Korpuslinguistik

COWboys im Wacky Wide Web: Webkorpora und CQP

Webkorpora

Die Offenbarung

- Internet als "a fabulous linguists' playground" (Kilgarriff & Grefenstette 2003)

Das Problem

- Einige frühe Studien benutzen Google als Korpus
- Welche Probleme bringt dieses Vorgehen mit sich?

"Googleology" (Kilgarriff 2007)

- Trefferzahlen bilden keine Tokenfrequenzen ab, sondern die Anzahl der Seiten, auf denen Suchwort gefunden wurde
- potentiell viele Duplikate
- "Boilerplate"-Text kann Frequenzen verfälschen

Zahnspange: Was essen bei neuen Bögen? Ich habe seit einer Weile meine feste Zahnspange und habe jetzt neue Bögen. Die Brackets sind noch die alten, deswegen weiß ich nicht, ob ich jetzt wie bei neuen Brackets ein paar Tage …

von tinycookie · vor 2 Min · Themen: essen, Zahnspange, Kieferorthopäde

2 Antworten

kann ich mir die bitch getten? voll bock auf die bitch.

von ObamaFanBoy007 · vor 2 Min · Themen: bitch, getten

3 Antworten

Mama ist depresiv? Meine Mutter hat Depression und ich weiß nicht was ich machen soll !! Bitte helft mir !!

von jeysammysam · vor 2 Min · Themen: Mutter, Depressiv

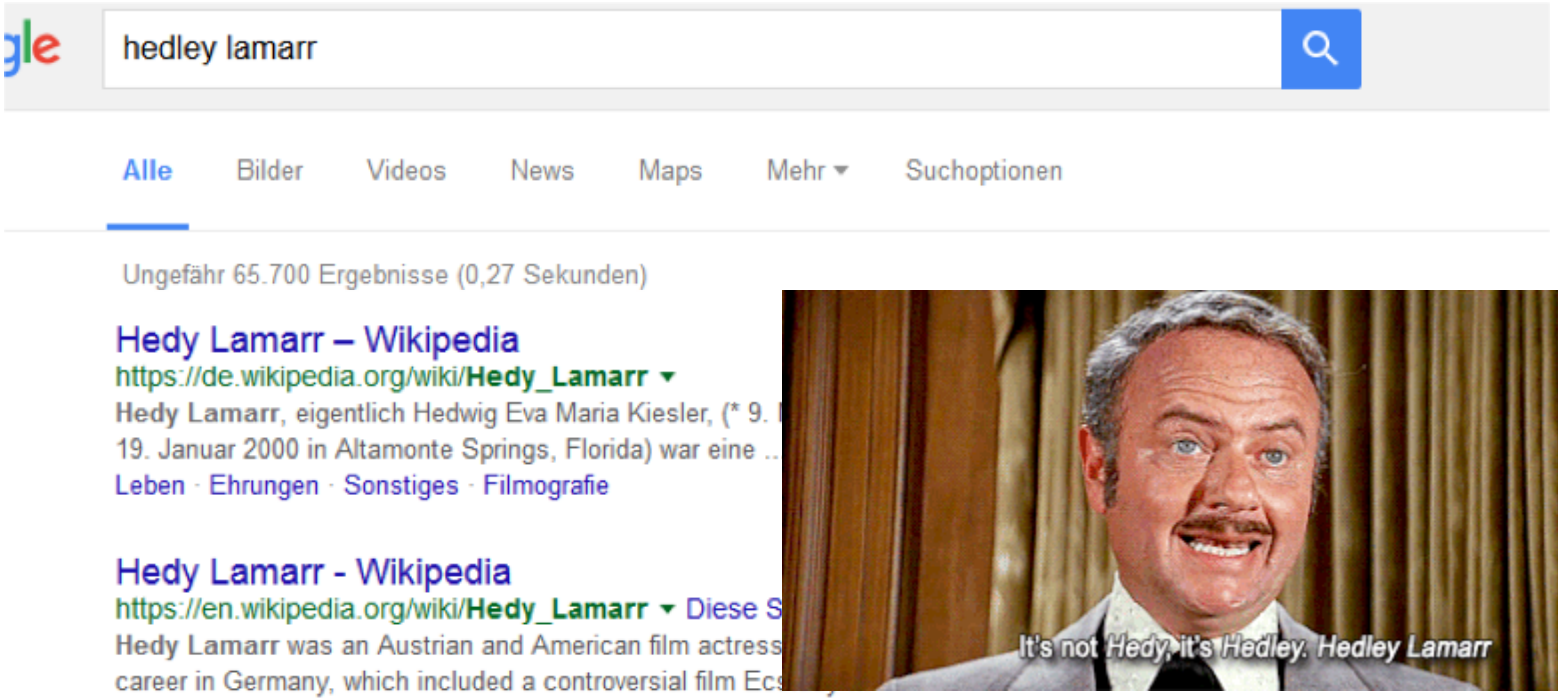
2 Antworten

Laufen gesund trotz harter Arbeit? Warum ist laufen auch dann gesund un sinnvoll, wenn man einen harten Anpackerjob hat ? Braucht der Körper nicht eher die Ruhe, oder sollte man trotz der Tatsache dass man jeden tag total kaputt …

von Kaffeemann88 · vor 2 Min · Themen: Sport, Medizin

"Googleology" (Kilgarriff 2007)

- Intransparante Algorithmen
- Teilweise Beschränkung der Trefferanzahl (wiederum aufgrund intransparenter Algorithmen)
- Ähnliche Ergebnisse werden mitgefunden

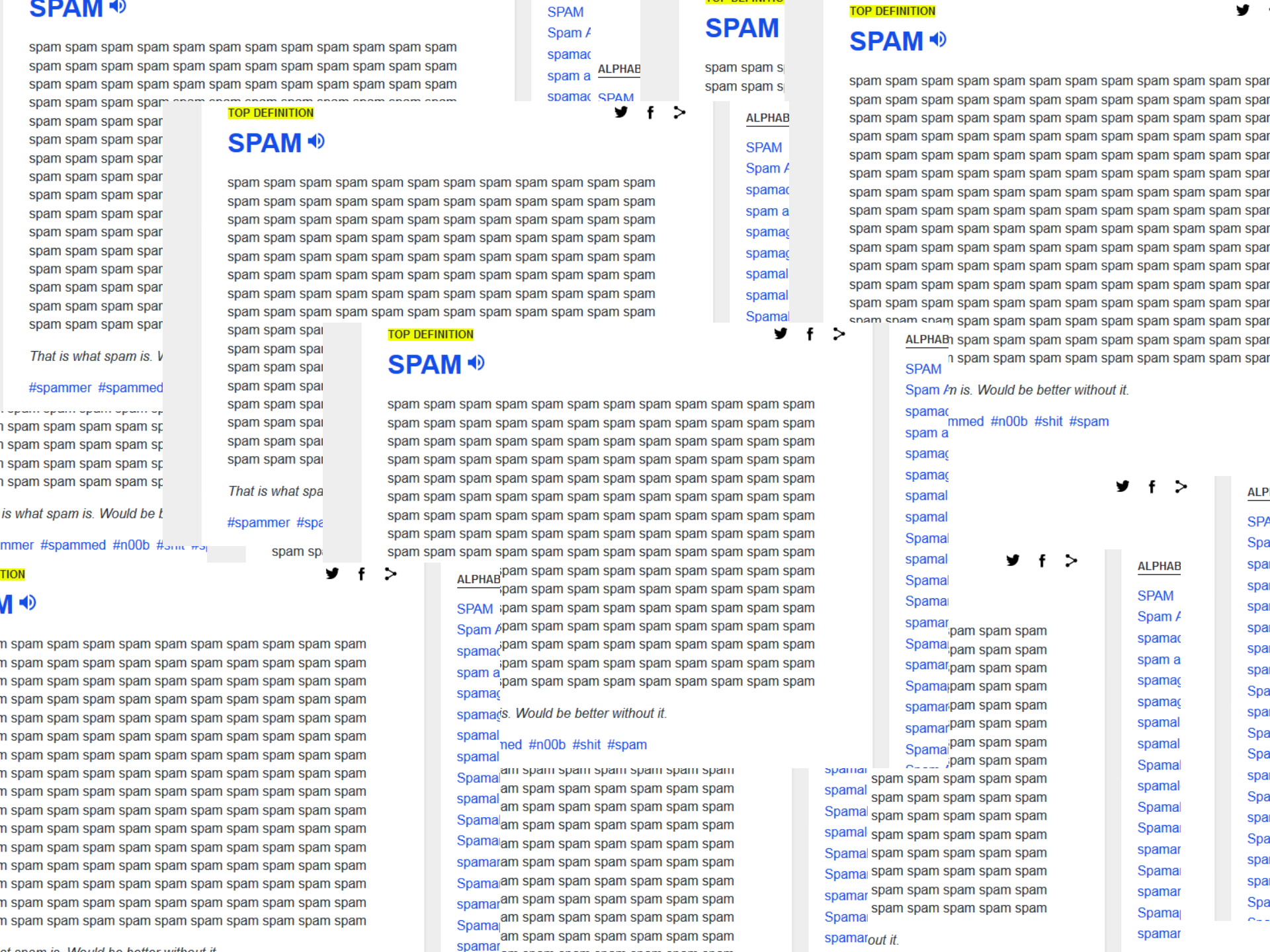


"Googleology" (Kilgarriff 2007)

- Fazit: "Googleology is bad science"

Wie kann man dieses Problem lösen?

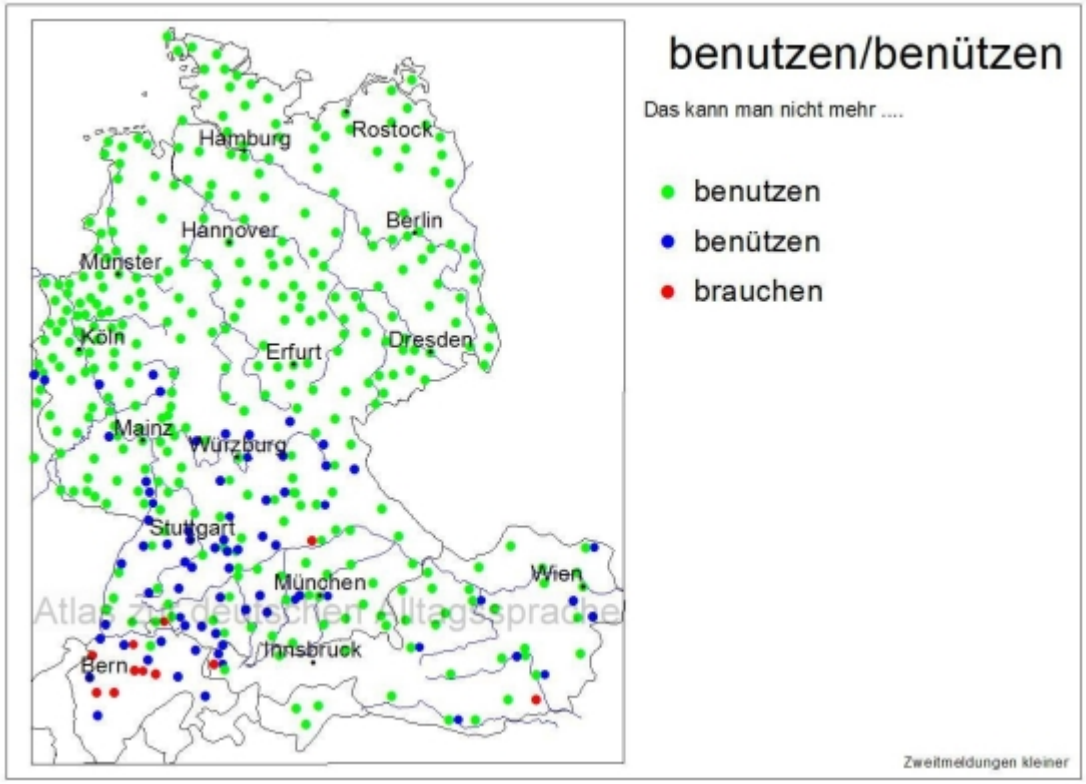
- Webkorpora: Textsammlungen aus dem Internet
- Welche Voraussetzungen müssen Webkorpora erfüllen, um die genannten Probleme zu umgehen?


SPAM
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
That is what spam is. W
#spammer #spammed
TOP DEFINITION
SPAM
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
That is what spa
#spammer #spa
TOP DEFINITION
SPAM
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
ALPHAB
SPAM
Spam A
spamac
spam a
spamag
spamag
spamal
spamal
Spamal
TOP DEFINITION
SPAM
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
spam spam spam spam spam spam spam spam spam spam spam spam
ALPHAB
SPAM
Spam A
spamac
spam a
spamag
spamag
spamal
spamal
Spamal
spamal
Spamal
spamar
Spamal
spamar
Spamal
spamar
Spamal
spamar
Spamal
spamar
ALPHAB
SPAM
Spam A
spamac
spam a
spamag
spamag
spamal
spamal
Spamal
spamal
Spamal
spamar
Spamal
spamar
Spamal
spamar
Spamal
spamar
ALP
SPA
Spa
spa
spa
spa
spa
Spa
spa
Spa
Spa
spa
Spa
spa
Spa
spa
Spa
spa
Spa

Deutsche Webkorpora

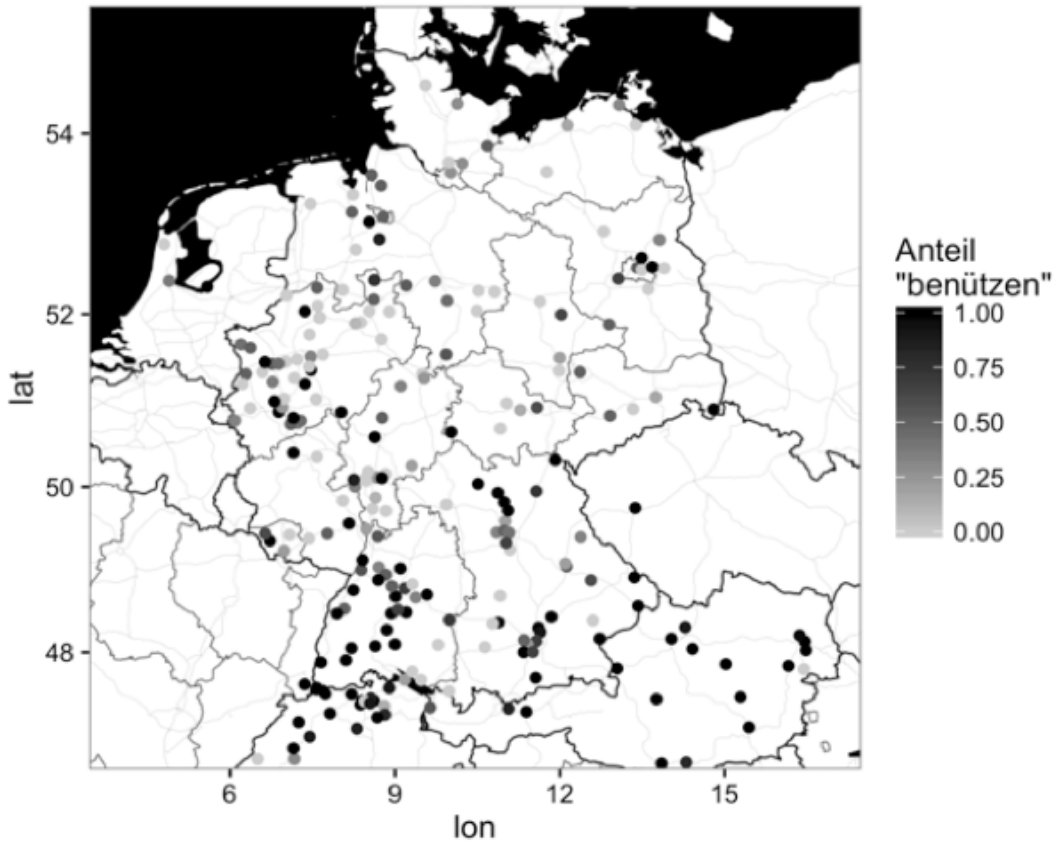
- DECOW14AX: 11 Mrd. Tokens, 624 Mio. Sätze
- Registrierung erforderlich
- umfangreiche Annotation, u.a. (grob) Textsorte und Textthemen, Anzahl und teilweise Geschlecht der AutorInnen, Geolocation
- Annotation natürlich nicht immer zuverlässig - aber z.B. im Falle der Geo-IP überraschend hohe Übereinstimmung mit dialektologischen Daten

benutzen vs. *benützen* (AdA)

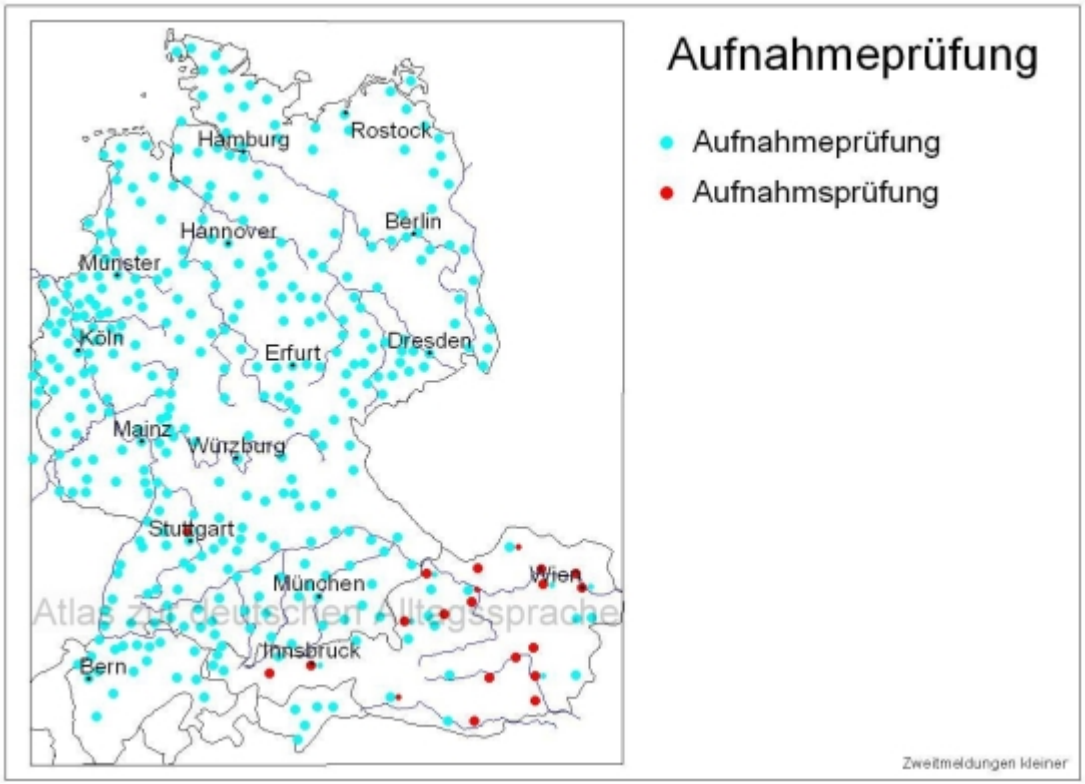


Quelle: http://www.atlas-alltagssprache.de/runde-3/f03b/

benutzen vs. *benützen* (DECOW)

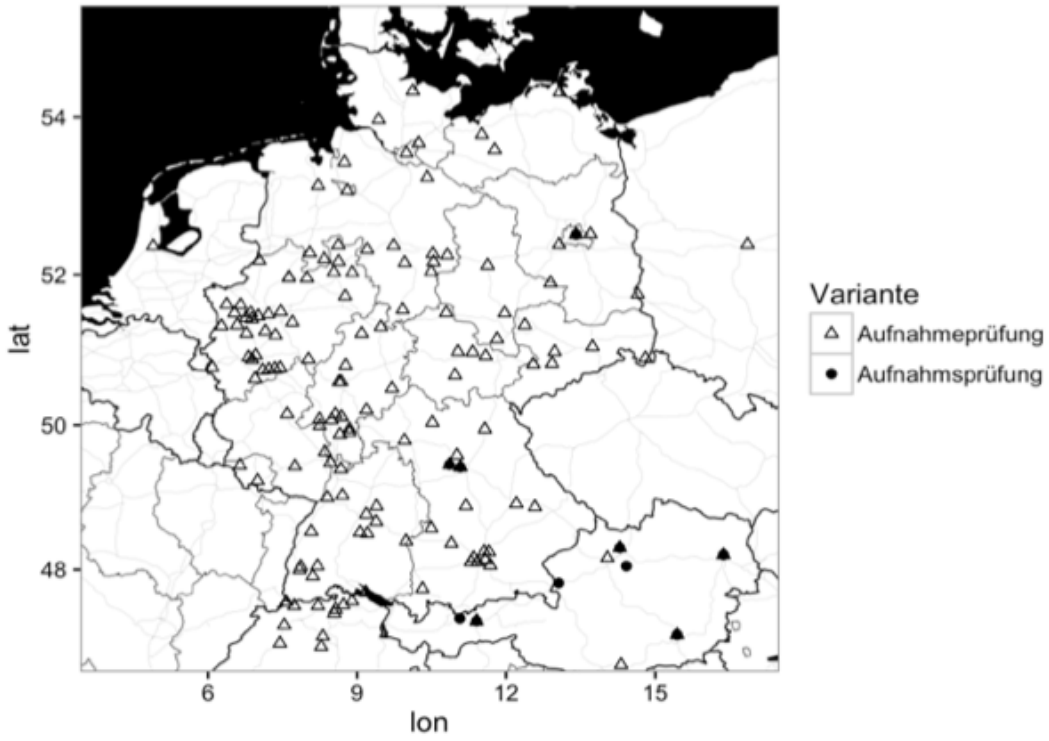


Aufnahme- vs. *Aufnahmsprüfung*



http://www.atlas-alltagssprache.de/runde-5/f13/

Aufnahme- vs. *Aufnahmsprüfung*



DECOW14AX

- nächste "Generation" derzeit in Arbeit: DECOW14AX
- u.a. mit automatischer Registererkennung auf Grundlage von >100 verschiedenen Merkmalen (Roland Schäfer, p.M.)

Beispiele für Arbeit mit COW

Kempf/Hartmann (demn.): Pseudopartizipien

- Partizipformen, zu denen kein entsprechendes Verb existiert: *bebrillt, betucht - *bebrillen, *betuchen*

(1) Wahrscheinlich sind in der Musik von Lexx, Obst und Wallace zu viele Gitarren für das **bepornosonnenbrillte** Housevolk

(2) Während ihr den Horden schwer schwankender Junggesellinnentrupps, die vor allem im Sommer wie eine der sieben Plagen über die Städte herfallen, peinlich berührt ausweicht, stößt eure Freundin bei der Sichtung eines **bebauchladeten** Junggesellinnenabschieds seit Jahren Verzückungsrufe wie "Oh wie cool!" aus.

Beispiele für Arbeit mit COW

- Vorgehen: Suche nach flektierten Formen von *be-x-t* in der Wortformenliste von DECOW
- → manuelles Aussortieren aus ca. 40.000 Belegen
- übrig blieben ca. 2000 Types
- alle Belege für diese Types wurden aus dem Gesamtkorpus extrahiert
- dabei wurden auch Komposita mit diesen Types als Bestimmungsglied einbezogen

Beispie…DW

Lemma	Freq
benachbart	124,662
beheimatet	45,422
bewaldet	16,678
beherzt	14,692
betagt	14,459
behaart	7,466
betucht	5,430
bewölkt	4,579
hochbetagt	3,205
begütert	3,091
beleibt	2,071
beringt	1,236
belaubt	1,123
bemoost	1,073
gutbetucht	1,011
bebrillt	858
behelmt	822
bemuskelt	768
unbehaart	768
behandschuht	730

Beispiele für Arbeit mit COW

- Insgesamt 273,242 Tokens (2,831 Types)
- 652 Types mit Kompositum als Basis
  - davon 514 vom Typ *sonnenbebrillt* (2164 Tokens)
  - und 138 vom Typ *besonnenbrillt* (266 Tokens)

- Bevorzugung des Musters mit aufgespaltenem Kompositum passt zum wortspielerischen Charakter des Musters

WaCky-Korpus

- "Web as Corpus kool ynitiative"
- ca. 1,7 Mrd. Tokens
- getaggt und lemmatisiert (wie COW)
- ohne Anmeldung zugänglich über NoSketchEngine

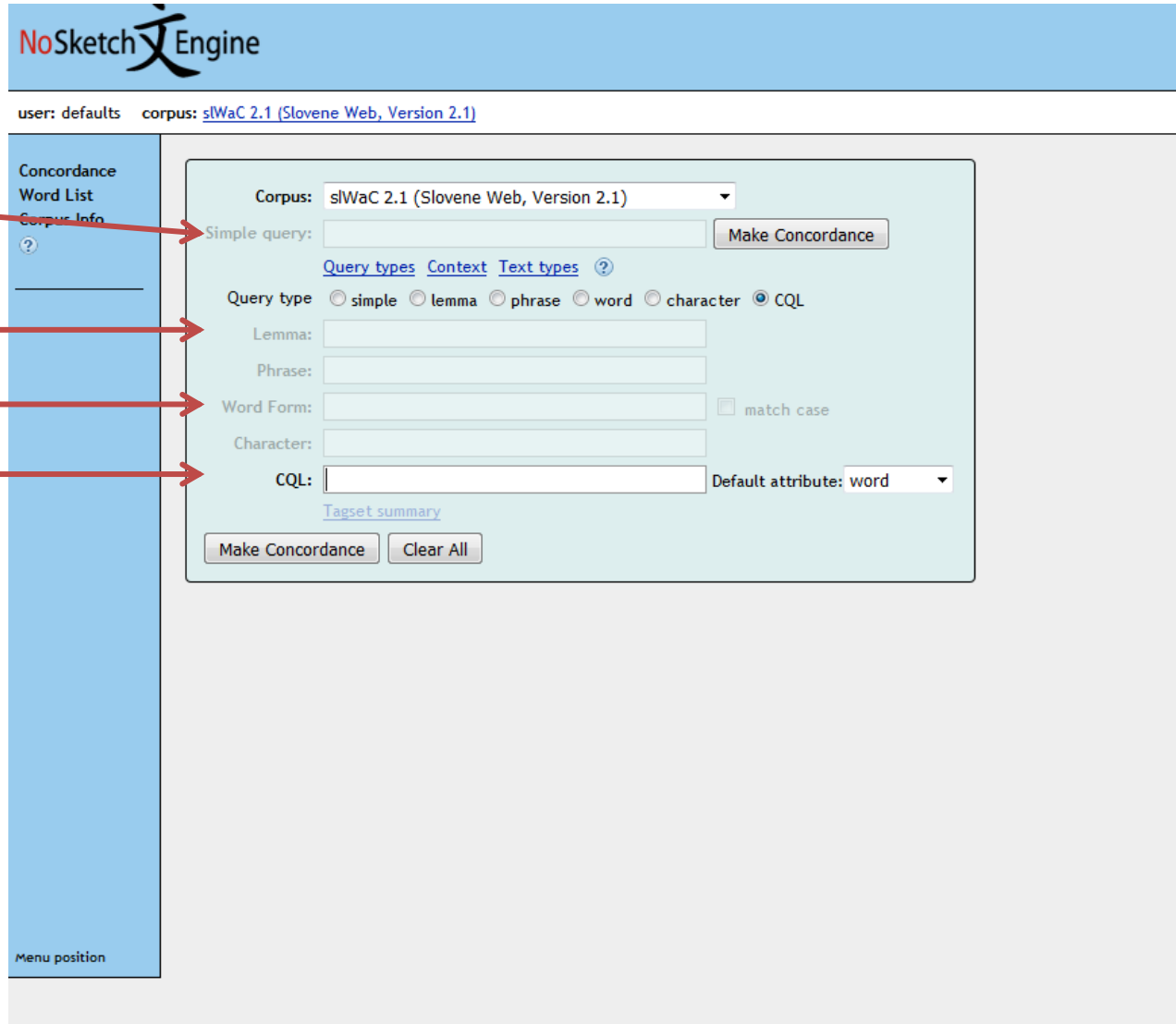
WaCky: Suchsyntax

einfache Abfrage: versucht zu erraten, was Sie suchen wollen

Lemmasuche

Wortform

Corpus Query Language



Corpus Query Language

- Suchabfragesprache des *Corpus Query Processor* des IMS Stuttgart
- Relativ intuitive Abfragesprache…
- … allerdings funktionieren nicht alle CQP-Befehle in der NoSketchEngine :-(

Corpus Query Language

- CQL-Grundprinzip: **1 Token = 1 eckige Klammer**
- in der eckigen Klammer jeweils **Attribut** und **Wert**

[word="Nacktnasenwombat"]

- word steht für die **Wortform**
- Was findet diese Anfrage also?

Nacktnasenwombat ✓

nacktnasenwombat ✗

Nacktnasenwombats ✗

Nacktnasenwombate ✗

Nackthasenkombat ✗

Corpus Query Language

- CQL-Grundprinzip: **1 Token = 1 eckige Klammer**
- in der eckigen Klammer jeweils **Attribut** und **Wert**

[lemma="Nacktnasenwombat"]

- lemma steht für die **Lemma** (duh)
- Was findet diese Anfrage also?

Nacktnasenwombat ✓

nacktnasenwombat ✓

Nacktnasenwombats ✓

Nacktnasenwombate ?

Nackthasenkombat ✗

Corpus Query Language

- Innerhalb der eckigen Klammer können auch **reguläre Ausdrücke** verwendet werden.

[lemma="(N|n)acktnasenwombat"]

Runde Klammern werden zur Gruppierung sog. **Klassen** benutzt, z.B. bei ODER-Operatoren innerhalb eines Tokens.

[lemma=".*wombat"]

Auch Wildcards und Wiederholungsoperatoren können benutzt werden.

Corpus Query Language

- In einer eckigen Klammer können auch mehrere Attribut-Wert-Kombinationen spezifiziert werden

[lemma="sieben" & tag="V.*"]

[lemma="sieben" & tag="CARD"]

tag ist die WacKy-Entsprechung für POS (Part of Speech = Wortart)

CARD steht für *Kardinalzahl.*

Corpus Query Language

- Direkt aufeinanderfolgende eckige Klammern stehen für direkt aufeinanderfolgende Tokens…

[lemma="gut"] [lemma="französisch"]

- … außer man nutzt Dummy-Token + Wiederholungsoperator:

[lemma="gut"] []{1,3} [lemma="französisch"]

nach dem System:

[]{mindestens,höchstens}

Corpus Query Language

- Randnotiz: Die Wiederholungsoperatoren sind "universell" einsetzbar…
- Einerseits lässt sich {min,max} auch innerhalb eines Tokens einsetzen:

[word="ne{1,100}i{1,100}n{1,100}"]

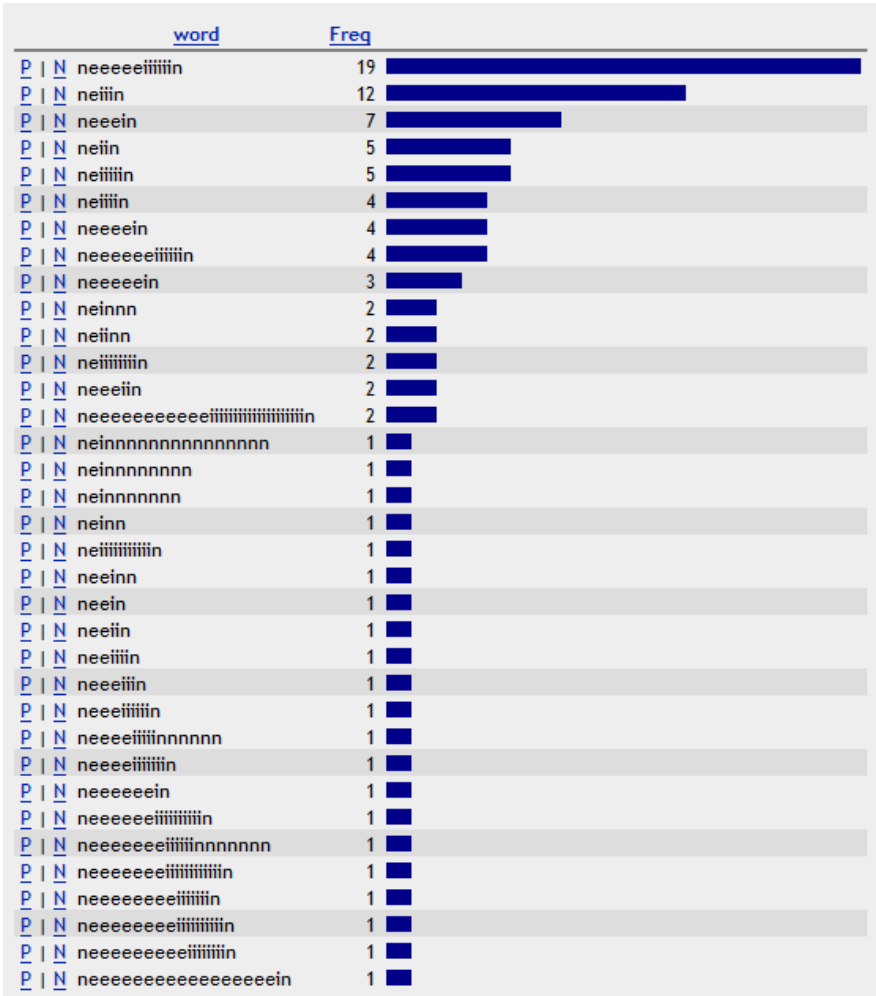
- … andererseits lassen sich auch die sonst eher innerhalb von Tokens genutzten Wiederholungsoperatoren wie * und + für ganze Sequenzen (= Tokens) einsetzen:

[lemma="gut"] []+ [lemma="französisch"]

Corpus Query Language

- Operatoren lassen sich auch mit ! **negieren**:

[word="ne{1,100}i{1,100}n{1,100}"
& word!="nein"]



word	Freq
P | N neeeeeiiiiiin	19
P | N neiiin	12
P | N neeein	7
P | N neiin	5
P | N neiiiiin	5
P | N neiiiin	4
P | N neeeein	4
P | N neeeeeeiiiiiin	4
P | N neeeeein	3
P | N neinnn	2
P | N neiinn	2
P | N neiiiiiiin	2
P | N neeeiin	2
P | N neeeeeeeeeeeiiiiiiiiiiiiiiiiiin	2
P | N neinnnnnnnnnnnnnnn	1
P | N neinnnnnnnn	1
P | N neinnnnnnn	1
P | N neinn	1
P | N neiiiiiiiiiin	1
P | N neeinn	1
P | N neein	1
P | N neeiin	1
P | N neeiiiin	1
P | N neeeiiin	1
P | N neeeiiiiiin	1
P | N neeeeiiiiinnnnnn	1
P | N neeeeiiiiiiin	1
P | N neeeeeein	1
P | N neeeeeeiiiiiiiiin	1
P | N neeeeeeeiiiiiinnnnnnn	1
P | N neeeeeeeiiiiiiiiiin	1
P | N neeeeeeeeiiiiiiin	1
P | N neeeeeeeeiiiiiiiiin	1
P | N neeeeeeeeeiiiiiiiin	1
P | N neeeeeeeeeeeeeeeein	1

Aufgabe

- Wir suchen nach Belegen für **Parallel- und Wechselflexion**.
- Worauf müssen wir achten?

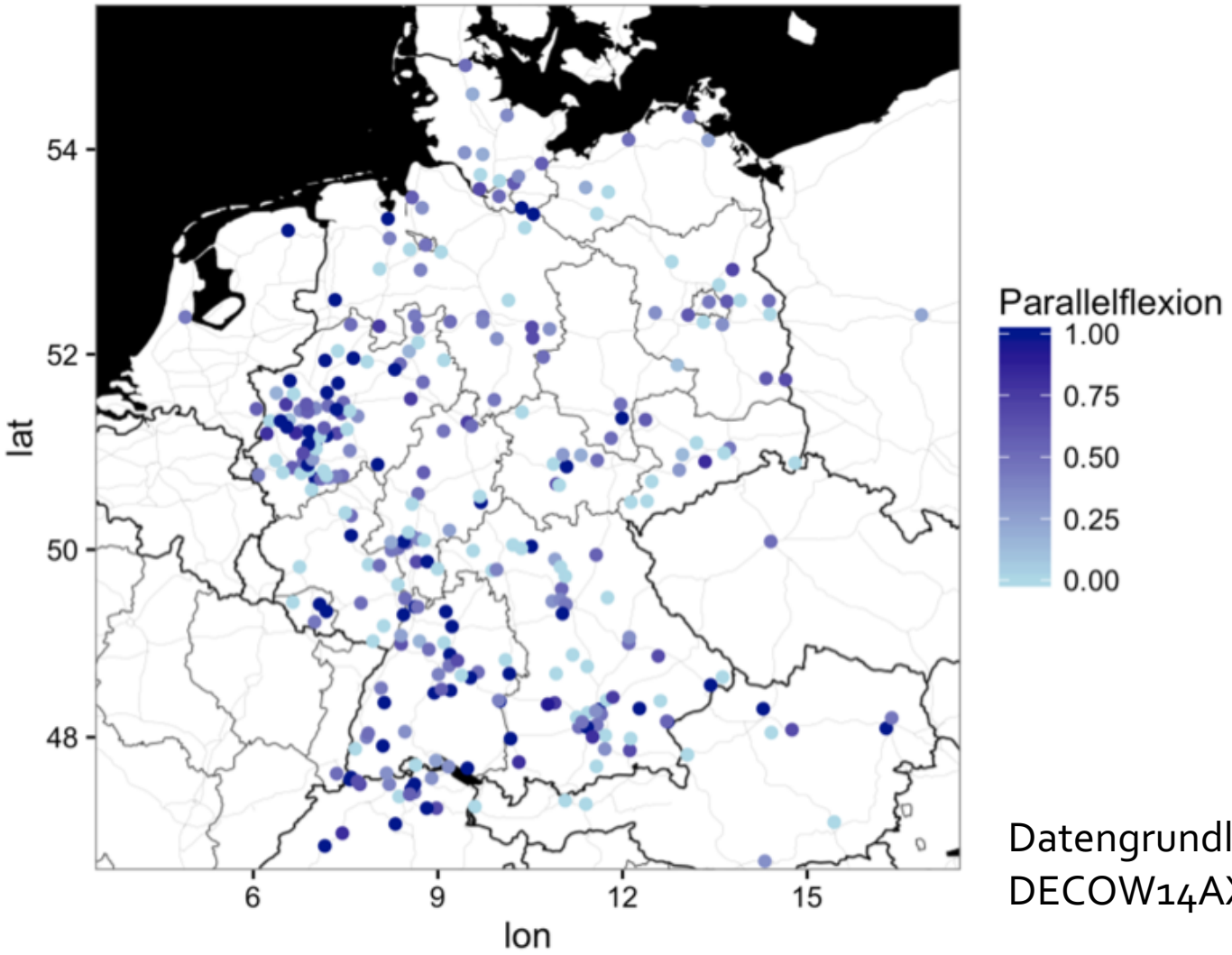
Auszug aus dem Tagset:

ADJA	attributives Adjektiv (der *schöne* Hund)
ADJD	adverbiales oder prädikatives Adjektiv (der Hund ist *schön*, er läuft *schnell*)
APPR	Präposition
NN	Substantiv, Appellativ (*Stein*)
NE	Substantiv, Eigenname (*Horst)*

Aufgabe

- **Problem:** Bei Wacky "empty result" - offenbar Bug, [tag="ADJA"] wird generell nicht gefunden
- (Grund: evtl. Timeout? Aber warum wird dann [tag="A.*"] gefunden?)
- [tag="A.*"] hingegen wird gefunden, [tag="A.*D"] ebenfalls
- --> unkonventionelle Lösung: Suche mit Wildcard und anschließendes Aussortieren

Parallel- vs. Wechselflexion: Geographische Distribution



Datengrundlage:
DECOW14AX