

Korpuslinguistik und Morphologie

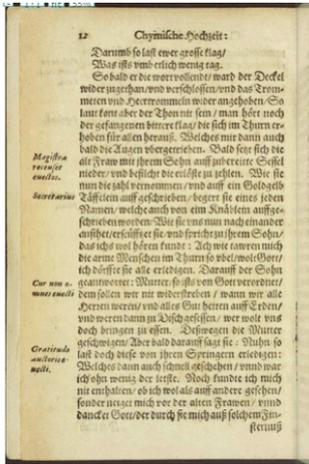
Zeitmaschinen...



Zeitmaschine



Zeitmaschinen...



Chymische Hochzeit:

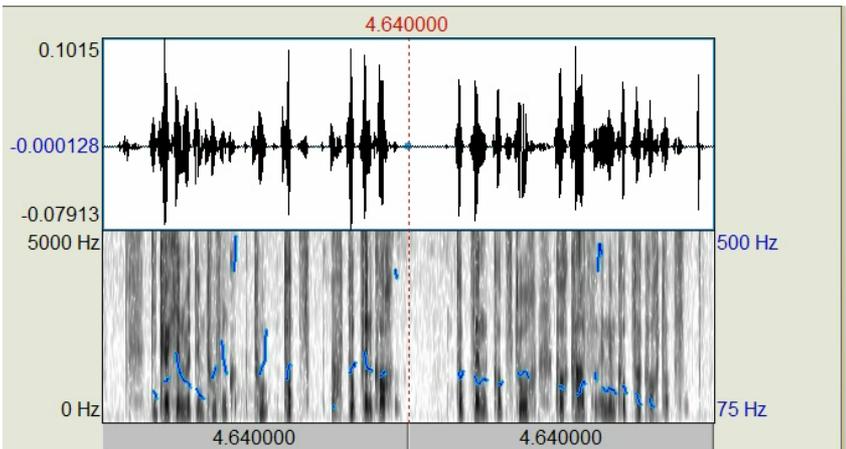
Darumb so laß ever große klag/
Was ich vmb etlich wenig tag.

So bald er die wort vollende/ ward der Deckel
wider zu gethan/ vnd verchlossen/ vnd das Trom-
men vnd Heertrommen wider angeholet/ So
laut kont aber der Thon nit kein/ man hört noch
der gefangenen bittere klag/ die sich im Thurn er-
hoben für allen freunt. Welcher mir dann auch
bald die Augen vbergetrieben. Bald fetz sich die
alt Fray mit ihrem Weib auff außereite. Gestel
nieder/ vnd bestich die erlöste ja ledten. Wie sie
mit die zahl vernommen/ vnd auff ein Ostbath
Lufflein auff geschriben/ begeret sie eines jeden
Namen/ welche auch von ein Käublen aufge-
schriben worden/ Wie sie vns nun nach einander
anlieh/ erfueffzet sie/ vnd sprich zu ihrem Sohn/
das ich wol hören kundt. Ach wie tawen mich
die arme Menschen im Thurn so viel/ wolt Got/
ich dorffte sie alle erledigen. Darauf der Sohn
geantwortet: Mutter/ so ills von Gott verordnet/
dem sollen wir nit widerthun/ wann wir alle
Heren weren/ vnd alles Gut hetten auff Erden/
vnd weren dann zu Diefel gefellen/ wer wolt vns
doch bringen zu offen. Dieweyen die Mutter
gedrungen/ Aber bald darauff laut sie: Stills so
laß doch diese von ihren Springern erledigen:
Welches dann auch schon geschriben/ vnd war
sch oben vmb der letzte. Also fante ich mich
nit erlösten/ ob ich wol als auß andere gefellen/
sonder nitze mich vee der alten Frauen/ vnd
danker Got/ der durch sie mich auß solchem Jün-
gerung

```
# Corpus: bnc (British National Corpus (XML edition))
# Name: BNC:Last
# Size: 21526 intervals/matches
# Context: 30 characters left, 30 characters right
#
# Query: BNC: [word="future"];
#
6145: s been discussing planning for ([[ future ]]) projects with ACET 's African
13688: s put at risk the security and ([[ future ]]) Of their nearest and dearest
15459: ging them to think about their ([[ future ]]) . So far we have visited over
15532: y infected . The effect in the ([[ future ]]) will be devastating . ACET 's
19223: HS his or her entire salary in ([[ future ]]) AIDS treatment costs alone .
19782: nt to AIDS prevention . In the ([[ future ]]) we hope also to be able to as
22396: nd discussed possibilities for ([[ future ]]) access by AI to Sri Lanka . A
27094: South Africa hold hope for the ([[ future ]]) , and countries abolishing th
33556: ewhere , always looking to the ([[ future ]]) . We will meet again one day
36114: nd discussed possibilities for ([[ future ]]) access by AI to Sri Lanka . A
42224: hat a reader will benefit in a ([[ future ]]) encounter with a work of art
90948: ft in the lurch , predicting a ([[ future ]]) whose likelihood the novel do
91260: elong to the town , to have no ([[ future ]]) , and they are parted when th
93201: ke of the past that shapes our ([[ future ]]) and present . & Fraser obse
95900: kward ( & hidden in the near ([[ future ]]) he was to be proved right 325
11567: ettle then for the foreseeable ([[ future ]]) . Their relationship is still
135864: mantent financial basis for the ([[ future ]]) . It is hoped that courses on
145184: en freed . He is relishing his ([[ future ]]) . He is a witty , ruthless ad
147215: y Parents wisely foreseeing my ([[ future ]]) Happiness in Country-pleasure ]
159731: f parts he/she may play in the ([[ future ]]) , or indeed may have played d
159398: 's going to be any use for the ([[ future ]]) . In the first place you are
161004: ay of commanding interest from ([[ future ]]) employers . Certainly the new
167071: 's a matter of trying to help ([[ future ]]) actors to gain a clearer focu
173221: quity could be resolved in the ([[ future ]]) & difficult as it is . Perh
180035: e to the present restricted or ([[ future ]]) enlarged republic , but it is
180580: e , chooses his words over the ([[ future ]]) of the North . He is careful
181720: likely to be forced on them by ([[ future ]]) events . The Monopoly of the
```

Deutsches Textarchiv

British National Corpus



4:08pm

was in 1333. >>
u kidding? it
all those years
ow it goes down
? that's
fire of 2012. and then we

paintings. >> this will
become a story, 200 years
in the future. well, that big

Alcohol Language Corpus

TV News Archive

1. Korpora sind Zeitmaschinen

Frag. Warumb haben die Weiber laenger Haar/ dann die Maenner?

Antwort. Diweil die Weiber mehr feuchtiger Natur sind/ dann die Maenner/ sind auch schnupffiger vnd fluessiger/ daher in jhnen mehr Saamens der Haar ist/ vnd folget endlich darauss die Laenge der Haar/ vnd darmit wird auch die Materi des Hirns ueberfluessiger von den jnnerlichen Gliedmassen/ sonderlich aber in der Weiber Vierwochenzeit/ diweil alssdann das Wesen auffsteiget/ dardurch die Feuchtigkeit der Haar gemehret wird/ wie solches Albertus meldet. Sagt auch/ dass/ wo eines Weibs (die jhre Zeit hat) Haupthaar vnder den Mist gelegt werde/ soll darauss ein giftige Schlange erwachsen.

Zum andern gibt man diese Antwort. Diweil die Weiber keine Baert haben/ vnd dass also die Natur vnd Zeug des Barts zu dem Wesen der Haupthaar komme/ auch dieselbige vollstrecke.

1. Korpora sind Zeitmaschinen

Frag. Warumb haben die Weiber laenger Haar/ dann die Maenner?

Antwort. Diweil die Weiber mehr **feuchtiger** Natur sind/ dann die Maenner/ sind auch **schnupffiger** vnd fluessiger/ daher in jhnen mehr **Saamens** der Haar ist/ vnd folget endlich darauss die Laenge der Haar/ vnd darmit wird auch die Materi des Hirns **ueberfluessiger** von den jinnerlichen Gliedmassen/ sonderlich aber in der Weiber Vierwochenzeit/ diweil alssdann das Wesen auffsteiget/ dardurch die Feuchtigkeit der Haar gemehret wird/ wie solches Albertus meldet. **Sagt auch/** dass/ wo eines Weibs (die jhre Zeit hat) Haupthaar vnder den Mist gelegt werde/ soll darauss **ein giftige Schlange** erwachsen.

Zum andern gibt man diese Antwort. Diweil die Weiber **keine Baert** haben/ vnd dass also die Natur vnd Zeug des Barts zu dem Wesen der Haupthaar komme/ auch dieselbige vollstrecke.

1. Korpora sind Zeitmaschinen

morphologischer Wandel:

feuchtiger Natur, keine Baert

graphematischer Wandel:

mehr Saamens; / (Virgel)

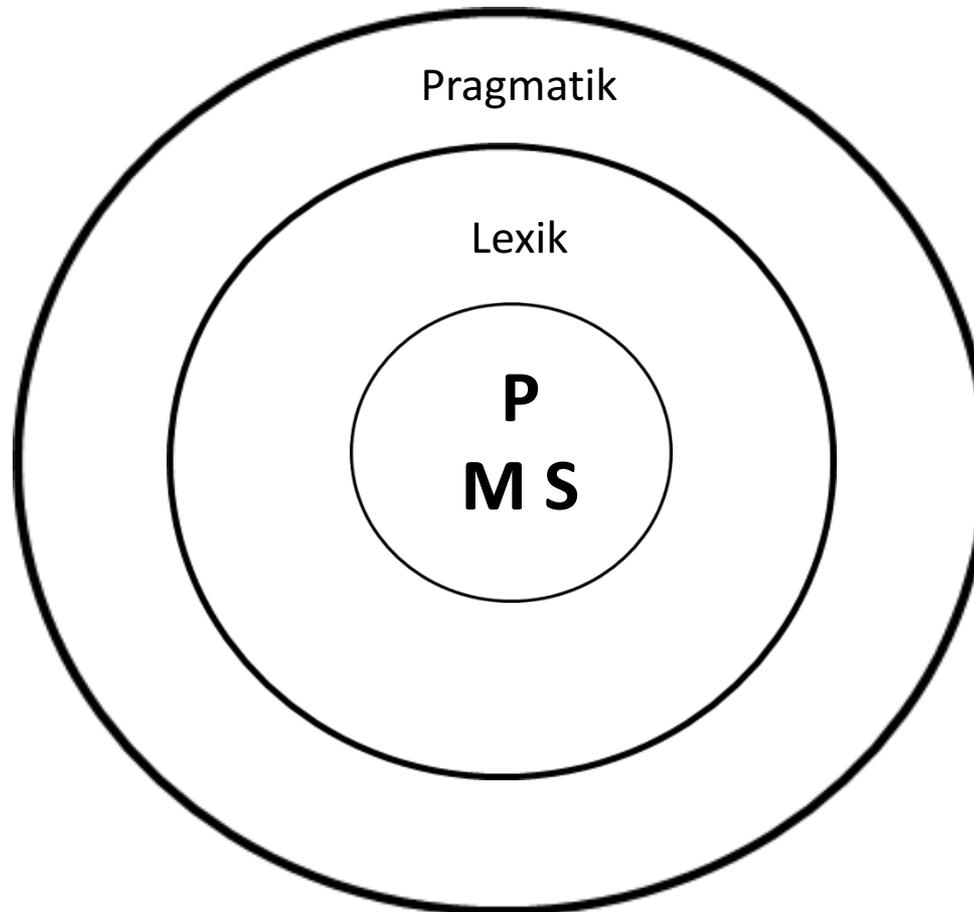
syntaktischer Wandel:

sind auch schnupffiger vnd fluessiger/
Sagt auch ...

semantischer Wandel:

vnd darmit wird auch die Materi des Hirns
ueberfluessiger (→ ‚stärker überfließend‘)

Exkurs: Die „sprachliche Zwiebel“



nach Nübling et al.
(2006: 2f.)

Korpora sind Zeitmaschinen

- Korpora ermöglichen Studien über Sprachgebrauch in der Vergangenheit
- Damit ermöglichen sie u.U. auch Voraussagen über mögliche zukünftige Entwicklungen (Beispiel: „flektierende“ Präpositionen)

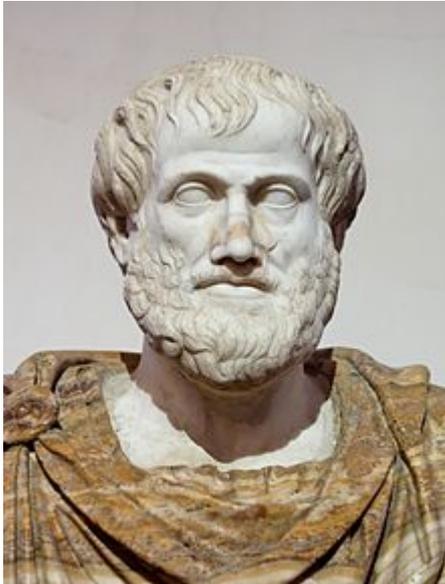
Die Zukunft beginnt heute...

- Korpuslinguistik als Disziplin „im Aufbruch“
- Verfügbarkeit von immer mehr Daten ermöglicht neue Fragestellungen
- Verfügbarkeit neuer methodischer Ansätze ermöglicht komplexere Fragestellungen
- ... gerade in der Morphologie, wo die Operationalisierung von Konzepten wie "Produktivität" notorisch schwierig ist!

Wieso, weshalb, warum?
Empirische Methoden und
Wissenschaftstheorie

(nach Maxwell & Delaney 1999)

Aristoteles



- **deduktive** (ableitende) Methode
- Ideal: Syllogismus

Alle Menschen sind sterblich.

Alle Griechen sind Menschen.

→ *Alle Griechen sind sterblich.*

Sir Francis Bacon



Novum Organon (1620):

- **Induktive** Methode
- aus Beobachtung von Naturphänomenen auf allgemeine Erkenntnis schließen
- Methode: Experimente!
- Ideal: Forscher ist objektiv und rational
- Explorativer Ansatz: Experimente nicht hypothesengeleitet

Vorannahmen

- Wissenschaft ist nie ganz frei von Vorannahmen
- Die wichtigsten davon:
 - Uniformität der Natur
 - Finite Kausalität

Uniformität der Natur



Uniformität der Natur

- Natur folgt gewissen Gesetzmäßigkeiten
- Daher sind Generalisierungen möglich.

Finite Kausalität



- Kausalkette, die zu einem Effekt führt, ist endlich.
- Damit ist der Effekt **replizierbar**.

Positivismus

- Vorläufer: David Hume: Inferenz einer kausalen Relation zwischen Unbeobachtbarem nie gerechtfertigt.
- Comte: Positivismus als ("ultimative") Religion



Auguste Comte

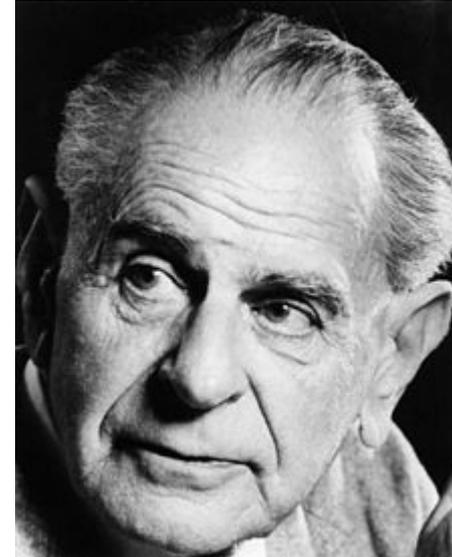


Logischer Positivismus

- Wiener Kreis (z.B. Rudolf Carnap, Herbert Feigl)
- Symbolische Logik als wichtigstes Analysewerkzeug
- Verifikationismus: Eine Proposition ist genau dann sinnvoll, wenn es eine empirische Methode gibt, um zu entscheiden, ob sie wahr oder falsch ist.
- Jedoch: Nicht alle wissenschaftlichen Fragestellungen lassen sich als universalgültige Propositionen formulieren.

Falsifikationismus

- Karl Popper: Wissenschaftlicher Fortschritt wird durch **Falsifikation** von Theorien erzielt.
- Rückkehr zur Deduktion, jedoch auf empirischer Grundlage.



Syllogismus der Bestätigung:

Wenn meine Theorie wahr ist, folgen die Daten dem von mir vorausgesagten Muster.
Die Daten folgen dem von mir vorausgesagten Muster.
~~Deshalb ist meine Theorie wahr.~~

Syllogismus der Falsifikation:

Wenn meine Theorie wahr ist, folgen die Daten dem von mir vorausgesagten Muster.
Die Daten folgen dem von mir vorausgesagten Muster *nicht*.
Deshalb ist meine Theorie falsch.

Occam's razor

- "*Entia non sunt multiplicanda praeter necessitatem*" (Johannes Clauberg, 17. Jh.)
- benannt nach Wilhelm von Ockham (13. Jh.)

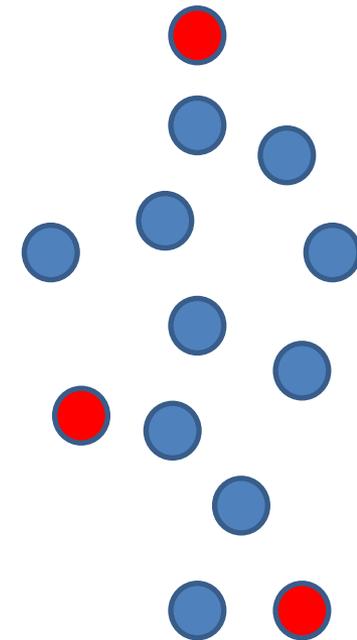
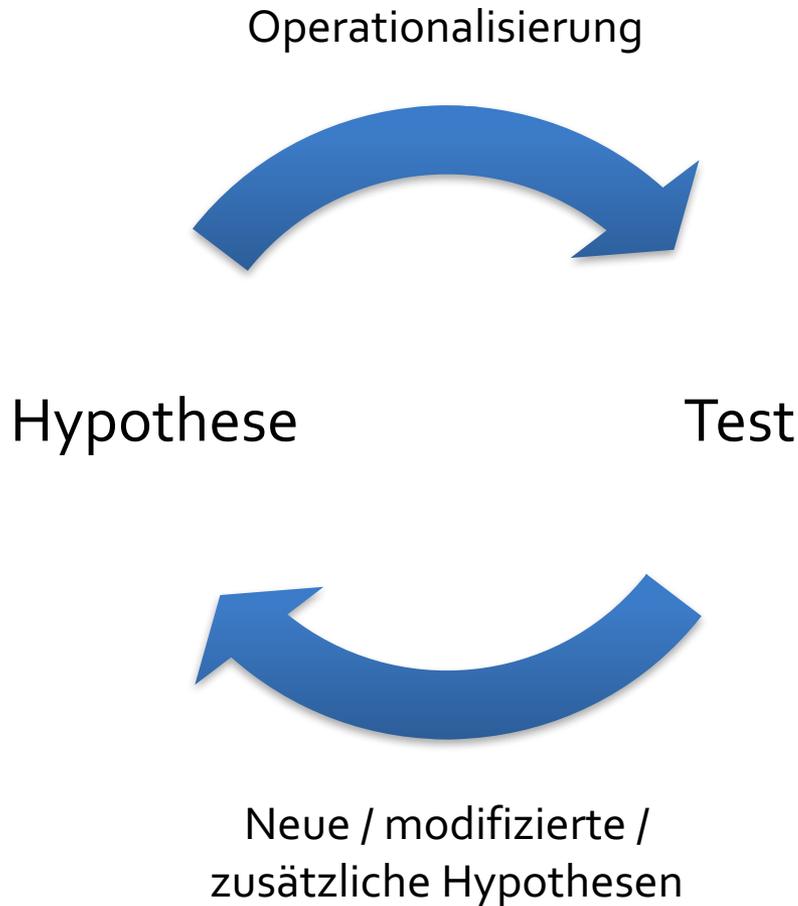
Zusammenhang zur Korpuslinguistik

- Jede Korpusuntersuchung ist im Prinzip ein Experiment.
- Gerade in der **quantativen** Korpuslinguistik geht es v.a. ums Hypothesentesten.
- Dabei wird ein **falsifikationistischer** Ansatz gewählt:
 - Ich formuliere eine Hypothese...
 - ...und überprüfe die **Nullhypothese**.

Aber...

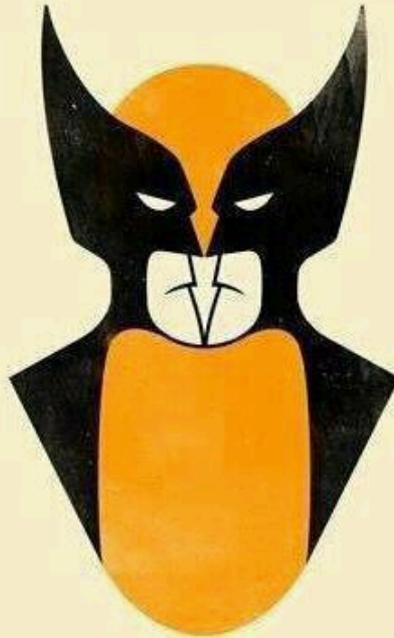
- Nicht alle (korpus-)linguistischen Methoden sind falsifikationistisch.
- Verbreitet ist auch **exploratives** Arbeiten: Muster in den Daten erkennen.
- Auch die derzeit sehr gängigen **Bayesschen Ansätze** sind nicht, oder zumindest nicht immer, falsifikationistisch orientiert.

Deduktive vs. induktive Methode



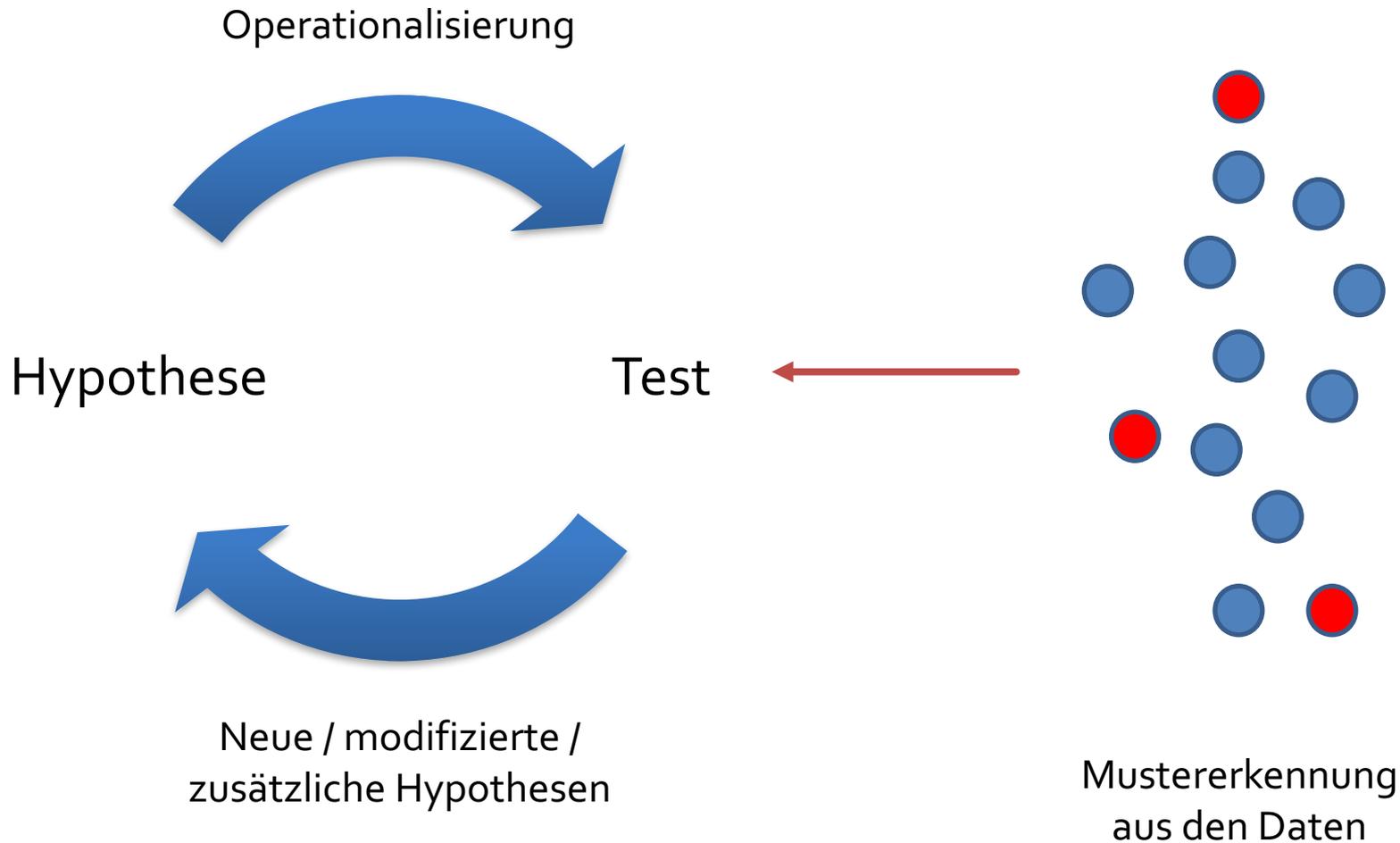
Mustererkennung
aus den Daten

WOLVERINE?.....



OR 2 BAT MEN?

Deduktive vs. induktive Methode



Warum eigentlich
Korpuslinguistik?



"Corpus
linguistics
doesn't mean
anything."
(Chomsky 2004)

Wozu?

Ich kann doch Deutsch
(Englisch, Französisch,
Mandarin.....) - warum
brauche ich dann ein
Korpus?



beizeiten, mitunter

Aus dem Korpus der Wikipedia-Diskussionen (verfügbar über COSMAS II):

- Ich werde **beizeiten** nach Quellen suchen
- Ich werde **beizeiten** die Gliederung noch ein wenig umstellen und mir das ganze nochmal mit etwas Abstand durchlesen,
- Dieser Artikel ist grausamst falsch. ich sollte mich **beizeiten** als Tropenmedizinerin mal selbst dransetzen...
- Vielleicht hat ja jemand ein vollständigeres Bild, das man **beizeiten** hier einfügen kann.

beizeiten, mitunter

Aus dem Korpus der Wikipedia-Diskussionen
(verfügbar über COSMAS II):

- Das dritte Zitat ist ja **mitunter** ein Grund für die Namensgebung
- Ich hab dazu nirgends was gefunden. Es sollte **mitunter** auch im Artikel erwähnt werden!

gleichwohl

Aus dem Korpus der Wikipedia-Diskussionen
(verfügbar über COSMAS II):

- ...gleichwohl noch nicht alle bereiche komplett dereguliert sind
- ... und andere Themen, gleichwohl sie sich im Kontext des ersten Themas befinden mögen, lieber unter den Tisch fallen lassen.

gleichsam

Internetbelege (z.T. eigene Funde, z.T. DECOW14AX)):

- [Dieses Vorgehen] ist **gleichsam** künstlerisch integer, wie konzernwirtschaftlich gerissen (<http://bit.ly/1LYhWka>)
- um ähnlich wie in Fassbinders CHINESISCHES ROULETTE (1976) die **gleichsam** dekadente wie misanthropische Upperclass abzubilden (<http://bit.ly/1a1Sw86>)
- Demnach ließe sich also leicht die Feststellung treffen, ein kongenialeres und **gleichsam** spannungsreicheres Duo als Joshua Redman und Brad Mehldau ließe sich nur schwerlich im 21. Jahrhundert auf einer Jazzbühne vereinigen. (<http://bit.ly/1PQ8m8U>)

Können wir unseren Intuitionen trauen?

- Intuition als notwendiger erster Schritt...
- ...aber Intuition ist erst der Anfang!
- ...oder *nur* der Anfang?

erst der Anfang vs. nur der Anfang

- deWaC: *erst* 1374, *nur* 1144; ganzer Satz: 278
erst vs. 138 *nur*

The screenshot shows the NoSketch Engine search interface. The search query is "Das | das, ist, erst, der, Anfang" with 272 results. The interface includes a search bar, navigation buttons (Page 1 of 2, Go, Next, Last), and a list of search results. The results are displayed in a table with columns for the source domain and the search snippet. The domains listed include baeng-2000.de, oreilly.de, exil.de, greenpeace-magazin.de, literaturline.stadt-muenster.de, konsument.at, bild.t-online.de, kika.de, zeit.de, djfl.de, e2ie2i.at, karlsruhe.de, politik-forum.at, freitag.de, literaturkritik.de, marktplatz-gp.de, maulkorbzwang.de, gazette.de, medizin-2000.de, bueso.de, ka-news.de, christoph-gaebler.de, djfl.de, and n24.at. The snippets show various contexts where the search terms appear, such as "genügen, doch dann findet sich unter einem Boot am Strand der skalpierte schwedische Ex-Justizminister", "entscheiden, sollten Sie beginnen, sich Notizen für eine Site mit mindestens 100 Seiten zu machen", "südlichen Afrika gilt. Eine Platzierung in den Top 10 der World Music Charts Europe folgte prompt", "entstehen. Das sieht der jüngst vom Kopenhagener Kabinett beschlossene Energie-21-Plan vor", "Portrait. HR: 1985. Stadt, Land, Fluß. Eine Kinderspielshow. ARD: 1982. Film, Video: ...und", "kommt. Nach zwei Monaten erstickt Europa an Überschüssen, und Brüssel ruft den Notstand aus", "Erdhalbkugel, lassen Autos durch die Luft wirbeln, pulverisieren Häuser und sogar ganze Wolkenkratzer", "stigmatisiert, dass sie nichts zum Erreichen der vom Staat erwünschten Geburtenplanziffer beitragen", "maskierter Schurke erscheint im Museum und macht Mystery Inc. für den Angriff verantwortlich", "atomar-fossile Energien mehr als verdoppelt, während sie sich für Erneuerbare Energien halbiert haben", "Neue in der nicht einfachen Klasse 3c und wird wegen seiner Sommersprossen von Florian gehänselt", "Colleges, Hotels, Krankenhäusern, Kreditkarten- und vielen anderen Unternehmen anfordern", "leistet er Widerstand, unberührt steht das Schnapsglas vor ihm, und dann trinkt er plötzlich doch", "hatte dieser sein Land vor vielen Jahren fluchtartig verlassen und blieb seither spurlos verschwunden", "unbeschriftetes Video eine wesentliche Rolle zu spielen scheint", "BETROFFEN SIND HIER ÜBER 25.000 WOHNUNGEN .. und", "45 biotechnologisch hergestellten Medikamente, wurden innerhalb der letzten drei Jahre eingeführt", "Wahlkommission festgestellt hatte, daß er in den letzten sieben Jahren rechtskräftig verurteilt worden war", "Stichwort: Lesernah. Die regelmäßige Umfrage ist eine Grundform der Leserbetätigung bei ka-news", "qualifizierten Schüler in den Lostopf. Bundesweit sind nach Uni-Angaben 400 Schulen bilingual", "TV-Film (RTL): Millionär und Stripperin - Regie: Donald Krämer - Rolle: Wolff 2000-07-27", "kämpfen gegen die Flammen. Hunderte Menschen wurden evakuiert, bargen in Notunterkünften aus".

Anwendungsbereiche...

- Zweifelsfälle
- ...was noch?

Anwendungsbereiche...

- Zweifelsfälle
- Historische Wandelprozesse
- Varietätenlinguistik und Dialektologie
- graphematischer Wandel
- Multimodalität und Interaktionsstudien
- Phonetik
-

Was ist Korpuslinguistik?

Corpus linguistics is the investigation of linguistic research questions that have been framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus.
(Stefanowitsch 2017)

Was ist Korpuslinguistik?

Corpus linguistics is the investigation of linguistic research questions that have been framed in terms of the **conditional** distribution of linguistic phenomena in a linguistic corpus.
(Stefanowitsch 2017)

- "Der Genitiv taucht in älteren Texten **häufiger** auf als in neueren Texten."
- "Ältere Sprecherinnen benutzen **seltener** Fremdwörter als jüngere."
- "Frauen benutzen **mehr** Diskurspartikeln als Männer."
- "Der Ausdruck *parkieren* wird **nur** im Schweizerdeutschen gebraucht."
- "Anglizismen werden im Deutschen **häufig** gebraucht."

Was ist Korpuslinguistik?

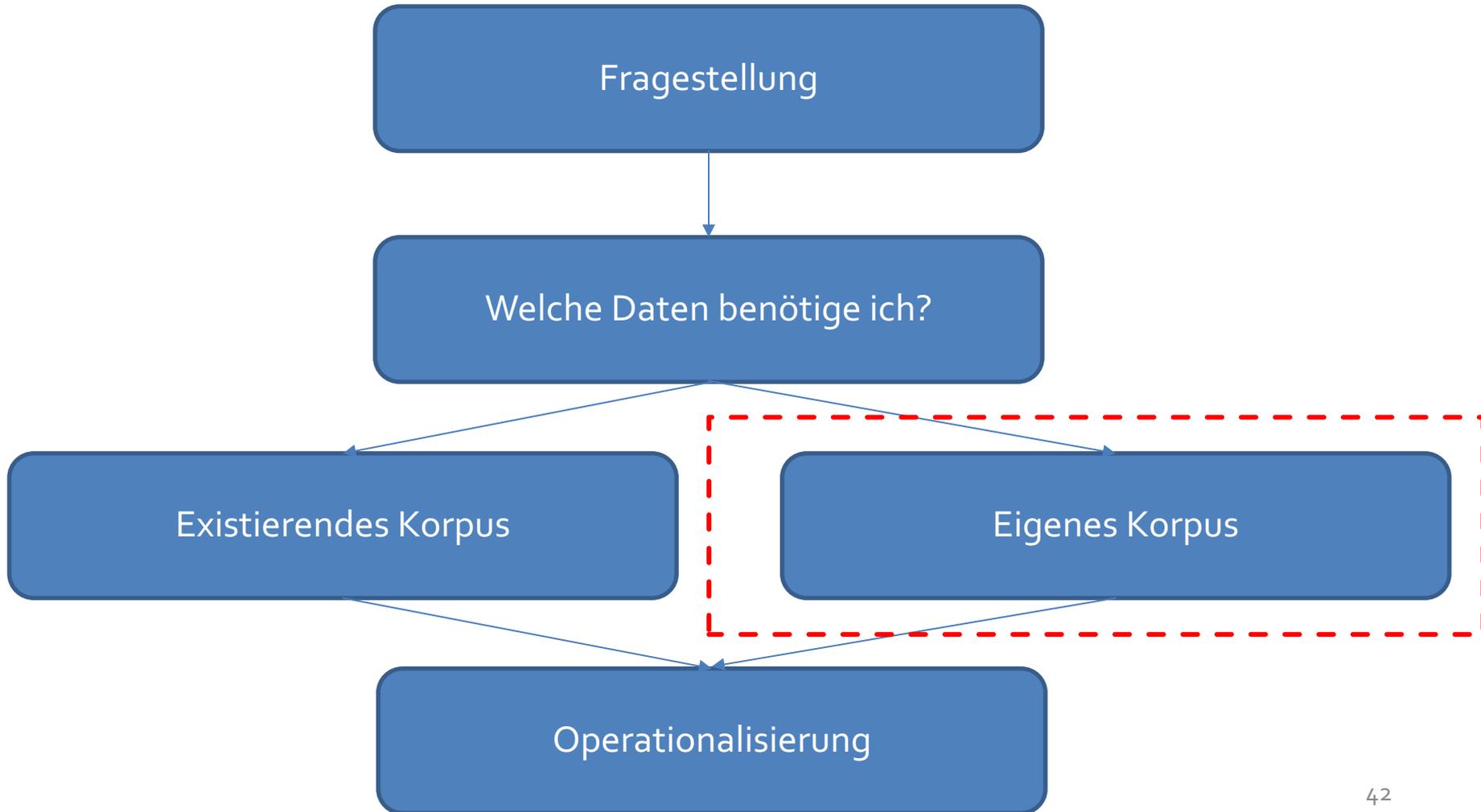
Korpusbasierte vs. korpus-illustrierte Ansätze

- "korpus-illustrierte" Ansätze sind qualitativ, benutzen aber selektiv ausgewählte Korpusbelege (z.B. viele Arbeiten von Bybee, Traugott, Trousdale)
- korpusbasierte Ansätze können rein **quantitativ** sein oder aber "**quantitativ-qualitativ**" (Lemnitzer & Zinsmeister 2015)

Was ist Korpuslinguistik

- Rein quantitative Ansätze stützen sich **ausschließlich** auf die Korpusdaten (z.B. n-Gramme, Latent-semantische Analyse...)
- Quantitativ-qualitative Ansätze stützen sich auf die Analyse und Interpretation der Daten (**Annotation**)

Arbeitsschritte in der Korpuslinguistik



Korpusdesign

Ein Wissenschaftler vom Mars bittet Sie darum, ein Korpus zusammenzustellen, das möglichst genau abbildet, wie die Leute in Hamburg sprechen.

Wie gehen Sie vor?



Korpusdesign

- Repräsentativität
- Ausgewogenheit
- Größe
- Angemessenheit für die jeweilige Forschungsfrage

bei transliterierten Texten:

- Qualität der Transliteration

Korpusdesign

Grundsätzliche Fragen:

- Was genau möchte ich untersuchen?
- Welche Art von Daten brauche ich dafür?
- Gibt es ein solches Korpus schon?
- In welcher Hinsicht muss das Korpus besonders akkurat sein?
 - z.B. bei graphematischen Untersuchungen: Graphie des Originals genau abbilden etc.

Korpusdesign

Falls ich ein eigenes Korpus zusammenstelle:

- Wie komme ich an Daten?
- Gibt es urheberrechtliche Bedenken?
- Gibt es sonstige moralische / ethische Bedenken?

Korpuserstellung

- Datensammlung und -aufbereitung
- Tokenisierung
- Lemmatisierung und POS-Tagging (z.B. TreeTagger)
- ggf. weitere Annotation

Exkurs: Wir basteln uns ein Korpus

NEWTICKER

+++ Neue Satzzeichen sollen Inflation von Frage- und Ausrufezeichen eindämmen!!! +++

Sonntagsfrage: Wen würden Sie zum Hundespräsidenten wählen?

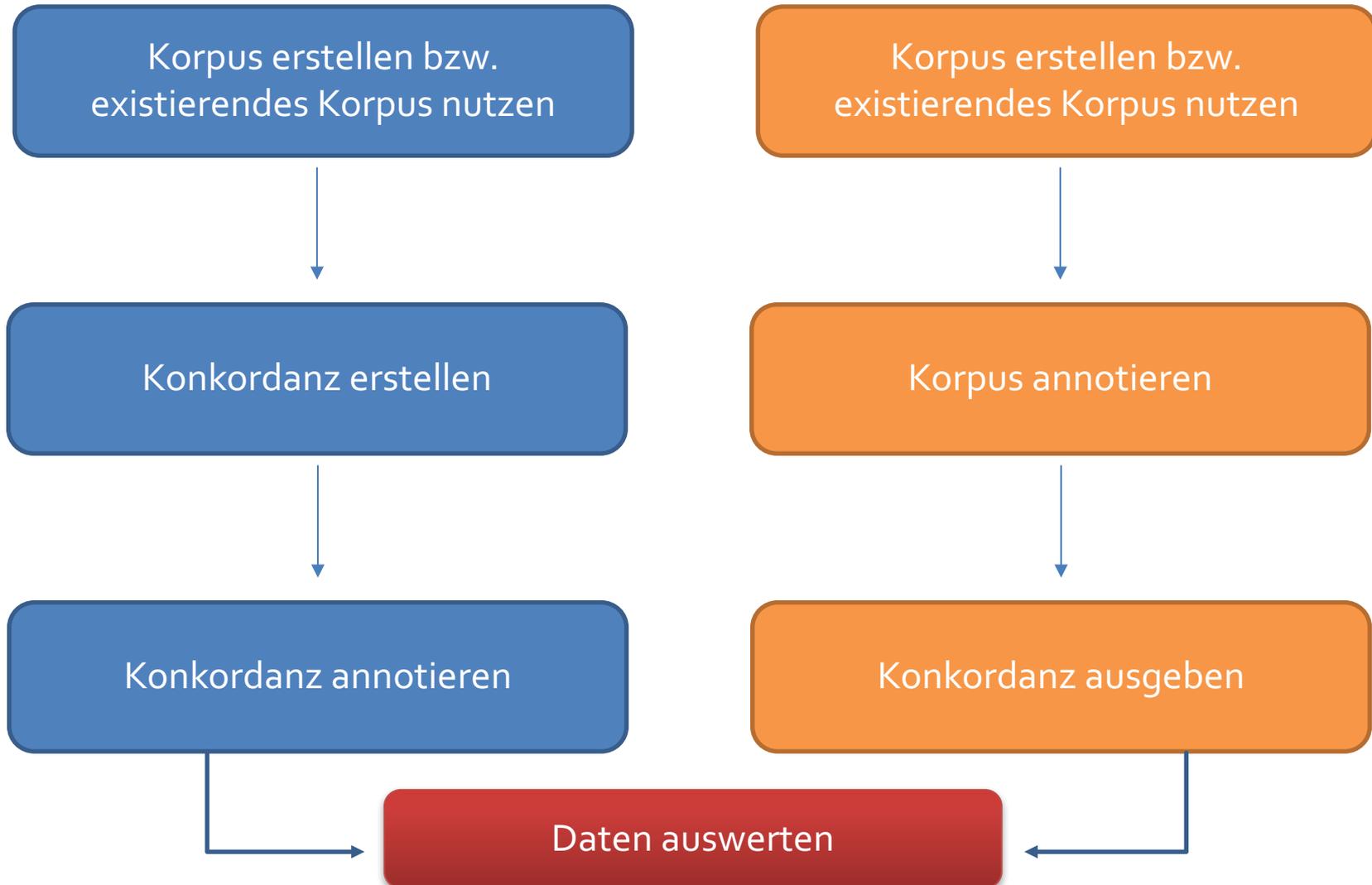


Ja, wer ist ein guter Hund? Ja, wer ist ein braver Hund? Wer ist der richtige Hund für Deutschland? Ja, wer? Ja, wer? Genau das will der *Gassillon* in dieser Woche (unterstützt vom Meinungsforschungsinstitut *Bone*

Wir basteln uns ein Postillon-Korpus

- Vorteil: Postillon-Texte unter Creative-Commons-Lizenz
- Text lässt sich recht einfach extrahieren
- Man muss nur den Seitenquelltext analysieren...

Mögliche Workflows



Annotation: Zwei Wege

Annotation
direkt im Korpus



Export

Export



Annotation
außerhalb des Korpus

A	B	C	D	E	F	G	H
URL		Left Key	Right	Lemma	POS	quasipoint	Tokens
http://www.alemannkande.de	Städte, nicht zugänglich für die Öffentlichkeit und eingeschlossen von einer hohen, efeubewachsenen Mauer, ist der Grund dafür, dass der Friedhof fast vergessen (unknown)		Augen, dasu noch ein Köstchen zwecks Beobachtung der Demos wahrheits	ADJA		2	
http://www.lauehies.de	Ein bittiger Fleum... wahlweise auch zum Schraubart verkommen... unter das... beibringen		?	ADJA		3	
http://spirituelle.de	In welchem Teil sind wir also jetzt?			beheimaten	VWPP	3	
http://www.waamte.de	sind in Deutschland mehrere führende Unternehmen aus dem Bereich der Agrartechnik.		die ihre Maschinen in aller Welt verkaufen.	beheimaten	VWPP	4	
http://www.wisens-ly.com	Das Wikin-Artikel und das... benachbarte		Erweitere Artikel im Pacific wurden als Topgezielte gewählt, weil benachbarte	ADJA		3	
http://www.waldorfschule.de	Den Namen hat das Haus von dem... benachbarten		Naturschutzgebiet "Mehorn-Teich ", einem Was benachbarte	ADJA		3	
http://www.zoo-owl.de	Ein so genannter "Zwergengrad " führt von der... benachbarten		Integriertes Kindererziehungsstätt zum Springbrunnen	beheimaten	VWPP	3	
http://www.wiki-lake.de	Beides... unbemanntes		Luftfahrzeugen gleich ist, dass der Betrieb innerhalb der Sicht unbemannt	ADJA		1	
http://deutschland-tourist.de	der Name an die hawaiianische Bezeichnung für "Wal " erinnert, steht der unbemannte		Station dort, wo es am Kältesten auf der Welt ist.	unbemannt	ADJA	4	
http://www.vogelwelt.de	, und da zu allem Überfluss der Kartensaal lediglich drei Gehminuten von meiner Wohnung		ist, wählte ich doch ganz spontan mal jener Aufführung bei	beheimaten	VWPP	1	
http://www.bangyou.com	Einerseits stellte sie ein Bindeglied zwischen dem jüngeren und dem etwas betagteren		Publikum dar, zum anderen waren die Burschen auch auf der betagt	ADJA		302	
http://www.welgen.com	sondern versteckt sich hinter jedem seiner Stopfbäume, in seinem Bart, seinen stopfelig behaartem		Kopf und seinen stets lustigen kohlrabensternen Augen, da behaart	ADJA		1	
http://www.negativ.com	Entwickler durchnah neue Gameplay-Konzepte abseits des Genre-Standards entfallen lassen. Ein... beherrschter		Druck sendet mehrere "Gates " im Head-Up-Display beherrschter	ADJA		304	
http://www.deutsche-grammatik.de	Auch in... benachbarten		Postleitzahlen ist Brüssel aktiv...	beheimaten	VWPP	2	
http://www.brunderra.de	So richtig ins Schwitzen kamen die mehr oder weniger... beibringen		Marschierer bei der Parade...	beheimaten	VWPP	5	2
http://www.glimmering.com	Auf der leicht... beibringen		Arbeitsfläche den Tag... ein bisschen richtigkeits ausrollen.	beheimaten	VWPP	697	
http://www.tv-zahn.de	mir sehr viel ", sagt der Springer, der seit fast zehn Jahren in der Deisterstadt... beheimaten		ist	beheimaten	VWPP	5	1
http://zeitsche.de	im 13. Jahrhundert errichteten die Fuggerei (Fuggerei) 1759 entstand im... benachbarten		Galtz eine Götterfigur, die später Rotschutzfarbe und von... benachbarte	ADJA		2	
http://focus.de	27 (links unten) symbolisiert ein Punkte den Mittelwert der Heißlufttemperatur... benachbarten		Beibourten	beheimaten	VWPP	201	
http://www.tabular.com	Im Kosovo und dem... benachbarten		Bonnie gilt und breitet es langsam, und auch das instabile Ma... benachbarte	ADJA		1	
http://www.eisenbahn-tarife.de	Vom Coach des Gagners erhielten unsere Jungs mächtig Lob für ihr... beheimaten		ist	beheimaten	VWPP	2	
http://www.sg-bobba.de	Vom Coach des Gagners erhielten unsere Jungs mächtig Lob für ihr... beheimaten		und jederzeit faires Spiel.	beheimaten	VWPP	225	1
http://www.marathon.com	In einer Sonderfahrt für... beheimaten		Sonderlinge, schlaube 50 Minuten...	(unknown)	ADJA	807	
http://www.leidert.de	So ist ja auch der (langweiligere) Teil der Fachhochschule Bonn-Bonn-Sieg in Rheinbach... beheimaten		beheimaten	VWPP		286	
http://www.hall-of-f33.com	000 Flüchtlinge in das nicht genutzte Sportstadion «Astrodome» nach Houston im... benachbarten		Bundesstaat Texas evakuiert werden.	beheimaten	VWPP	793	
http://www.wocamp.com	Berlin und auch nördlich davon eher am... beheimaten		Wort geht auf... beheimaten	ADJA		897	
http://www.keller-fa.de	Küme bewandene		180 m hohe Erhebung über dem Katzthal bei der Riemberg... bewandert	ADJA		2	

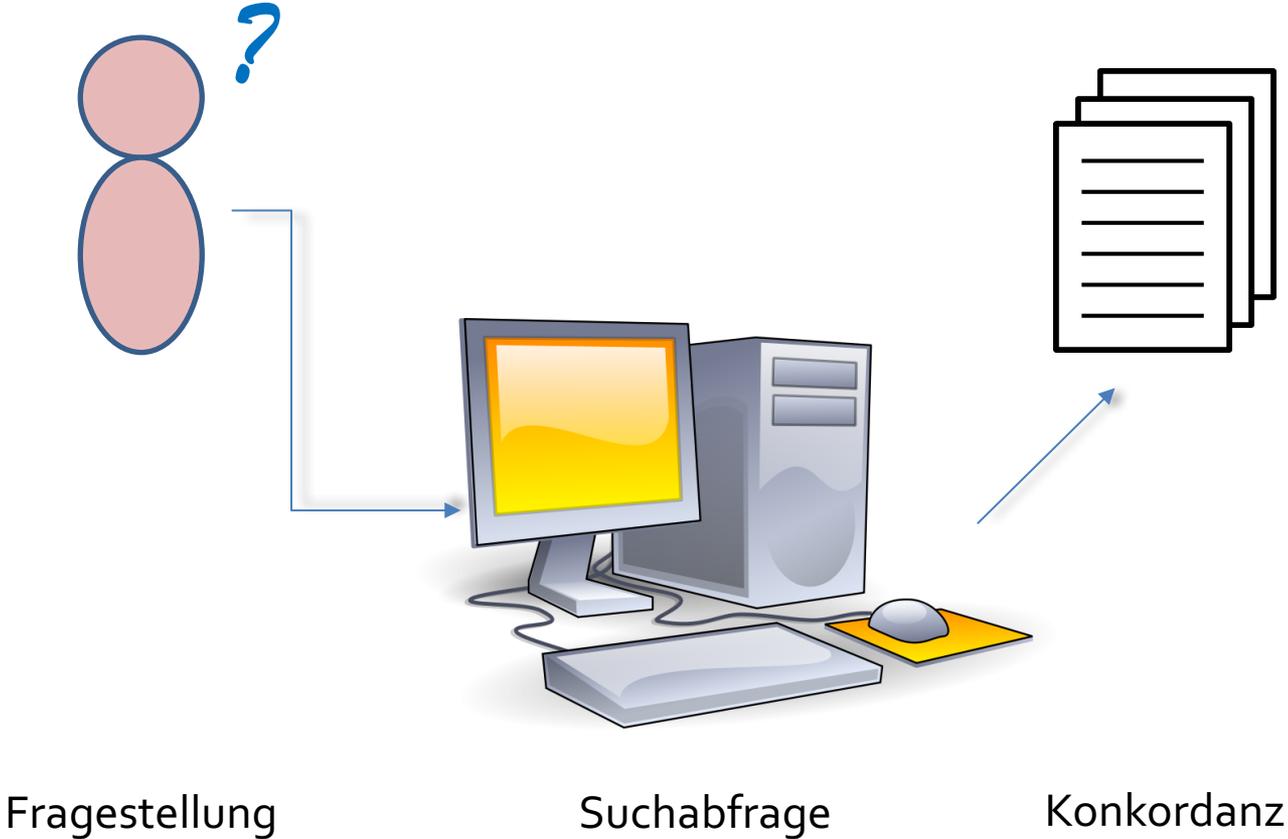
Korpus- oder Konkordanzannotation?

- **Korpusannotation:** Die Annotationen werden für das gesamte Korpus direkt in den Korpusdateien vorgenommen.
- **Konkordanzannotation:** Die Annotation erfolgt nach Ausgabe der Konkordanz in einem Tabellenkalkulationsprogramm.

Korpusauswertung

- In diesem Seminar beschränken wir uns auf die Arbeit mit **existierenden** Korpora.
- Daher werden wir uns auf den zweiten Annotationsweg beschränken (Annotation in Spreadsheet-Programmen).

Vorgehen



Korpuslinguistische Grundbegriffe

POS-Tagging & Lemmatisierung

Dieweil	ADV	dieweil
die	ART	die
Weiber	NN	Weib
mehr	ADV	mehr
feuchtiger	ADJA	feuchtiger
Natur	NN	Natur
sind/	VVFIN	sind/
dann	ADV	dann
die	ART	die
Maenner/	ADJA	Maenner/
sind	VAFIN	sein
auch	ADV	auch
schnupffiger	ADJA	schnupffiger
vnd	NN	vnd
fluessiger/	VVFIN	fluessiger/
daher	PAV	daher
in	APPR	in
jhnen	ADJA	jhnen
mehr	PIAT	mehr
Saamens	NN	Saamens
der	ART	die
Haar	NN	Haar
ist/	ADJA	ist/
nd	NN	nd

- oft automatisch, z.B. mit TreeTagger
- Vorteil: extrem schnell und effizient
- Nachteil: ungenau
- für historische Daten z.T. eigene Tagger verfügbar
- z.B. eigenes TreeTagger Parameter File für Mhd.

Tagsets

- unterschiedliche Tagsets für POS
- am verbreitetsten jedoch: Stuttgart-Tübingen Tagset (STTS)
- Übersicht: <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>

Types und Tokens

Wort	Freq	
die	9	
der	5	
vnd	5	
Weiber	4	
auch	3	
Antwort.	2	
dann	2	
darauss	2	
den	2	
des	2	
Dieweil	2	
Haar	2	
Haar/	2	
Haupthaar		2
in	2	

Types vs. Tokens



Types vs. Tokens



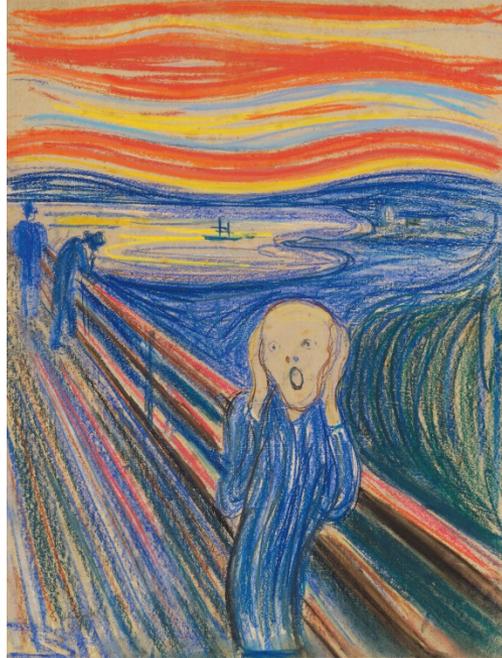
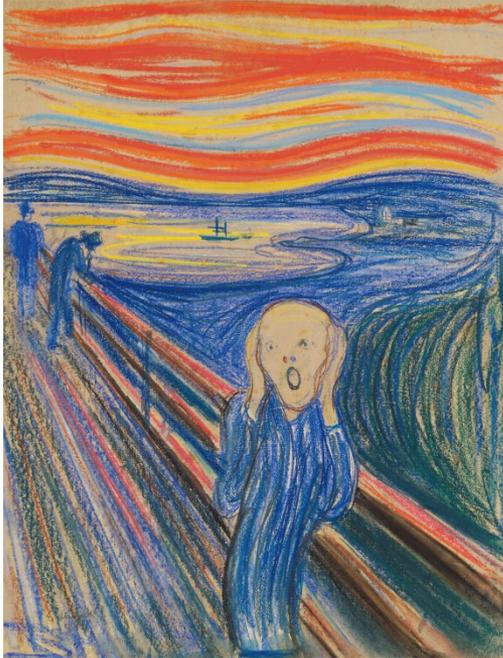
Types vs. Tokens



Types vs. Tokens



Types vs. Tokens



Types vs. Tokens



Wie viele Types...?

Es kommt drauf an...

Types und Tokens

Wenn Fliegen neben Fliegen fliegen, fliegen
Fliegen neben Fliegen.

Lemma	Tokens
Fliege	4
fliegen	2
wenn	1
neben	2

Methoden der Korpusanalyse

Korpusauswertung

qualitative Analyse:

- Beobachtungen auf Grundlage einzelner Belege
- kann sich auf alle Aspekte von der Semantik über die Morphologie bis hin zur Syntax beziehen
- gerade für semantische und pragmatische Analysen geeignet

Korpusauswertung

quantitative Analyse:

- Einbezug zahlreicher Belege statt Einzelbeobachtungen
- Quantifizierung z.B. durch
 - Zählen von Wörtern, Wortarten, grammatischen Mustern usw.
 - statistische Methoden (z.B. Kollokationsmaße)

Korpusauswertung

Frag. Warumb haben die Weiber laenger Haar/ dann die Maenner?

Antwort. Dieweil die Weiber mehr feuchtiger Natur sind/ dann die Maenner/ sind auch schnupffiger vnd fluessiger/ daher in jhnen mehr Saamens der Haar ist/ vnd folget endlich darauss die Laenge der Haar/ vnd darmit wird auch die Materi des Hirns ueberfluessiger von den jnnerlichen Gliedmassen/ sonderlich aber in der Weiber Vierwochenzeit/ dieweil alssdann das Wesen auffsteiget/ dardurch die Feuchtigkeit der Haar gemehret wird/ wie solches Albertus meldet. Sagt auch/ dass/ wo eines Weibs (die jhre Zeit hat) Haupthaar vnder den Mist gelegt werde/ soll darauss ein giftige Schlange erwachsen.

Zum andern gibt man diese Antwort. Dieweil die Weiber keine Baert haben/ vnd dass also die Natur vnd Zeug des Barts zu dem Wesen der Haupthaar komme/ auch dieselbige vollstrecke.

Wie können wir diesen Text untersuchen?

Qualitativ vs. quantitativ

Bitte überlegen Sie: Welche Vor- und Nachteile haben **qualitative** bzw. **quantitative** Ansätze?

Wofür würden Sie welchen Ansatz wählen?

1. Wandel der Genitivstellung (*des Vaters Haus* > *das Haus des Vaters*)
2. Rassismus in Leserbriefen
3. Semantischer Wandel von *geil*

~~und~~ Qualitativ vs. quantitativ

- Die meisten korpuslinguistischen Ansätze sind zugleich qualitativ und quantitativ
- Operationalisierung einzelner (z.B. semantischer) Variablen erfordert in der Regel eine (qualitative) **Interpretation** der einzelnen Belege
- Beispiel: Belebtheitsannotation

~~und~~ Qualitativ vs. quantitativ

- Sogar syntaktische Annotation erfordert oft Interpretation der Daten

Gefragt worden,

Ob sie nicht einmal Wolfgang Söhnen des Oberschulteisen
dochterlein angrieffen

Sie nichts gestehenn wollen

vnd da sie weiter gefragt worden

ob sie nicht domals geredt man könne dem kindt nicht wieder
noch wol helffenes

sey den weil der 9 te noch nit furuber
diesesauchnicht gestehen wollen

(SiGS-Korpus, Gaugrehweiler 1610)

Von der Konkordanz zur Analyse

- Operationalisierung von Hypothesen
- --> klare und nachvollziehbare Annotationskriterien!

Produktivität

Nr.	Ansatz	Definition
I		Produktivität einer Wortbildungsregel ist eine Funktion der Menge der zu einem bestimmten Zeitpunkt nach dieser Wortbildungsregel realisierten Bildungen (direkte Proportionalität)
II		Produktivität einer Wortbildungsregel ist die Funktion der Menge ihrer möglichen Basen (direkte Proportionalität)
III		Produktivität als Verhältnis von usuellen zu möglichen Bildungen
IV		Produktivität als Möglichkeit von Neubildungen
V		Produktivität als Wahrscheinlichkeit von Neubildungen
VI		Produktivität als Anzahl der Neubildungen in einem bestimmten Zeitraum

(Rainer 1987)

Produktivität

- Ist die Definition **synchron, diachron** oder **unbestimmt**?
- Ist die Definition **quantitativ** oder **qualitativ** orientiert?
- Welche Definition kommt dem **Scherer (2006)** zugrundeliegenden Produktivitätsbegriff am nächsten?

Produktivität

Nr.	Ansatz	Definition
I		Produktivität einer Wortbildungsregel ist eine Funktion der Menge der zu einem bestimmten Zeitpunkt nach dieser Wortbildungsregel realisierten Bildungen (direkte Proportionalität)
II		Produktivität einer Wortbildungsregel ist die Funktion der Menge ihrer möglichen Basen (direkte Proportionalität)
III		Produktivität als Verhältnis von usuellen zu möglichen Bildungen
IV		Produktivität als Möglichkeit von Neubildungen
V		Produktivität als Wahrscheinlichkeit von Neubildungen
VI		Produktivität als Anzahl der Neubildungen in einem bestimmten Zeitraum

(Rainer 1987)

Produktivität

Nr.	Ansatz	Definition
I	synchron	Produktivität einer Wortbildungsregel ist eine Funktion der Menge der zu einem bestimmten Zeitpunkt nach dieser Wortbildungsregel realisierten Bildungen (direkte Proportionalität)
II	synchron	Produktivität einer Wortbildungsregel ist die Funktion der Menge ihrer möglichen Basen (direkte Proportionalität)
III	synchron	Produktivität als Verhältnis von usuellen zu möglichen Bildungen
IV	unbestimmt	Produktivität als Möglichkeit von Neubildungen
V	unbestimmt	Produktivität als Wahrscheinlichkeit von Neubildungen
VI	diachron	Produktivität als Anzahl der Neubildungen in einem bestimmten Zeitraum

(Rainer 1987)

Usuelle vs. mögliche Wörter

- „Mögliche“ Wörter: z.B. *Durstigkeit, Fleißigkeit*
- Usuelles Wort: „Wort, das zu einem bestimmten Zeitpunkt zum Wortschatz eines bestimmten Sprechers gehört“ (Rainer 1987)
- (Wie) Lassen sich diese Begriffe **empirisch** operationalisieren?

Produktivität

Nr.	Ansatz	Definition
I	synchron	Produktivität einer Wortbildungsregel ist eine Funktion der Menge der zu einem bestimmten Zeitpunkt nach dieser Wortbildungsregel realisierten Bildungen (direkte Proportionalität)
II	synchron	Produktivität einer Wortbildungsregel ist die Funktion der Menge ihrer möglichen Basen (direkte Proportionalität)
III	synchron	Produktivität als Verhältnis von usuellen zu möglichen Bildungen
IV	unbestimmt	Produktivität als Möglichkeit von Neubildungen
V	unbestimmt	Produktivität als Wahrscheinlichkeit von Neubildungen
VI	diachron	Produktivität als Anzahl der Neubildungen in einem bestimmten Zeitraum

(Rainer 1987)

Produktivität

- synchron: Fähigkeit eines Musters, Neubildungen hervorzubringen
- diachron: Fähigkeit eines Musters zum Zeitpunkt t_i , Neubildungen hervorzubringen, im Vergleich zum Zeitpunkt t_{i-1} , t_{i-2} , \dots , t_{i-n} .

Produktivität

- binär: ein Wortbildungsmuster ist entweder produktiv oder unproduktiv
- graduell: ein Wortbildungsmuster kann in verschiedenem Maße produktiv sein

Produktivität

- „Doppelexistenz“: Jedes Wortbildungsprodukt ist **zugleich** ein eigenes Wort **und** Instantiation eines Wortbildungsmusters
- Wortbildungsprodukte sind in unterschiedlichem Maße **transparent**
- vgl. *Jungtier* ‚junges Tier‘ vs. *Junggeselle* *‚junger Geselle‘; *Landung* vs. *Zeitung*

Produktivität

- Voraussetzung dafür, dass ein Wortbildungsmuster produktiv ist, ist, dass es **erkennbar** ist
- z.T. jedoch phonologischer Schwund (z.B. *vrouw-ida* > *Freude*)
- z.T. noch erkennbar, aber sehr infrequent und (wohl deshalb) unproduktiv: *Kehricht*, *Dickicht*.

„Grammatikalisierung“?

- Grammatikalisierung: Lexikalisch(er)e Einheiten werden zu grammatisch(er)en Einheiten.
- Grammatikalisierung i.e.S.: Entstehung von Einheiten, die grammatische **Kategorien** kodieren (Kriterium der Obligatorizität!)
- Aber: Auch WB-Affixe können „im weiteren Sinne als grammatische Morpheme gelten“ (Munske 2002: 28)

Was ist eigentlich Grammatikalisierung?



**Elizabeth
Closs Traugott**



Paul J. Hopper

- the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical function

Hopper & Thompson (2003): Grammaticalization. 2nd ed. Cambridge.

Was ist eigentlich Grammatikalisierung?

Ich **habe** ein Smartphone.



Ich **habe** Blumenkohl gegessen.

Was ist eigentlich Grammatikalisierung?



Elizabeth
Closs Traugott



Paul J. Hopper

- the change whereby lexical items and constructions **come** in certain linguistic contexts **to serve grammatical functions** and, once grammaticalized, continue to develop new grammatical function

Hopper & Thompson (2003): Grammaticalization. 2nd ed. Cambridge.

Was ist eigentlich Grammatikalisierung?



**Elizabeth
Closs Traugott**



Paul J. Hopper

- the change whereby lexical items and constructions **come** in certain linguistic contexts **to serve grammatical functions** and, once grammaticalized, continue to develop new grammatical function

Hopper & Thompson (2003): Grammaticalization. 2nd ed. Cambridge.

Was ist eigentlich Grammatikalisierung?



Elizabeth
Closs Traugott

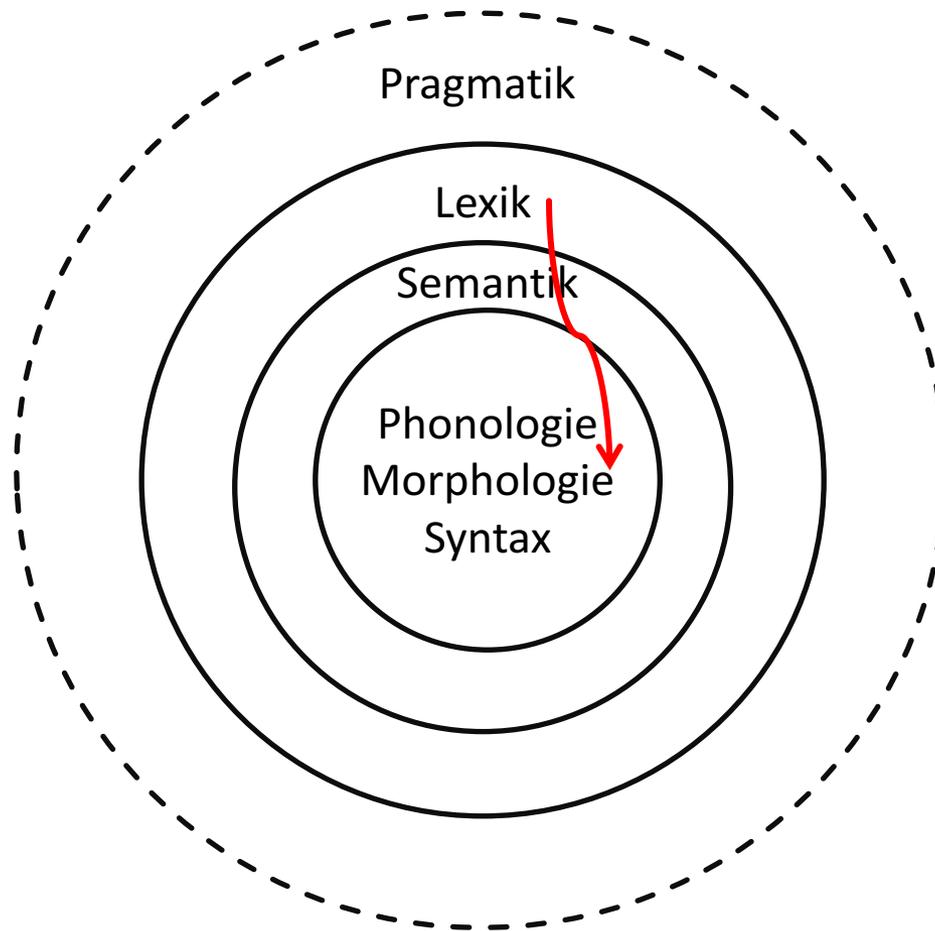


Paul J. Hopper

- the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical function

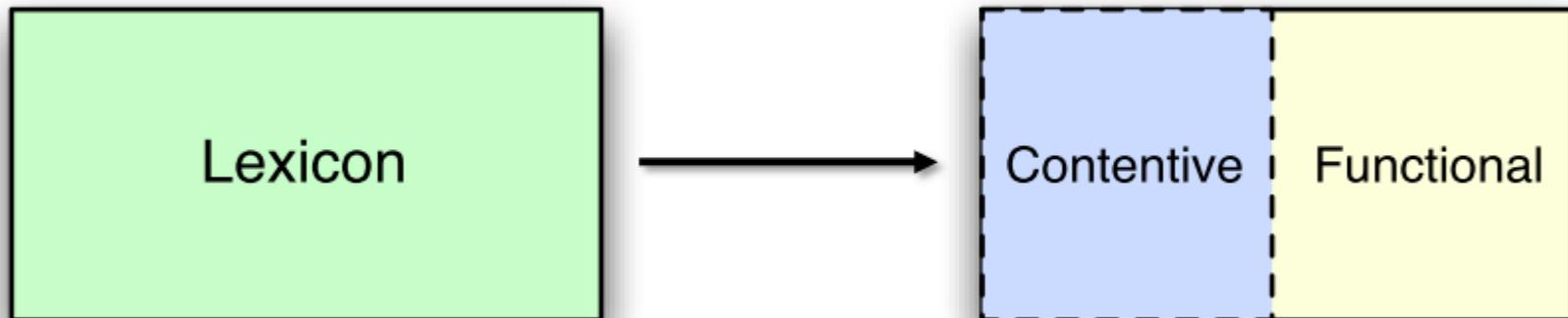
Hopper & Thompson (2003): Grammaticalization. 2nd ed. Cambridge.

Was ist eigentlich Grammatikalisierung?



Was ist eigentlich Grammatikalisierung?

- Inhalts- vs. Funktionswörter / -einheiten
- z.B.: Hilfsverb *werden* (*Wir werden heute über Grammatikalisierung sprechen*) vs. Vollverb *werden* (*Ich werde Lehrerin*)



Was ist eigentlich Grammatikalisierung?

- Vor vielen Jahren kaufte Petra ein Fahrrad. Jetzt hat sie ihr Fahrrad wieder verkauft.

Was ist eigentlich Grammatikalisierung?

- Vor vielen Jahren kaufte Petra ein Fahrrad. Jetzt hat sie ihr Fahrrad wieder verkauft.

Was ist eigentlich Grammatikalisierung?

Desemantisierung

Verlust an semantischem Gehalt



Extension

Verwendung in neuen Kontexten (z.B. morphologischer Umlaut)



Dekategorisierung

Verlust des Status als unabhängiges Wort (Verlust der morphosyntaktischen Eigenschaften des Ursprungswortes)



Erosion

Reduktion phonologischer Substanz

Grammatikalisierungszyklus

- Hypothese: unidirektionaler / zyklischer Wandel
- Diskurs > Syntax > Morphologie > Morphonologie > Schwund (nach Givón 1979 und Lehmann 1995)

Konstruktionalisierung



**Elizabeth
Closs Traugott**



**Graeme
Trousdale**

Konstruktionalisierung

Konstruktionsgrammatik:

- Sprache als Netzwerk von **Konstruktionen** (Form-Bedeutungs-Paaren)

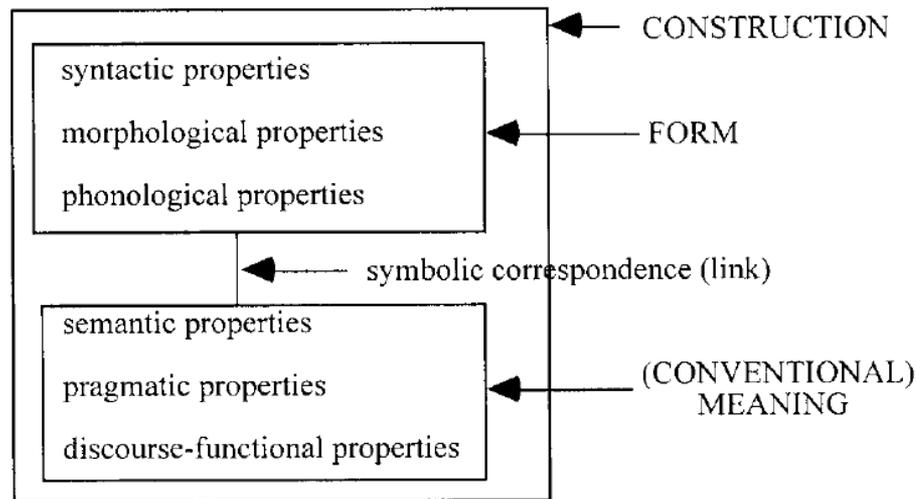


Figure 18.2. The symbolic structure of a construction

(aus Croft 2007)

Konstruktionalisierung

- Lexikon-Syntax-Kontinuum (aus Croft 2001):



Construction type	traditional name	example
complex and (mostly) schematic	syntax	Passivkonstruktion
complex and (mostly) specific	idiom	jn. auf den Arm nehmen, je X-er desto Y-er
complex but bound	morphology	Land-ung; Fisch-e
atomic and schematic	syntactic category	[DEM], [ADJ]
atomic and specific	word / lexicon	<i>Hund, Katze</i>

Konstruktionalisierung

Traugott & Trousdale (2013)

- Konstruktionalisierung: Entstehung eines neuen Knotens im Konstruktikon
- Beispiel: *Was macht XY?* - Konstruktion (vgl. auch Bybee 2010)
- Was machst du da? fuhr Kronbecher sie an. (Jentzsch, Kerstin: Ankunft der Pandora | DWDS)
- Was macht der Blüm jetzt schon wieder? (Der Spiegel, 05.10.1987 | DWDS)

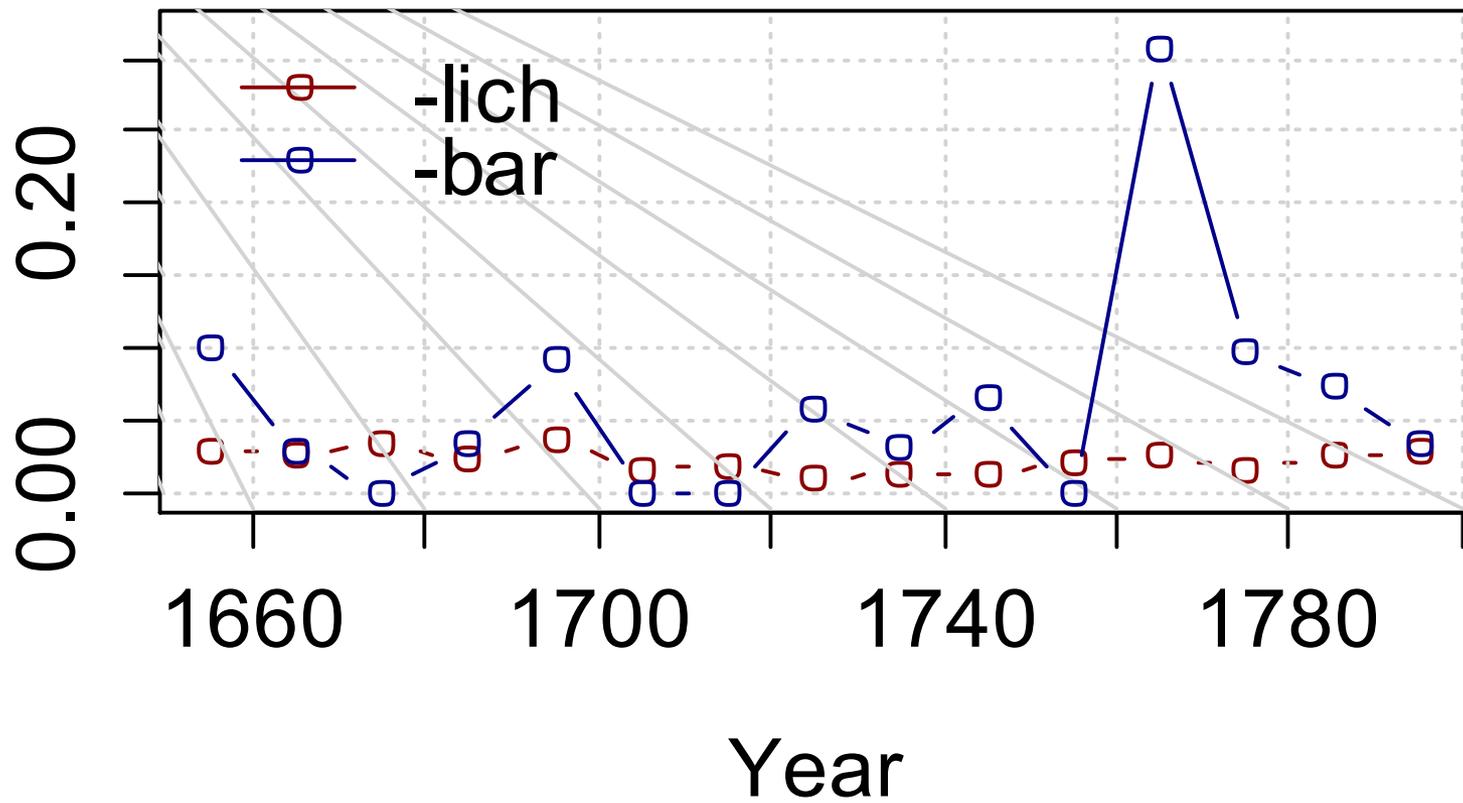
Konstruktionalisierung

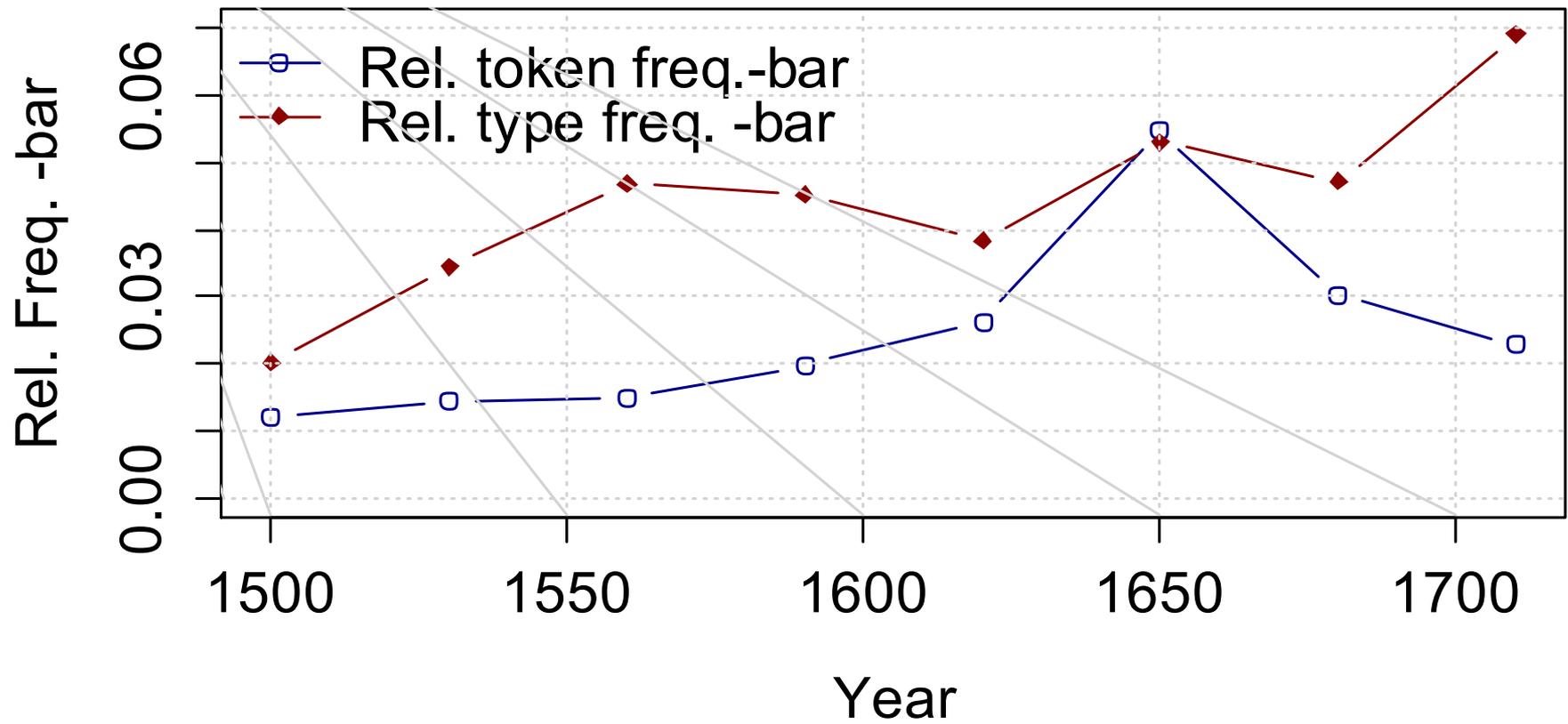
- Konstruktionalisierung in der Wortbildung:
 - Entstehung neuer WB-Konstruktionen
 - Lexikalisierung bestehender WB-Produkte
- Morphologische vs. lexikalische Konstruktionalisierung
- In beiden Fällen entstehen neue Form-Bedeutungs-Paare

Produktivität messen

- Baayen (2009: 902):
 - **realized productivity** $P = V(C, N)$
 - **expanding productivity** $P = V(1, C, N) / V(1, N)$
 - **potential productivity** $P = V(1, C, N) / N(C)$
- $V(C, N)$: Zahl der Types, die zur morphologischen Kategorie C gehören, in einem Korpus mit N Tokens
- $V(1, C, N)$: Hapax Legomena, die zur morphologischen Kategorie C gehören
- $V(1, N)$: Gesamtzahl der Hapax Legomena im Korpus
- $N(C)$: Zahl der Tokens, die zur morphologischen Kategorie C gehören

Potential Productivity



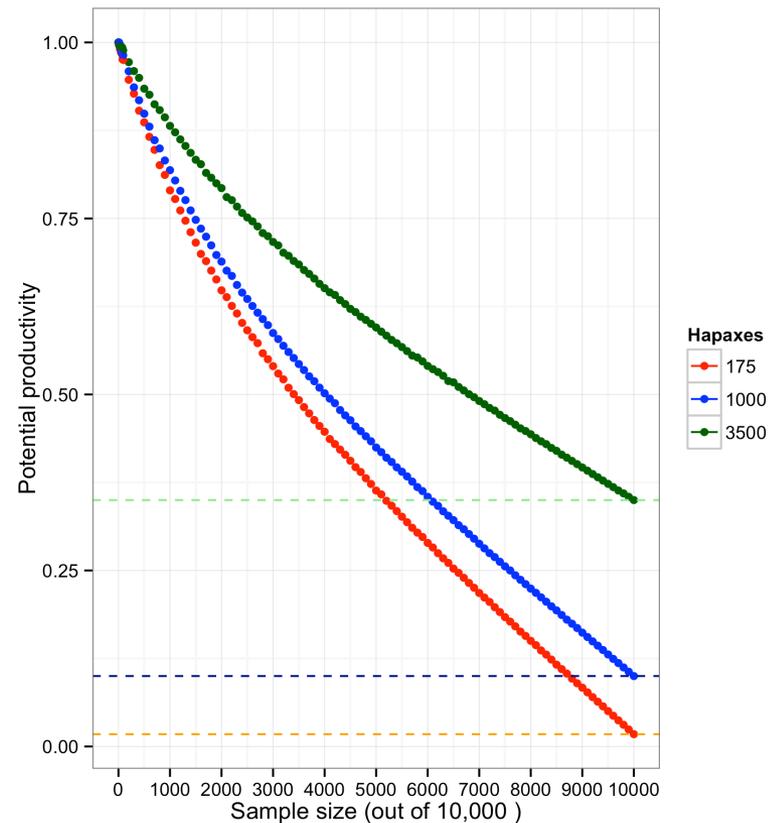


Probleme der Produktivitätsanalyse

- Maß der potentiellen Produktivität ist von der Korpusgröße und von der Anzahl der Tokens abhängig (Baayen 2001; Gaeta & Ricca 2006; Hilpert 2013)
- wenn sich Anzahl der Tokens stark verändert, kann dies den Produktivitätswert verzerren.

Probleme der Produktivitätsanalyse

- Beispiel: Modell mit 10000 fiktiven Wortbildungsprodukten
- Algorithmus zieht Stichproben von sukzessive zunehmender Größe und berechnet potentielle Produktivität
- drei verschiedene Modelle mit unterschiedlicher Anzahl an Hapaxen



Finites Zipf-Mandelbrot-Modell

- Zipfsches Gesetz: Frequenz von Wörtern ist invers proportional zu ihrer Position in Frequenzrangliste
- Mandelbrots Generalisierung des Zipfschen Gesetzes:

$$\pi_z = \frac{c}{(z+b)^a}$$

- Evert (2004), Evert & Baroni (2007): Zipf-Mandelbrot-Modell zur Extrapolation von Produktivitätswerten für größere Datenmengen (vgl. auch Schneider-Wiejowski 2011; Kempf 2016)