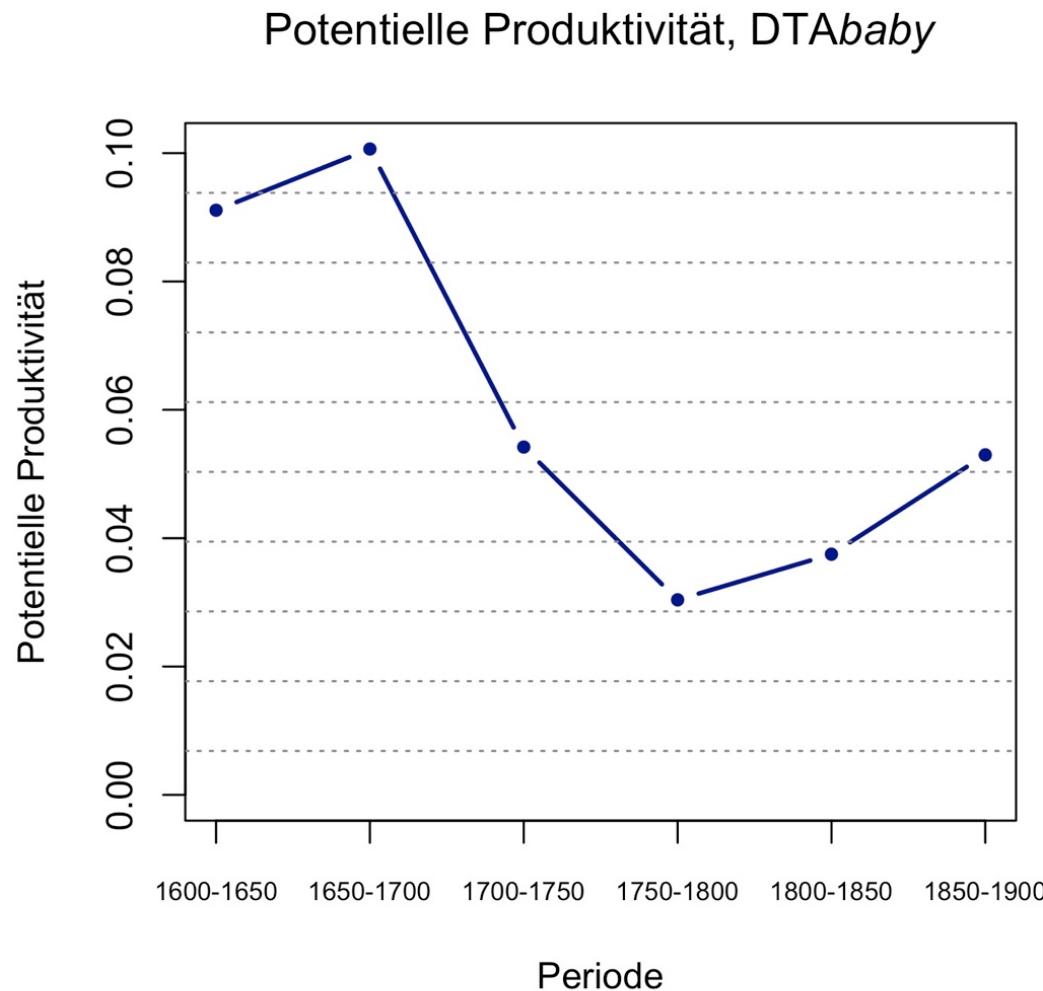
Anzahl der Hapaxe, die zum
Wortbildungsmuster gehören

(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

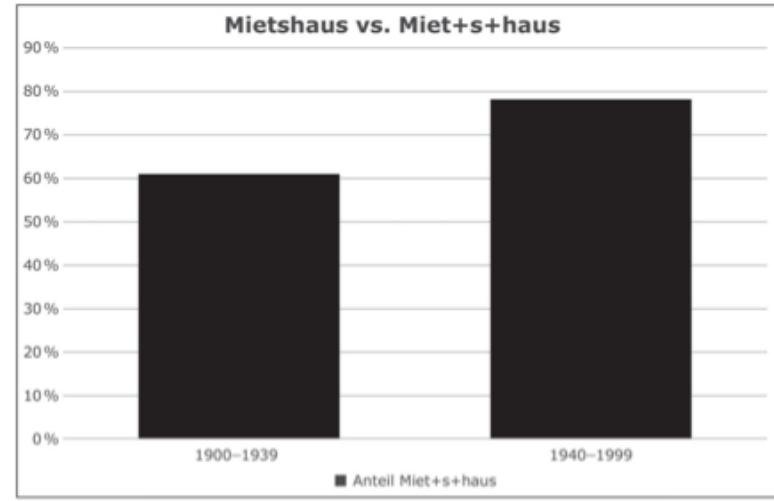
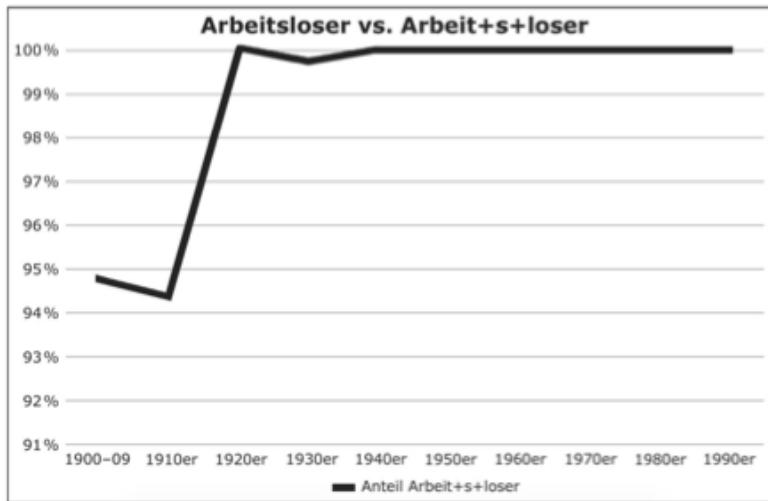


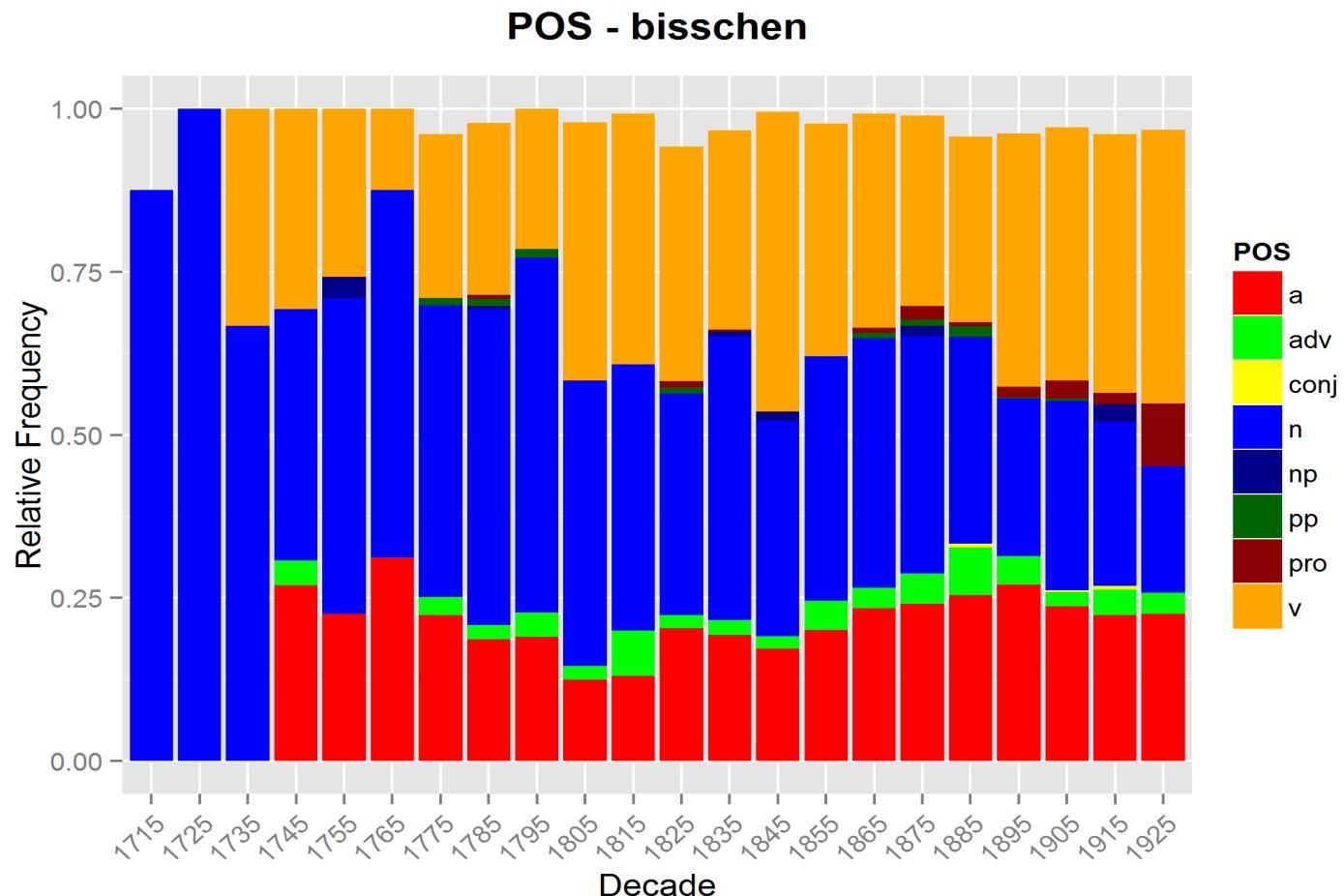
Tabelle 6: *Die Häufigkeit des Fugen-s nach präfigierten Erstgliedern.*

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)

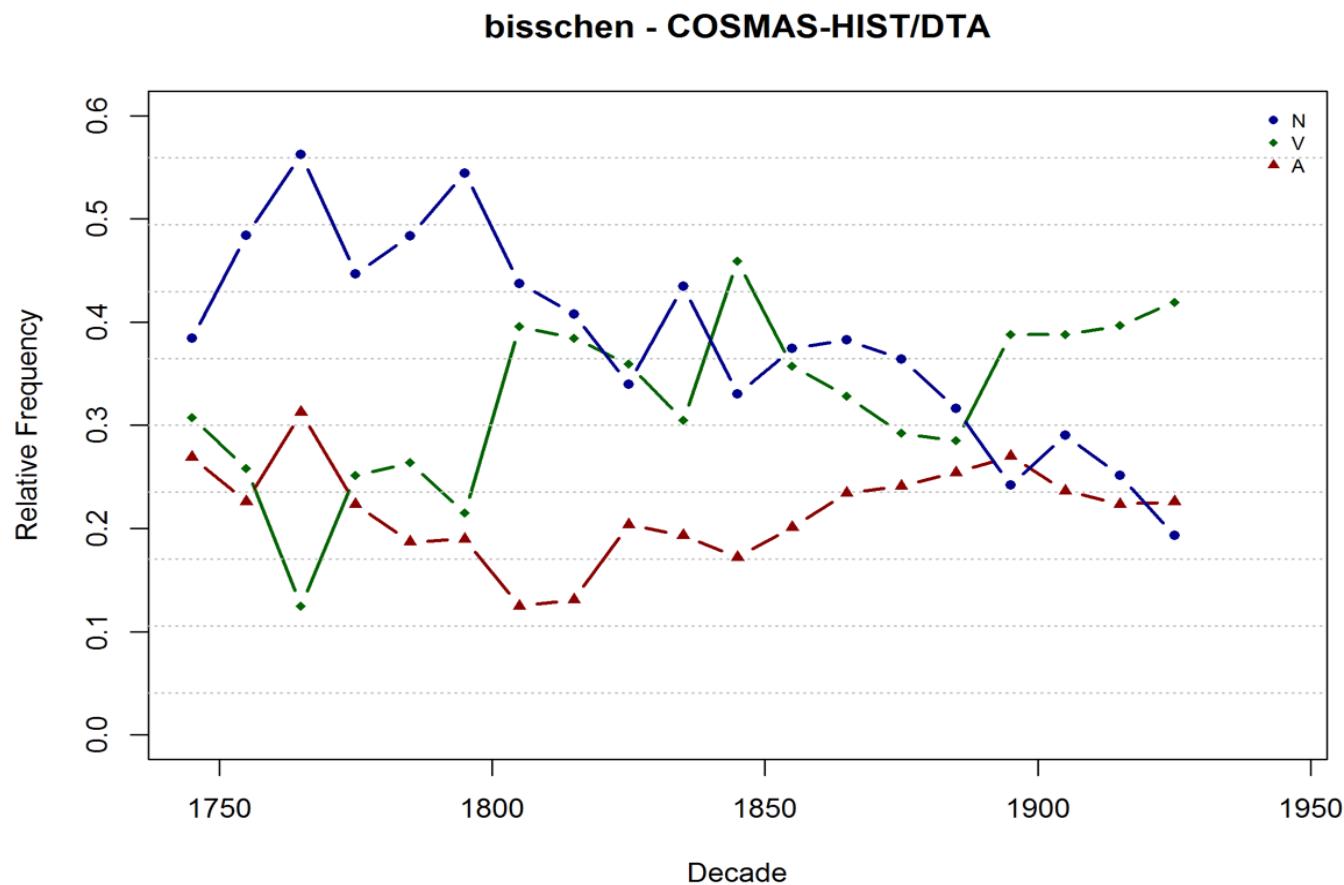
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)

- 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

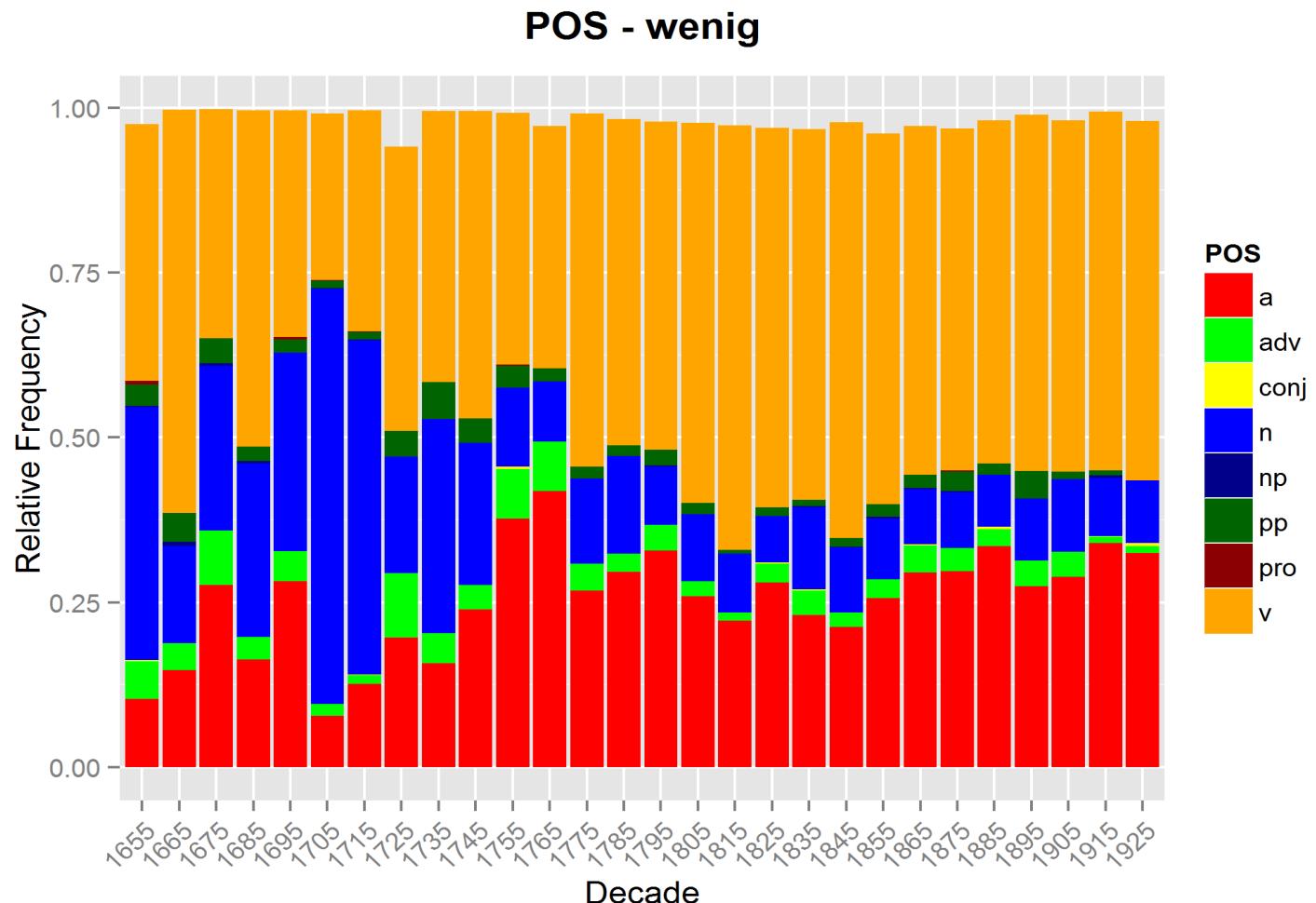
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



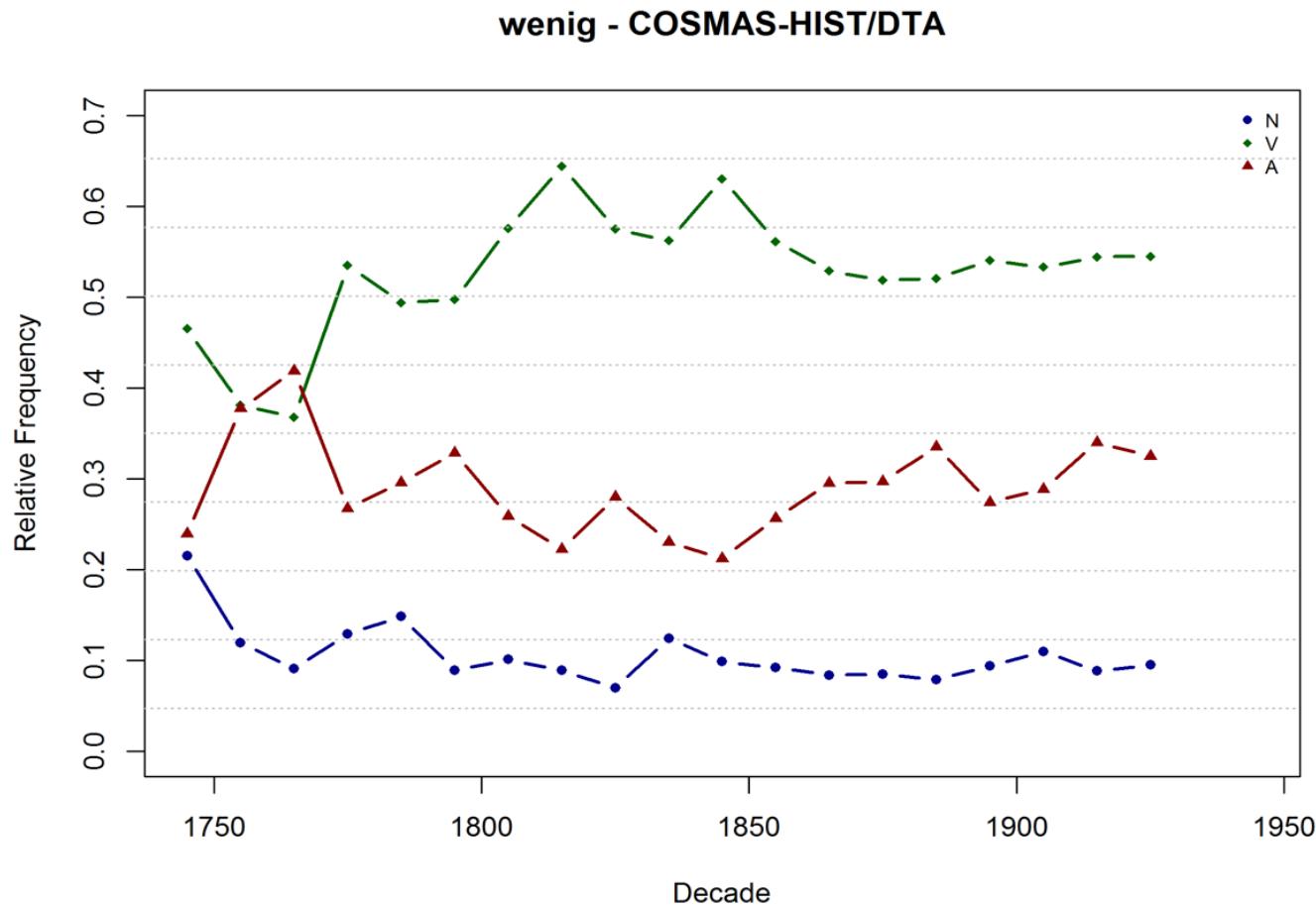
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



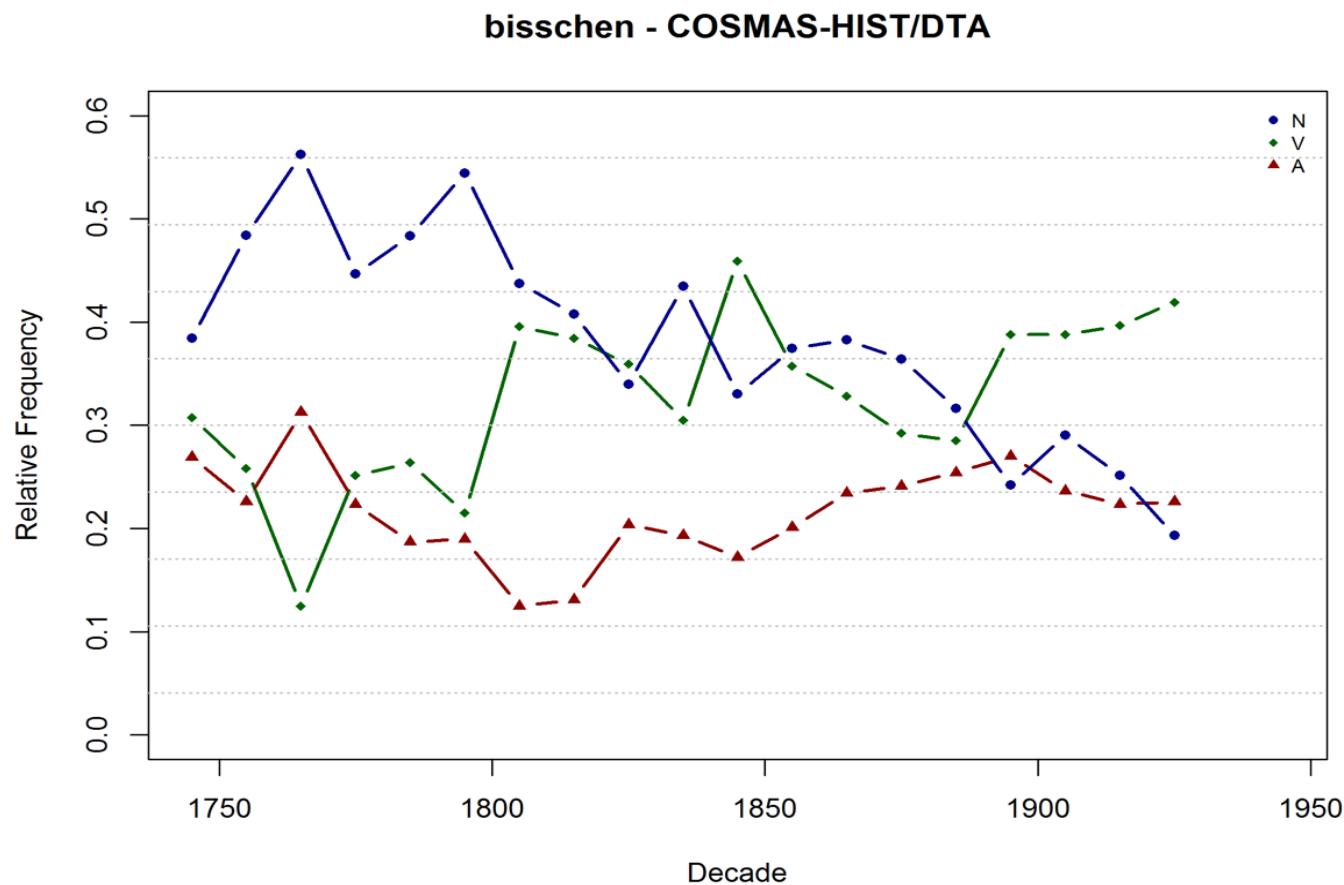
Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Beispiel Gradmodifikatoren (Neels & Hartmann 2017)



Explorative Methoden

- Unterscheidung zwischen **hypotesentestenden** und **explorativen** Methoden
- Zur Erinnerung: Hypothesentesten
 - Ich formuliere eine **Nullhypothese**...
 - ... und versuche sie zu **falsifizieren**
- Explorativer Ansatz:
 - Keine explizite Hypothese
 - stattdessen: Ich versuche, **aus den Daten selbst** Muster zu erkennen

Bag-of-words vector space model

Das mit der erfolgreichen Eroberung und er erzählt von der Eroberung und als Mittel zur Eroberung und » Es geht um Eroberung , vielleicht auch

"Zivilisierung" des "wilden Westens" einhergehende Bild des aufrechten eines halben Kontinents . « fremden Bodens zu führen . Zivilisierung Zivilisierung Zivilisierung .

Dabei ist es doch die eigene Kultur und Kultur ist Seele , ist Zeit ; Was wir menschliche Kultur und menschliche Die europäische

Zivilisation Zivilisation Zivilisation Zivilisation , die sich als zu schwach erweist ist Geist , ist Raum . nennen , sind Manifestationen solch einer Welt . hat den Endsieg über alle anderen Kulturen

erfolgreich

Eroberung

Boden

wild
Kontinent

Westen

füh

ramification
grassation
enjoyment

Seele

Kultur

Manifestation

menschlich

Raum

erweisen

europäisch

Zeit

Welt



Wie weihnachtlich ist Spekulatius?

- Datenbasis: DECOLW14AX
- Suche nach 9 Lemmata:
 - *Banane*
 - *Bonbon*
 - *Christstollen*
 - *Keks*
 - *Lebkuchen*
 - *Marzipan*
 - *Schokoriegel*
 - *Spekulatius*
 - *Zimt*

Wie weihnachtlich ist Spekulatius?

- Datenbasis: DECOW14AX
- Suche nach 9 Lemmata:
 - *Banane*
 - *Bonbon*
 - *Christstollen*
 - *Keks*
 - *Lebkuchen*
 - *Marzipan*
 - *Schokoriegel*
 - *Spekulatius*
 - *Zimt*

Wie weihnachtlich ist Spekulatius?

- insgesamt 61.115 Belege
- Methode: 5 Lemmas aus dem linken Kontext, 5 Kommas aus dem rechten Kontext (keine Satzzeichen) gehen in die Analyse ein

von Spekulatius im September , auch **Spekulatius** und Lebkuchen echt lecker sein
dies typisch Weihnachtsleckerei wie Lebkuchen **Spekulatius** oder (unknown) mögen ich so
kalt , es duften nach Lebkuchen **Spekulatius** wir schlendern über Weihnachtsmarkt und
d Stadt und da sein deutsch **Spekulatius** unter ander sehr belieben

...

Kookkurrenz-Matrix

- Matrix aus Kookkurrenzfrequenzen
- insg. 204.075 Types im linken und rechten Kontext der 9 Keywords

Lemma	Keywords									
	Banane	Bonbon	Christstollen	Keks	Lebkuchen	Marzipan	Schokoriegel	Spekulatius	Zimt	
abgucken	0	0		0	0	0	0	1	0	0
abhaben	0	0		0	0	2	0	0	0	0
abhalten	1	0		0	1	1	0	0	0	0
abhanden	0	0		0	0	0	0	1	0	0
Abhandlung	1	0		0	0	0	0	0	0	0
abhängen	1	0		0	0	1	0	0	0	0
abhängig	4	0		0	3	0	1	0	0	1
Abhängigkeit	2	0		0	0	0	0	0	0	0
abheben	0	1		0	0	0	0	0	0	0
Abheften	0	0		0	1	0	0	0	0	0
abhelfen	0	1		0	0	0	0	0	0	0
abhetzen	0	1		0	0	0	0	0	0	0
Abhilfe	0	2		1	0	0	0	0	0	1
abholen	1	3		0	13	1	1	1	0	0

Von Kookkurrenzen zu Vector Spaces

- Drei Schritte (vgl. Levshina 2015):
 - Errechnen von *Pointwise Mutual Information*-Werten durch den Vergleich von beobachten und erwarteten Kookkurrenz-Frequenzen,
 - Errechnen des Ähnlichkeitswerts mit Hilfe der Kosinus-Ähnlichkeit,
 - explorative Analyse der Ähnlichkeitswerte.

Schritt 1: (Positive) Pointwise Mutual Information (PPMI)

- $PMI(x, y) = \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) = \log_2\left(\frac{O_{xy}}{E_{xy}}\right)$
- z.B.: Keks und abhängig...

Lemma	Banane	Bonbon	Christstollen	Keks	Lebkuchen	Marzipan	Schokoriegel	Spekulatius	Zimt
abgucken	0	0	0	0	0	0	1	0	0
abhaben	0	0	0	0	2	0	0	0	0
abhalten	1	0	0	1	1	0	0	0	0
abhanden	0	0	0	0	0	0	1	0	0
Abhandlung	1	0	0	0	0	0	0	0	0
abhängen	1	0	0	0	1	0	0	0	0
abhängig	4	0	0	3	0	1	0	0	1
Abhängigkeit	2	0	0	0	0	0	0	0	0
abheben	0	1	0	0	0	0	0	0	0
Abheften	0	0	0	1	0	0	0	0	0
abhelfen	0	1	0	0	0	0	0	0	0
abhetzen	0	1	0	0	0	0	0	0	0
Abhilfe	0	2	1	0	0	0	0	0	1
abholen	1	3	0	13	1	1	1	0	0

Schritt 1: Positive Pointwise Mutual Information (PPMI)

- Keks und *abhängig* treten dreimal zusammen auf
- Was wäre der **erwartete** Wert bei Zufallsverteilung?
- Antwort: 0.86558412
- (... warum, erfahren wir bei der Einführung in die Statistik Anfang nächsten Jahres!)

Schritt 1: Positive Pointwise Mutual Information

- **Positive** Pointwise Mutual Information: Wo immer $PMI < 0$, wird der Wert auf 0 gesetzt
- (Grund: bessere Ergebnisse, vgl. Bullinaria & Levy 2007)

Schritt 1: Positive Pointwise Mutual Information

- Was sagt uns der PPMI-Wert?
- Beispiel *Keks* und *Banane*: erwartet: 0.87, beobachtet: 3
- Formel also: $\text{PPMI} = \log_2(3 / 0.87) = 1.79$
- Je höher der beobachtete Wert im Vergleich zum erwarteten, desto höher PPMI
- Ist der beobachtete Wert kleiner als der erwartete, so ist $\text{PMI} < 0$, also $\text{PPMI} = 0$.

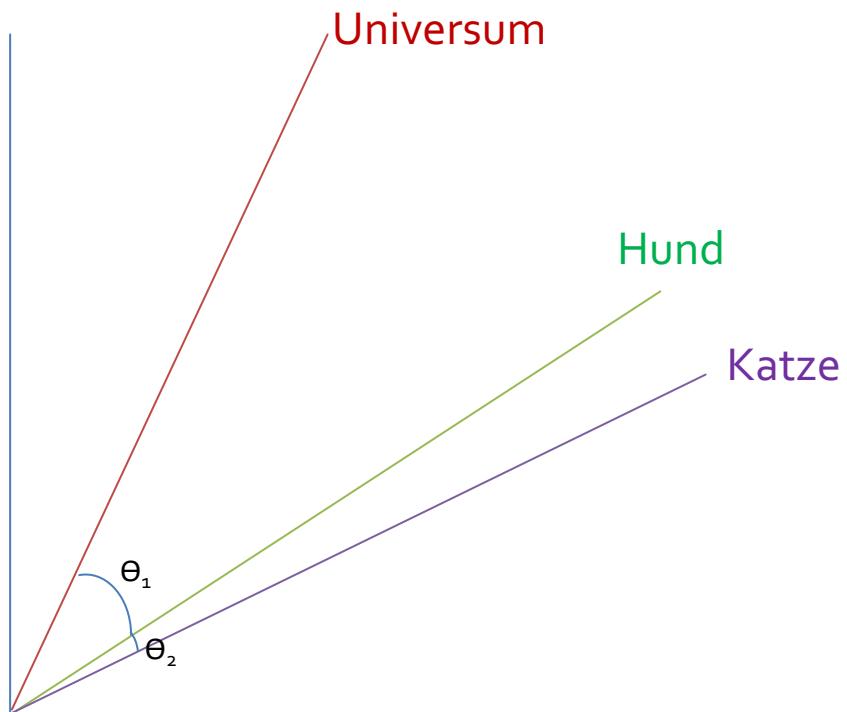
Schritt 2: Kosinus-Ähnlichkeit

- Wir haben nun eine Reihe von **Vektoren** mit PPMI-Werten

quadratisch	anfertigen	Podcast	nahend	Gesine	Paniermehl	mißlingen
1.397237	0.000000	1.397237	0.000000	0.000000	0.000000	0.000000

- hier: winziger Ausschnitt aus dem Vektor für *Banane*
- Die Vektoren werden quasi in Winkel überführt

Schritt 2: Kosinus-Ähnlichkeit



Fiktive Distributionsvektoren nach Levshina (2015)

Schritt 2: Kosinus-Ähnlichkeit

$$\cos(\theta) = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n (a_i)^2} \cdot \sqrt{\sum_{i=1}^n (b_i)^2}}$$

- Wert reicht von -1 (genau entgegengerichtet) bis 1 (genau gleichgerichtet)

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Manuelle Inspektion der Resultate:

```
> round(christmas.cos,2)
```

	Banane	Bonbon	Christstollen	Keks	Lebkuchen	Marzipan	Schokoriegel	Spekulatius	Zimt
Banane	1.00	0.05		0.03	0.05	0.04	0.04	0.06	0.04 0.05
Bonbon	0.05	1.00		0.04	0.06	0.05	0.06	0.07	0.04 0.04
Christstollen	0.03	0.04		1.00	0.04	0.12	0.07	0.05	0.09 0.03
Keks	0.05	0.06		0.04	1.00	0.06	0.05	0.07	0.06 0.03
Lebkuchen	0.04	0.05		0.12	0.06	1.00	0.08	0.06	0.12 0.05
Marzipan	0.04	0.06		0.07	0.05	0.08	1.00	0.05	0.05 0.07
Schokoriegel	0.06	0.07		0.05	0.07	0.06	0.05	1.00	0.05 0.03
Spekulatius	0.04	0.04		0.09	0.06	0.12	0.05	0.05	1.00 0.04
Zimt	0.05	0.04		0.03	0.03	0.05	0.07	0.03	0.04 1.00

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

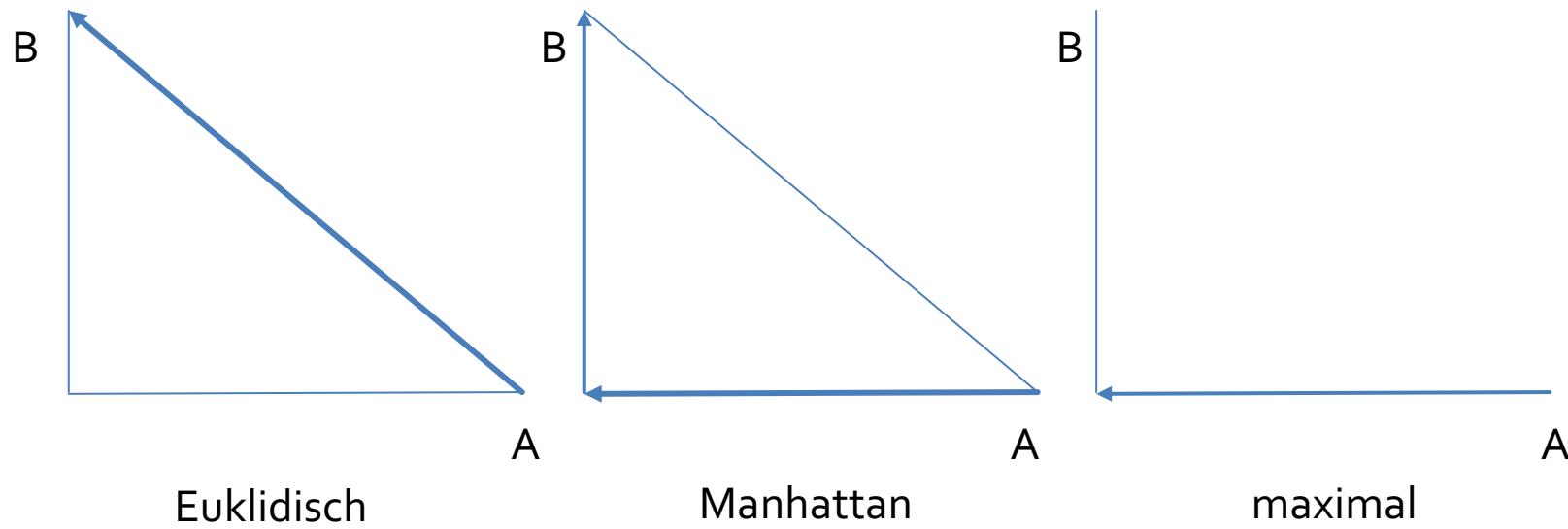
- Ähnlichkeitswerte können in **Distanzen** überführt werden.
 - Einfachste Möglichkeit: $1 - \text{Ähnlichkeitswert}$
 - zwecks Normalisierung jedoch besser geeignet:
 $1 - (\text{aktueller Ähnlichkeitswert} / \text{maximaler Ähnlichkeitswert} < 1)$

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Wir wollen die Wörter nun in **Clustern** anordnen.
- Wie können wir die Cluster aus den Daten identifizieren?
- Lösung: "Partitioning around medoids"

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- PAM teilt die Elemente um Medoide herum auf, wobei ein Medoid ein zentrales Exemplar ist, von dem aus die Distanz zu allen anderen Mitgliedern des Clusters minimal ist
- Distanz kann auf verschiedene Weise bestimmt werden



Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Anzahl der Cluster lässt sich manuell bestimmen:
 - *Banane*
 - *Bonbon*
 - *Christstollen*
 - *Keks*
 - *Lebkuchen*
 - *Marzipan*
 - *Schokoriegel*
 - *Spekulatius*
 - *Zimt*

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Anzahl der Cluster lässt sich manuell bestimmen:
 - *Banane*
 - *Bonbon*
 - *Christstollen*
 - *Keks*
 - *Lebkuchen*
 - *Marzipan*
 - *Schokoriegel*
 - *Spekulatius*
 - *Zimt*

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

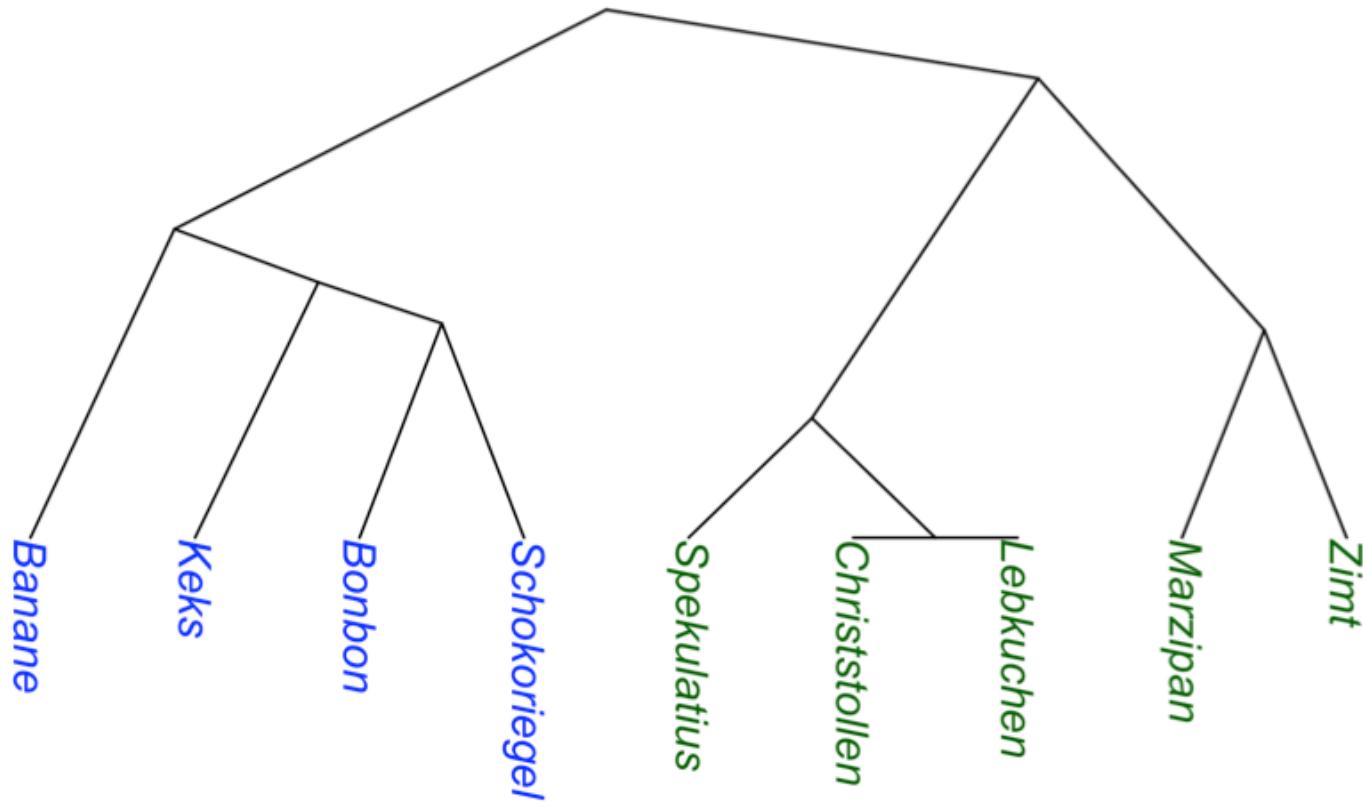
- Anzahl der Cluster lässt sich manuell bestimmen:
 - *Banane*
 - *Bonbon*
 - *Christstollen*
 - *Keks*
 - *Lebkuchen*
 - *Marzipan*
 - *Schokoriegel*
 - *Spekulatius*
 - *Zimt*

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Welche Anzahl an Clustern ist die beste?
- "Average Silhouette Width" gibt Auskunft
- zeigt, wie wohlgeformt die Cluster sind
- Wert zwischen 0 und 1
- 0: keine Clusterstruktur in den Daten
- 1: klare Struktur, vollständige Trennung zwischen den Clustern.
- bester Wert hier: zwei Cluster.

Ergebnis: Hierarchical Clustering

Cluster Dendrogram



Fazit

- Hohe Übereinstimmung zwischen "intuitivem" und datenbasiertem Clustering
- Daher interessante Methode gerade für ältere Sprachstufen, bei der "intuitive" Urteile über die Semantik von Wörtern und Konstruktionen nur sehr bedingt möglich sind.

Übung: ngram viewer / DWDS

- Einfache Aufgabe: Wir replizieren die Suche von Nübling/Szczepaniak zu *Arbeitsloser* vs. *Arbeitloser* im Google ngram viewer und in DWDS



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's a dark bar with the text "Anmelden (DTAQ)". Below it is the DTA logo. On the right side, there are social media icons for Twitter and Google+. The main search area has a large input field and a "suchen" button. Below the input field are three radio buttons for search scope: "in den Titeldaten" (selected), "im Korpus", and "in der Dokumentation". To the right of these are links for "Hilfe" and "Beispielanfragen". A code snippet for a search query is also displayed.

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
- 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

- Aufgrund der besseren Exportmöglichkeiten benutzen wir für unsere Beispielanfragen eins.dwds.de
- Suchsyntax auf der Homepage.

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

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

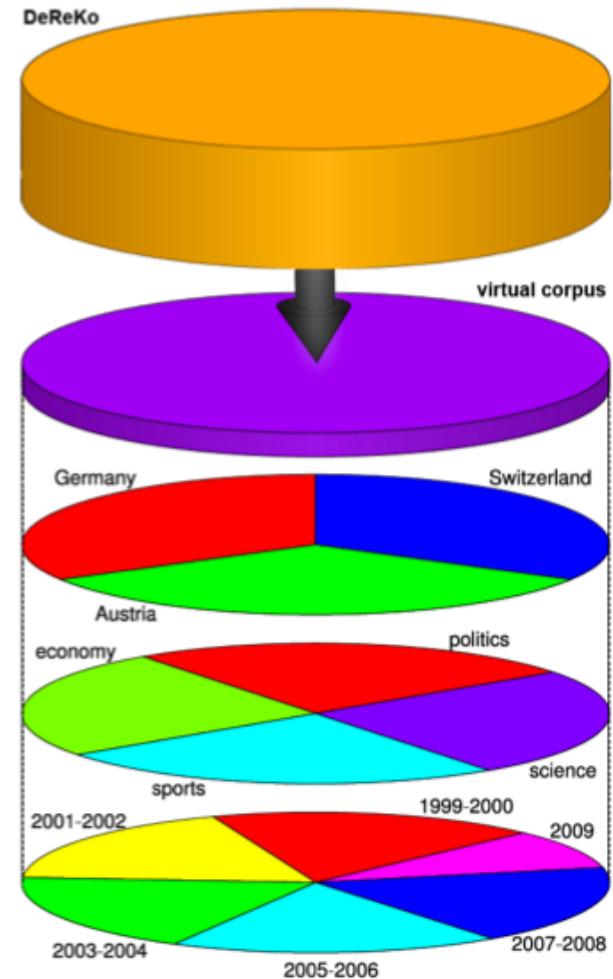


Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

(aus Kupietz et al. 2010)

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 lemmatisiert oder 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

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

ANNIS

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

ANNIS Query Language (AQL)

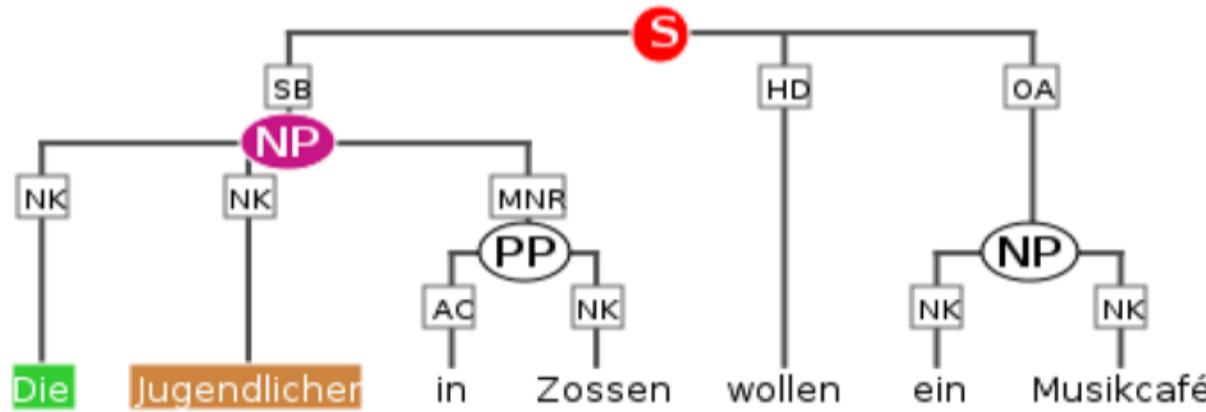
- Ähnlich logisch wie CQL...
- ...aber (m.E.) etwas komplizierter.
- Vorteil: extrem detaillierte syntaktische Suchanfragen möglich.
- vgl. "Cheat Sheet für ANNIS"



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 ANNIS web interface. On the left, there is a red box highlighting the 'Eingabefeld' (Input Field) where users can enter AQL queries. On the right, there is a red box highlighting the 'Hilfe/Tutorial' (Help/Tutorial) section, which contains a link to a tutorial and example queries. Below these, another red box highlights the 'Korpora' (Corpora) section, which displays a list of available corpora (mo, mu) with their respective text and token counts.

Eingabefeld

Hilfe/Tutorial

Korpora

Please enter AQL query

Query Builder

Example Queries

Example Query Description

Q " " search for the word " " open corpus browser

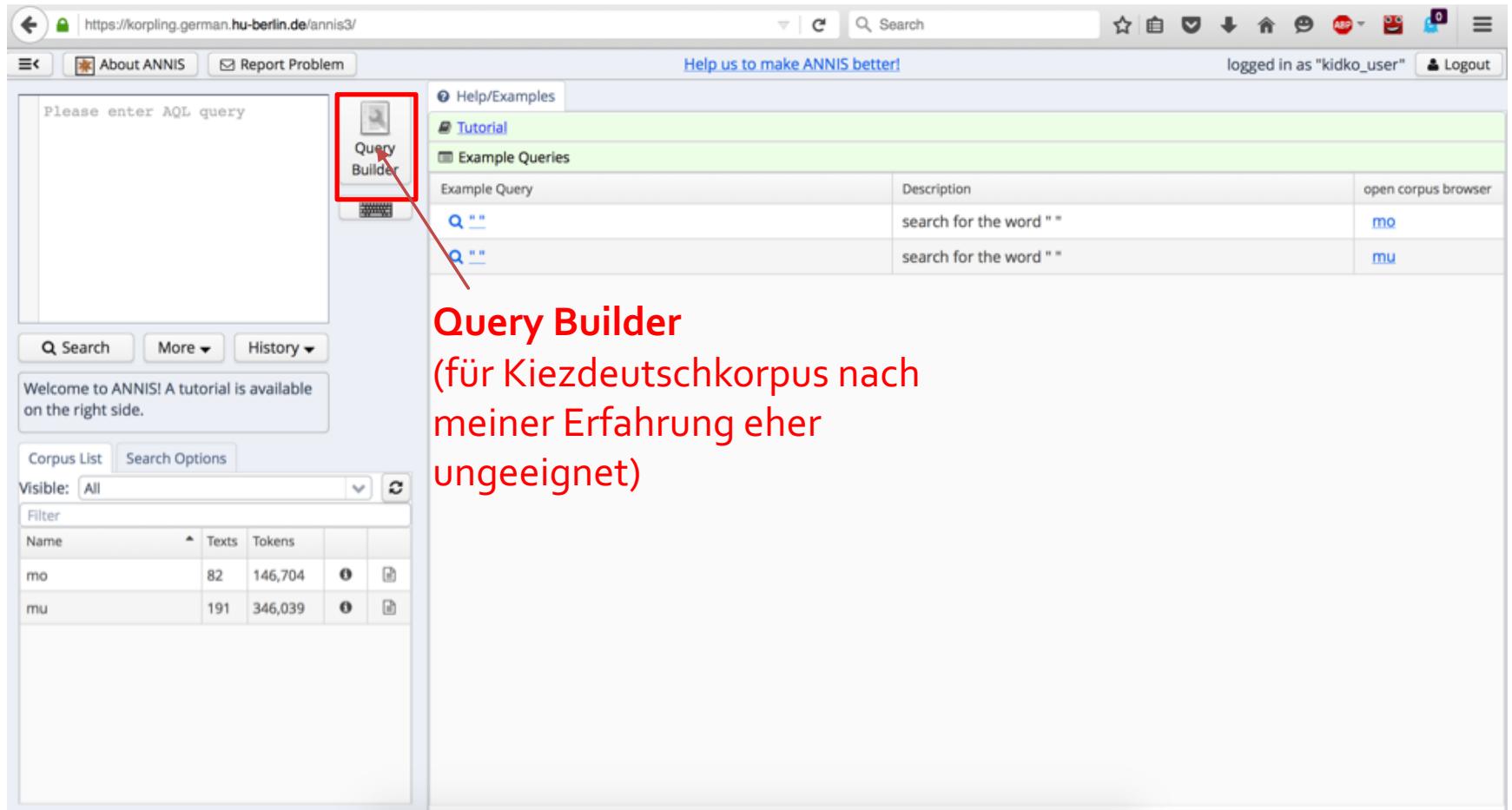
Q " " search for the word " " mo

mu

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

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

ANNIS Query Language



The screenshot shows the ANNIS3 web interface at <https://korpling.german.hu-berlin.de/annis3/>. A red box highlights the 'Query Builder' button in the top navigation bar, which is accompanied by a red arrow pointing towards the main content area. The main content area features a red title 'Query Builder' and a subtitle '(für Kiezdeutschkorpus nach meiner Erfahrung eher ungeeignet)'. The interface includes a search bar, a sidebar with a corpus list, and a central panel displaying example queries.

Please enter AQL query

Help us to make ANNIS better!

logged in as "kidko_user" Logout

Query Builder

Example Queries

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

Search Options

Visible: All

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

REM: Ein Korpus wählen

Corpus List Search Options

Visible: ▼ ↻

Filter	Name	Count	Size	Info	File
All	MiGraKo	5	956	i	file
	eReM	7	44,986	i	file
	11-12_1- uuu-rvv-	22	20,451	i	file
	11-12_1-rhfrhess-PV-X	5	956	i	file
	11_2-12_1-obd-PV-G	7	44,986	i	file
	12-13_1-mdnd-PV-X	4	36,281	i	file
	12-13_1-mdnd-V-X	5	33,992	i	file
	12_2-alem-PV-G	4	30,309	i	file

- 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 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

	6a,8	6a,9				6a,10				
reference	.	der	vbel	hellevvarte	.	ern	vverde	ir	geverte	.
tok_dipl	.	der	vbel	hellevvarte	.	er	n	vverde	ir	g
tok_anno	.	der	vbel	hellevvarte	.	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		29817000	174642000	71433000		40380000	119841000	225138000	83982000	55
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.*"`
- Results:** Displaying Results 1 - 2 of 2
- Result 1 (Path: mo > Mo05WD_11-2):**
 - Annotations:
 - SPK15::v WACHS nur
 - SPK39::v a sein kle:ner WILli da unten is kle:n
 - SPK67::v ich bin SCHWUL
 - Left context: 5, Right context: 5
- Result 2 (Path: mu > MuH9WT_10):**
 - Annotations:
 - MuH9WT::v doch die ganze zeit voll (SCHWUL) (halt) wallAH er soll
 - Left context: 5, Right context: 5
- Corpus List:** Visible: All
Corpora:

Name	Texts	Tokens	Annotations	Actions
mo	82	146,704	View Edit	Delete
mu	191	346,039	View Edit	Delete

Wortfolgen suchen

.* = geht indirekt voran

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

- Query:** `n="schwul" .* POS=/AD.*`
- Results:** Displaying Results 1 - 10 of 204
- Result 1 (Path: mo > Mo05WD_11-2 (tokens 1920 - 1931)):**
 - Left context: 5, right context: 5
 - Table:

SPK15::v	WACHS	nur	a	sein	kle:ner	WILli	da	unten	is	kle:n
SPK39::v			ich	bin	SCHWUL					
annotation										
- Result 2 (Path: mo > Mo05WD_11-2 (tokens 1920 - 1933)):**
 - Left context: 5, right context: 5
 - Table:

SPK15::v	WACHS	nur	a	sein	kle:ner	WILli	da	unten	is	kle:n	(-)
SPK39::v			ich	bin	SCHWUL						
annotation											
- Result 3 (Path: mo > Mo05WD_11-2 (tokens 1920 - 1934)):**
 - Left context: 5, right context: 5
 - Table:

SPK15::v	WACHS	nur	a	sein	kle:ner	WILli	da	unten	is	kle:n	(-)
SPK39::v			ich	bin	SCHWUL						ich
annotation											
- Result 4 (Path: mo > Mo05WD_11-2 (tokens 1920 - 1936)):**
 - Left context: 5, right context: 5
 - Table:

SPK15::v	WACHS	nur	a	sein	kle:ner	WILli	da	unten	is	kle:n	(-)
SPK39::v											
annotation											

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

The interface includes a Query Builder button, search history, and various context settings like left and right context sizes of 5 tokens.

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:

- Query Builder:** A sidebar on the left containing two search queries:
 - `tok=/[jJ]e/ . pos=/A.*/ .2,10`
 - `tok=/[dD]esto|[jJ]e|[Uu]mso/`
- Help/Example:** A link to help examples.
- Base text:** A section showing the results of the search queries. It includes:
 - Result for: `tok=/[jJ]e/ . pos=/A.*/ .2,10 tok=/[dD]esto|[jJ]e|[Uu]mso/`
 - left context: 5 right context: 5
 - Sample text:
 - 1 1 handlungen gibt , desto gewaltiger wird die Gesellschaft .
IN VVFIN \$, KON ADJD VAFIN ART NN \$.
 - mehr gew
 - mehr gew
 - ADV ADJ
 - Expansion of ZH1 (grid):
 - mus argumentiert , desto negativer wird vielleicht diese Öffentlichkeit
 - VVPP \$, KON ADJD VAFIN ADV PDAT NN
 - Expansion of learner (grid):
 - mus argumentieren , desto negativ werden vielleicht dies Öffentlichkeit
 - VVPP \$, KON ADJD VAFIN ADV PDAT NN
 - Expansion of ZH2 (grid):
 - mus argumentieren , desto negativ werden vielleicht diese Öffentlichkeit
 - VVPP \$, KON ADJD VAFIN ADV PDAT NN
 - Expansion of ZHverb (grid):
 - eigenen Materialien beschäftigt , desto hat man grösere
 - eigen Material beschäftigen , desto haben man groß
 - ADJA NN VVPP \$, KON VAFIN PIS ADJA
 - Expansion of falko (grid):
 - wird
 - werden
 - VAFIN \$
- Search:** Buttons for Search, More, and History.
- Export:** A button to download the results.
- Frequency Analysis:** A button to perform frequency analysis.
- Corpus List:** A list of available corpora: FalkoEssayL2WHIGv2.0, FalkoGeorgetownL2v1.0, FalkoSummaryL1v1.2, FalkoSummaryL2v1.2, FalkoWHIGL2v2.1, and FNHD contexts.
- Search Options:** Buttons for Visible (All) and a dropdown menu.
- Grid View:** A table showing the results of the search queries, with columns for path, tokens, and various linguistic features.

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` and `tok=/[dD]esto|/[jJ]e|[Uu]mso/`. Below the search bar, it says "39 matches in 36 documents". At the bottom, there are tabs for "Corpus List" and "Search Options", and a "Visible: All" dropdown. A "Filter" section is present. On the right, the "Query Result" tab is active, showing settings for "Exporter: GridExporter", "Left Context: 20", "Right Context: 20", "Annotation Keys: tok, pos", and "Parameters". Buttons for "Perform Export", "Cancel Export", and "Download" are available. A blue circular progress indicator is visible.

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

39 matches
in 36 documents

Corpus List Search Options

Visible: All

Filter

Name	Texts	Tokens	
DDD-Tatian	245	54,677	

Help/Examples Query Result Export

Exporter: GridExporter

Left Context: 20

Right Context: 20

Annotation Keys: tok, pos

Parameters:

The Grid Export each annotation To display only:

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 details:

- Left Panel:** Displays a search query: `n="sein" .* POS=/AD.*/`. Below it, a message says "3078 matches in 154 documents".
- Toolbar:** Includes "About ANNIS", "Report Problem", "Query Builder", and "Logout".
- Export Dialog (Top Right):**
 - Exporter:** Set to "GridExporter" (highlighted with a red arrow).
 - Left Context:** Set to 10.
 - Right Context:** Set to 10.
 - Annotation Keys:** Empty.
 - Parameters:** Empty.
- Help Text:** Describes the Grid Exporter's function: "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", and "Download".
- Status:** "exported 350 items in 35,08 s".
- Bottom Panel:** Shows a "Corpus List" with "Visible: All" and a "Filter" table. The filter table has columns: Name, Texts, Tokens, and two icons. It contains two rows:

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

Bedingte Formatierung in Excel

Screenshot of the Microsoft Excel ribbon showing the "Formatvorlage" (Format) tab selected. A context menu is open over a row of data in column A, with the option "Formatvorlage" highlighted.

The "Formatvorlage" dialog box is displayed, titled "Neue Formatvorlagentyp".

Formatvorlage: Klassisch

Formel für die Ermittlung der zu formatierenden Zellen verwenden:

=Istleer(A1)

Formatieren mit: Hellrote Füllung mit dunkler...

AaBbCcYyZz

Buttons: Abbrechen, OK

Table Data (Rows 0 to 6):

0	tok	von Kriminalität ist nur kurzfris...	
	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] VVINF[3-3] \$.[4-4] KON[5-5] PWS[6-6] VVFIN[7-7] ADV[8-8] \$.[9-9] VVINF[10-10] \$.[11-11] \$.[12-12] PIS[13-13] VMFIN[14-14] ADV[15-15] VVINF[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	