

Korpuslinguistik



Materialien

Github:

- hartmast.github.io/korpling-siegen/

Dropbox:

- tinyurl.com/korpling-siegen2018

Vorbemerkungen

- Keine Angst vor Computern!
- Keine Angst vorm Programmieren!

Ziele

- verstehen, was Korpuslinguistik ist und wofür man sie braucht;
- eigenständigen Umgang mit Korpora erlernen;
- "best practices" für korpuslinguistische Studien kennenlernen.

Organisatorisches

- Prüfungsform: Hausarbeit
- Themen frei wählbar, aber korpuslinguistische Herangehensweise notwendig
- Bearbeitung in Kleingruppen möglich
- Mehr zur Hausarbeit morgen Vormittag!
- Kontakt:

sic!

stefan1.hartmann@uni-bamberg.de

Was ist ein Korpus
und wie kann man es benutzen?

Zeitmaschinen...



Zeitmaschinen...



Zeitmaschinen...



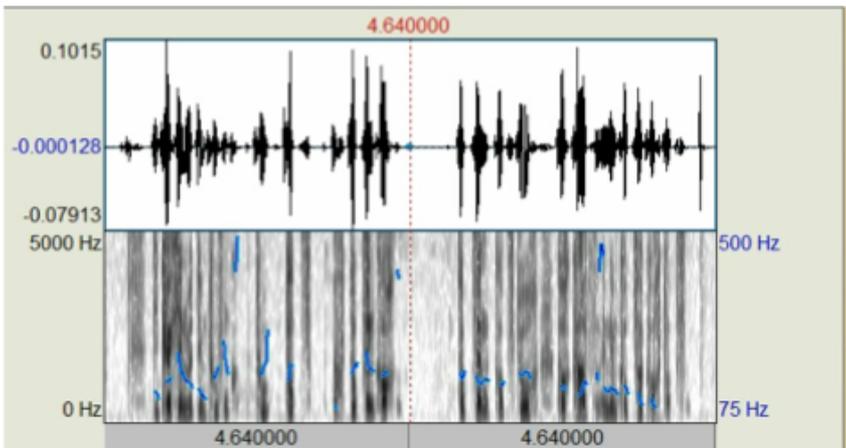
Chymische Poeterey:
 Darumb so laß euch große klag/
 Was ihl vnd etlich wenig sag.
 So bald er die wort vollend/ wand der Deckel
 wider zu gethan/ vnd verschloß/ vnd das Trom-
 men vnd Hornen mehr wider angelohet/ So
 laß koch aber der Thon nit fei/ man hier noch
 der gefangenen bitter klag/ die sich im Thon er-
 hoben für allen berath. Welches mir dann auch
 bald die Augen vbergetriden. Bald fetz sich die
 alt Frau mit ihrem Sohn auff mehrere Seßel
 nieder/ vnd heische die erliche zu schen. Wie sie
 nun die zahl vernommen/ vnd auff ein Goldgub
 Taffel auff geschriben/ begert sie eines jeden
 Namen/ welche auch von ein Knecht auffge-
 schriben worden: Wie sie vns nun nach einander
 andert/ erwidert sie/ vnd heische zu ihrem Sohn/
 das ich wol hören kantz. Ach wie turen mich
 die arme Mordchen im Thon so vbel/ vnd Got/
 ich dierfür se alle erlögen. Daruff der Sohn
 gawantet: Mutter/ so ihl von Gott verordnet/
 dem sollen wir nit widerstehen/ wann wir alle
 Herten weren/ vnd alles Gut hetten auff Erden/
 vnd weren dann zu Dinch gelisten/ wer wolt vns
 doch bringen zu offen. Darwegen die Mutter
 geschwiegen/ Aber bald daruff sagt sie: Nahn ihl
 laß doch diese von jhren Spitzigen erlögen:
 Welches dann auch schnell geschehen/ vnd war
 ich ohn wenig der kette. Noch kantz ich mich
 nit enthalten/ ob ich wol als auff andere gelisten/
 sonder neigt mich vor der dem Herten/ vnd
 dancket Got/ der durch sie mich auff solchem Fin-
 (firtung)

```

# Corpus: bnc (British National Corpus (BNC edition))
# Name: BNC:Last
# Size: 21524 intervals/matches
# Context: 30 characters left, 30 characters right
# Query: BNC: [word="future"]:
6145: s been discussing planning for [future] projects with ACEI 's African
13688: s put at risk the security and [future] of their nearest and dearest
15532: gng them to think about their [future] . So far we have visited over
15532: y infected . The effect in the [future] will be devastating . ACEI 's
19223: ED his or her entire salary in [future] AIDS treatment costs alone .
19782: sr 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: eshere , 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 obs
95900: kward ( & hidden in the near [future] he was to be proved right
115567: ettle them 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
156731: f parts he/she may play in the [future] , or indeed may have played
159398: 's going to be any use for the [future] . In the first place you are
162004: ay of commanding interest from [future] employers . Certainly the new
166701: '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
20350: e , chooses his words over the [future] of the North . He is careful
281720: 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
 paintings. >> this will
 become a story, 200 years
 in the future. well, that big
 fire of 2012. and then we

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 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

morphologischer Wandel:

feuchtiger Natur, keine Baert

graphematischer Wandel:

mehr Saamens; / (Virgel)

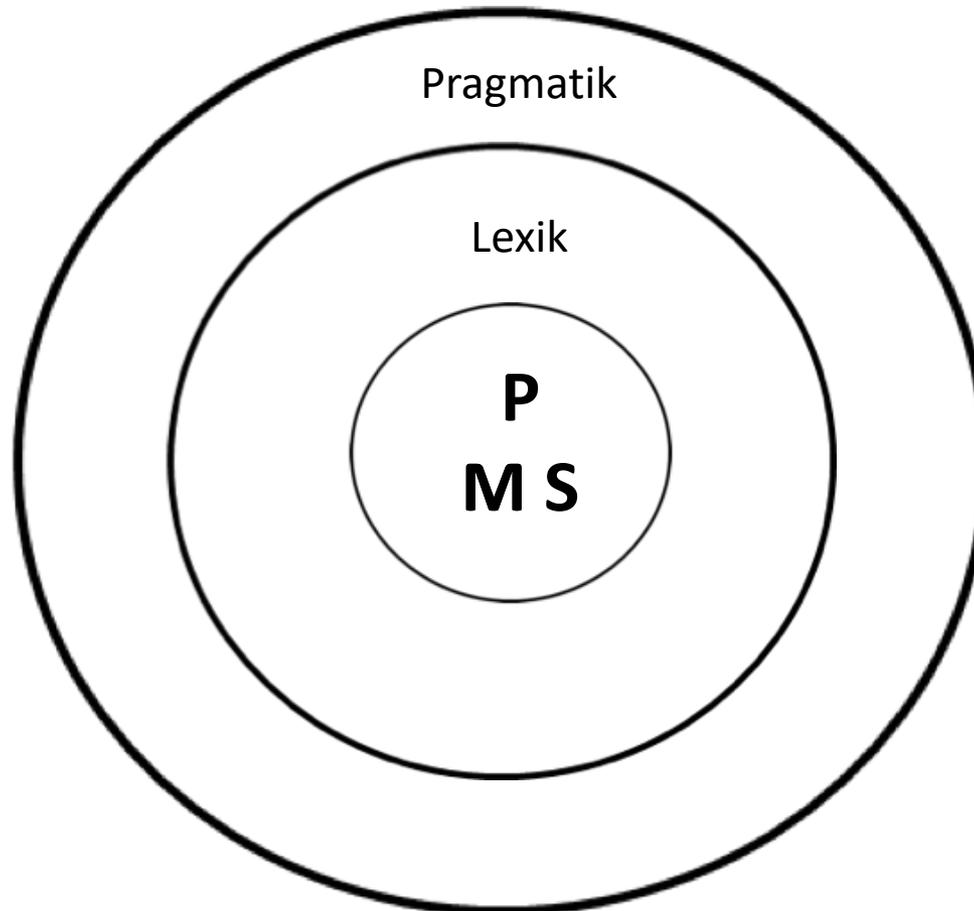
syntaktischer Wandel:

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.)

1. 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)

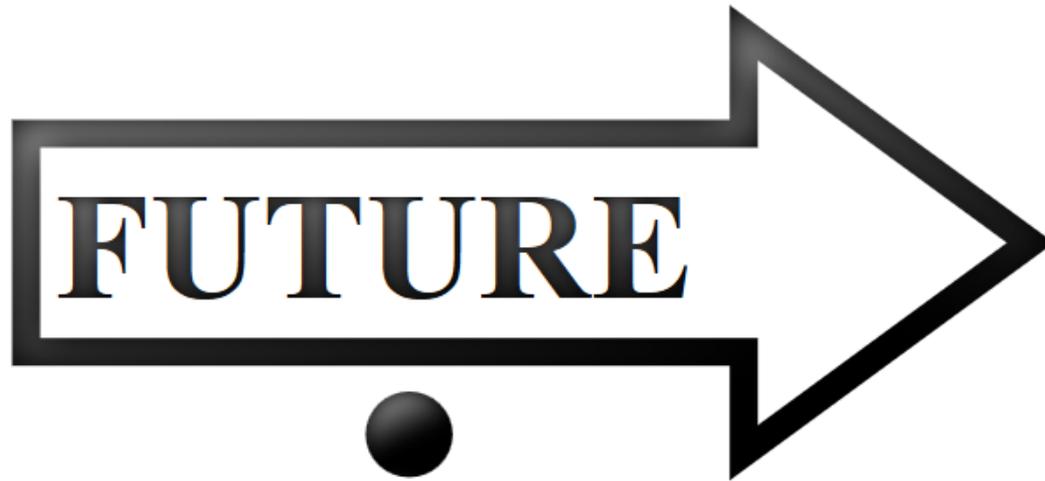
2. Die Zukunft beginnt heute...

- Korpuslinguistik als Disziplin „im Aufbruch“
- Verfügbarkeit von immer mehr Daten ermöglicht neue Fragestellungen
- Verfügbarkeit neuer methodischer Ansätze ermöglicht komplexere Fragestellungen

Was die Zukunft bringt...

In diesem Kurs wollen wir...

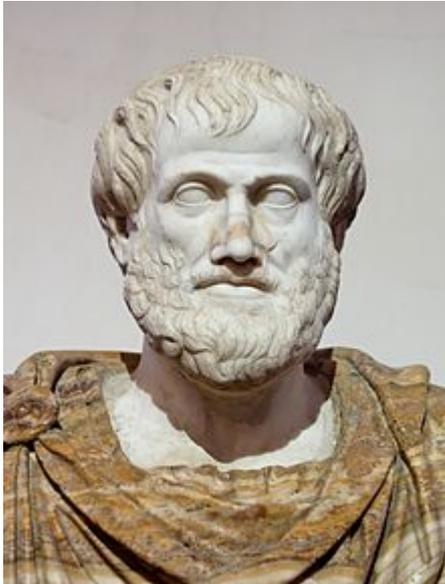
- erfahren, wie ein gutes Korpus erstellt wird
- **bestehende Korpora kennenlernen**
- mit Korpora umgehen lernen
- grundlegende statistische Methoden kennenlernen



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.*

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

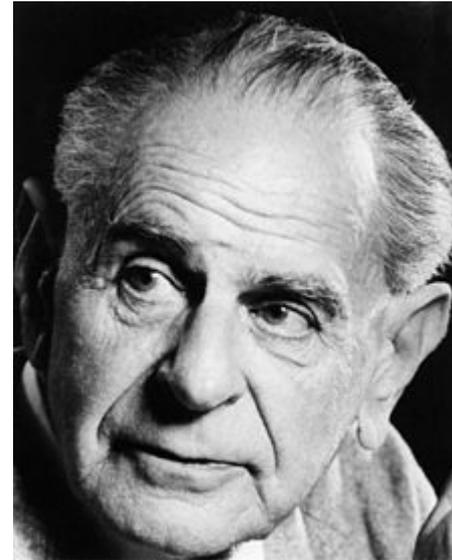


Logischer Positivismus

- 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.)

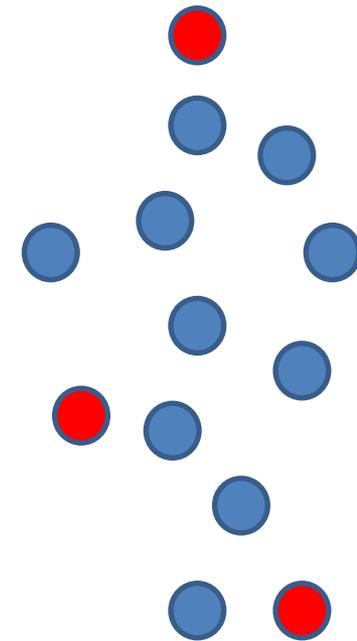
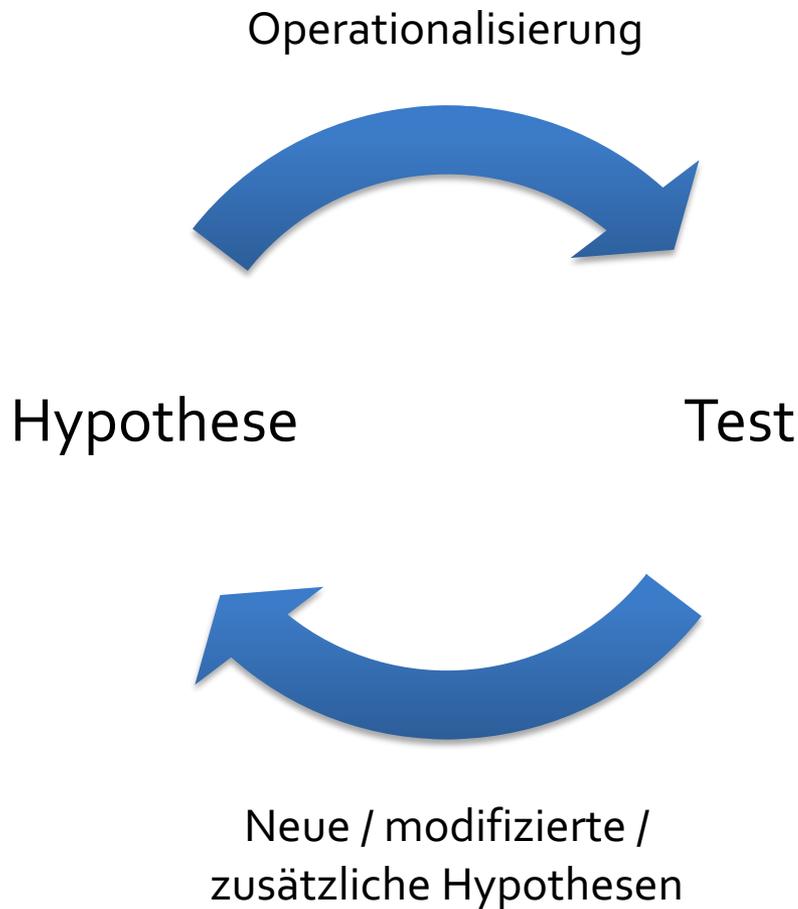
Zusammenhang zur Korpuslinguistik

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

Aber...

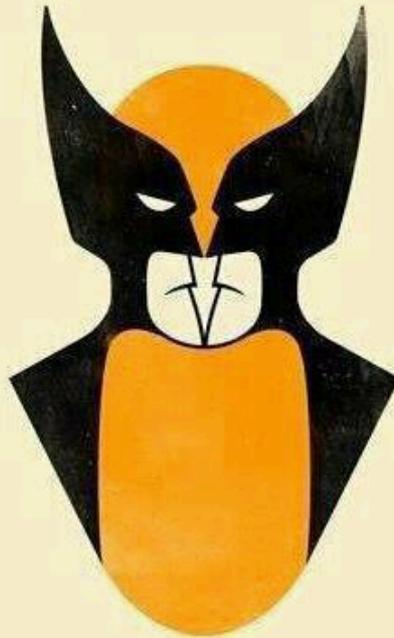
- Nicht alle (korpus-)linguistischen Methoden sind falsifikationistisch.
- Verbreitet ist auch **exploratives** Arbeiten: Muster in den Daten erkennen.
- Auch die derzeit sehr gängigen **Bayesschen Ansätze** sind nicht, oder zumindest nicht immer, falsifikationistisch orientiert.

Deduktive vs. induktive Methode



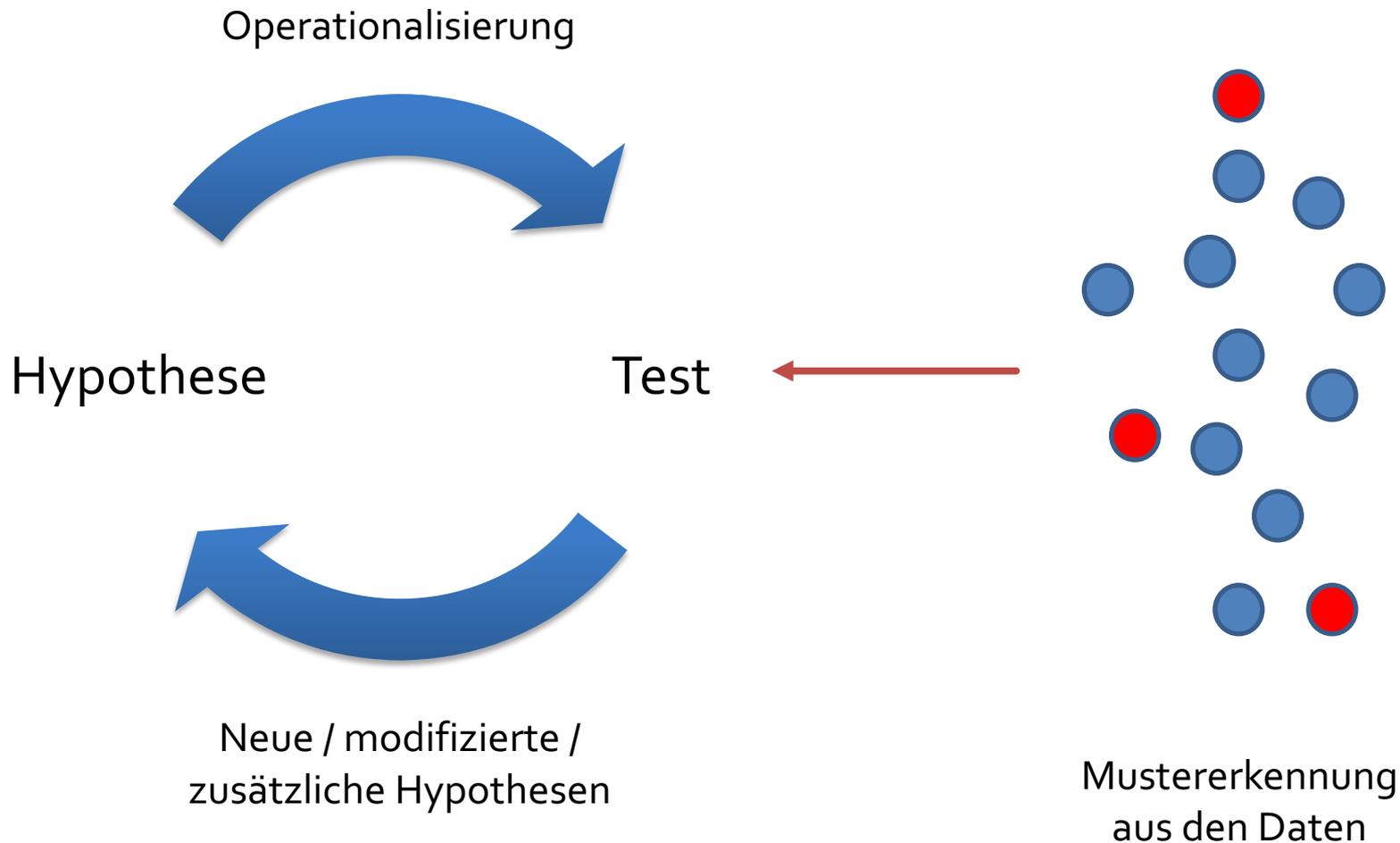
Mustererkennung
aus den Daten

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, a navigation menu on the left, and a list of search results. The results are displayed in a table with columns for the source domain, the search query, and a snippet of the text containing the query.

Source	Snippet
baeng-2000.de	genügen, doch dann findet sich unter einem Boot am Strand der skalpierte schwedische Ex-Justizminister . Das ist erst der Anfang der Katastrophe Mord reiht sich an Mord , e
oreilly.de	entscheiden, sollten Sie beginnen, sich Notizen für eine Site mit mindestens 100 Seiten zu machen . Und das ist erst der Anfang . Hiermit sind 100 Seiten mit » echtem « Inhal
exil.de	südlichen Afrika gilt . Eine Platzierung in den Top 10 der World Music Charts Europe folgte prompt , und das ist erst der Anfang ... Vier- und Marschlande à Kulturlandschaft i
greenpeace-magazin.de	entstehen . Das sieht der jüngst vom Kopenhagener Kabinett beschlossene Energie-21-Plan vor . Doch das ist erst der Anfang : Bis zum Jahr 2030 soll es Dutzende Windfarm
literaturline.stadt-muenster.de	Portrait . HR : 1985 . Stadt , Land , Fluß . Eine Kinderspielshow . ARD : 1982 . Film , Video : ...und das ist erst der Anfang . Spielfilm 2000 : Regie : Pierre Franck . Jahre
konsument.at	kommt . Nach zwei Monaten erstickt Europa an Überschüssen , und Brüssel ruft den Notstand aus . Aber das ist erst der Anfang . Niemals wieder hatte ein Hörspiel eine so di
bild.t-online.de	Erdhalbkugel, lassen Autos durch die Luft wirbeln, pulverisieren Häuser und sogar ganze Wolkenkratzer à und das ist erst der Anfang ! Trailer Sehen Sie hier den Film-Trailer Se
kika.de	das Rätsel um den Drachen lösen und nebenbei noch ihren rücksichtslosen Verfolgern entkommen . Aber das ist erst der Anfang der Abenteuer , die auf die Freunde zukommi
zeit.de	stigmatisiert, dass sie nichts zum Erreichen der vom Staat erwünschten Geburtenplanziffer beitragen . Und das ist erst der Anfang einer groß angelegten Sündenbockabstempel
djfl.de	maskierter Schurke erscheint im Museum und macht Mystery Inc. für den Angriff verantwortlich . Doch das ist erst der Anfang : Der rätselhafte Bösewicht hat eine Maschin
e2ie2i.at	atomar-fossile Energien mehr als verdoppelt, während sie sich für Erneuerbare Energien halbiert haben Und das ist erst der Anfang einer Entwicklung, die sich noch dramatisch
karlsruhe.de	Neue in der nicht einfachen Klasse 3c und wird wegen seiner Sommersprossen von Florian gehänselt . Doch das ist erst der Anfang und die Lehrer scheinen nichts zu sehen . Mi
politik-forum.at	, Colleges, Hotels, Krankenhäusern, Kreditkarten- und vielen anderen Unternehmen anfordern . Und das ist erst der Anfang . Erst letzte Woche hat Bundesanwalt Genera
freitag.de	leistet er Widerstand, unberührt steht das Schnapsglas vor ihm, und dann trinkt er plötzlich doch . Aber das ist erst der Anfang des Unglücks, an dessen Ende die totale Des
literaturkritik.de	hatte dieser sein Land vor vielen Jahren fluchtartig verlassen und blieb seither spurlos verschwunden ? Das ist erst der Anfang einer ständig wachsenden Fülle von Fragen, ,
marktplatz-gp.de	unbeschriftetes Video eine wesentliche Rolle zu spielen scheint . Rachel ist zutiefst schockiert . Doch das ist erst der Anfang . Denn schon bald wird ihr klar, dass hier wie
maulkorbzwang.de	===== BETROFFEN SIND HIER ÜBER 25.000 WOHNUNGEN .. und das ist erst der Anfang !!! [Alles wegen diese Beißmaschinen] §
gazette.de	seit diesem Sommer bietet amazon.com neben Büchern auch Spielzeug und Unterhaltungselektronik . Und das ist erst der Anfang . Amazon.com will etwas werden, sagt sein C
medizin-2000.de	45 biotechnologisch hergestellten Medikamente , wurden innerhalb der letzten drei Jahre eingeführt . Das ist erst der Anfang . Laut PHRMA (Pharmaceutical Research and
bueso.de	Wahlkommission festgestellt hatte, daß er in den letzten sieben Jahren rechtskräftig verurteilt worden war . Und das ist erst der Anfang : Wie Ha'aretz am 24.12. unter Hinweis auf Po
ka-news.de	Stichwort : Lesernah . Die regelmäßige Umfrage ist eine Grundform der Leserbeteiligung bei ka-news . Doch das ist erst der Anfang . Das Team des Nachrichtenmagazins für Karls
christoph-gaebler.de	qualifizierten Schüler in den Lostopf . Bundesweit sind nach Uni-Angaben 400 Schulen bilingual . Und das ist erst der Anfang , wenn man Wissenschaftlern glauben darf, d
djfl.de	TV-Film (RTL) : Millionär und Stripperin - Regie : Donald Krämer - Rolle : Wolff 2000-07-27 : Und das ist erst der Anfang 2000 : TV-Film (RTL) : Der Millionär und die S
n24.at	kämpfen gegen die Flammen . Hunderte Menschen wurden evakuiert , barren in Notunterkünften aus . Und das ist erst der Anfang Kahn : Scheidung / Simone : " Will endlich e

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?

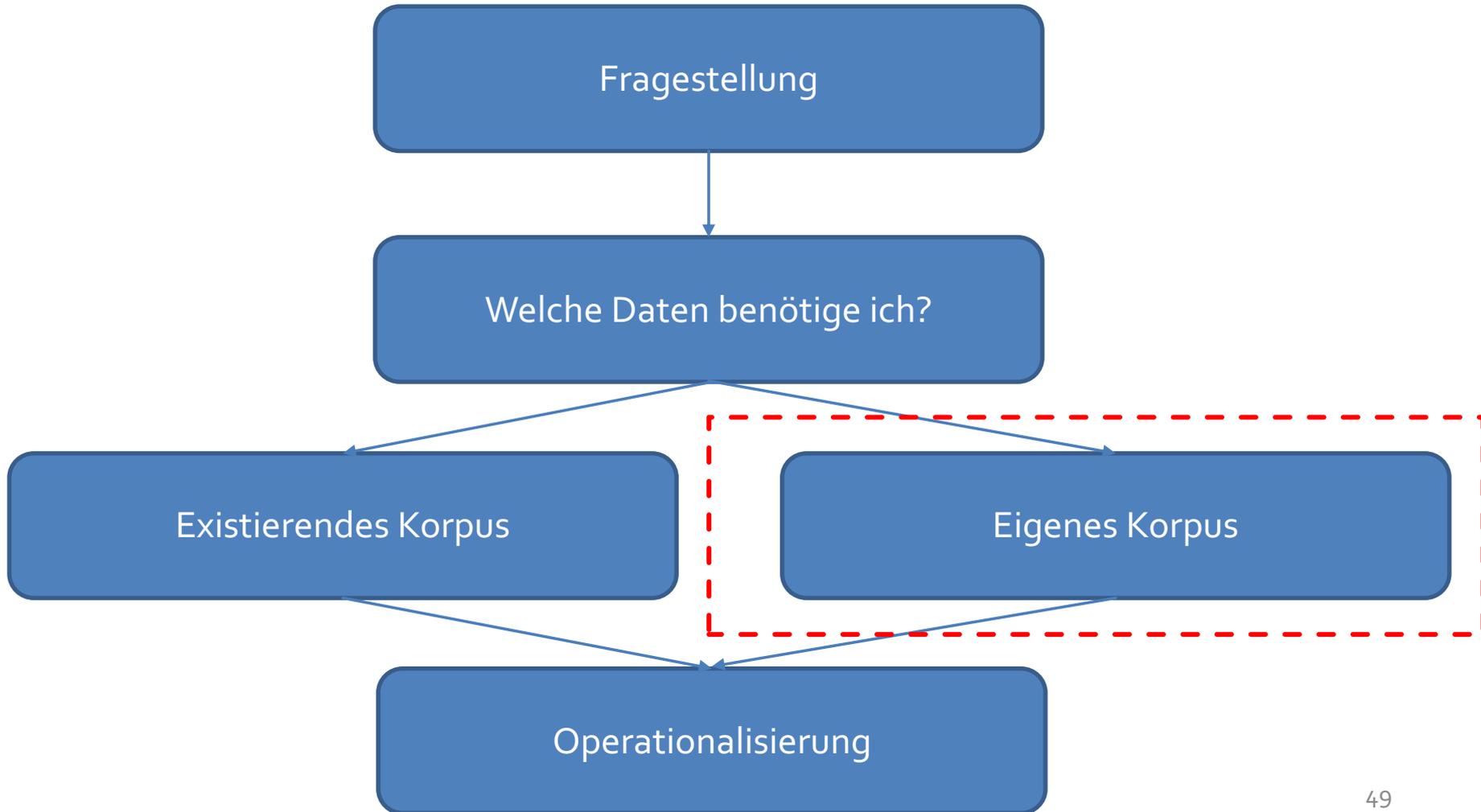
Korpusbasierte vs. korpus-illustrierte Ansätze

- "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

Aufgabe:

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

Psychologen warnen: Kostenloser Nahverkehr würde tausende sadistisch veranlagte Ticketkontrolleure auf Gesellschaft loslassen



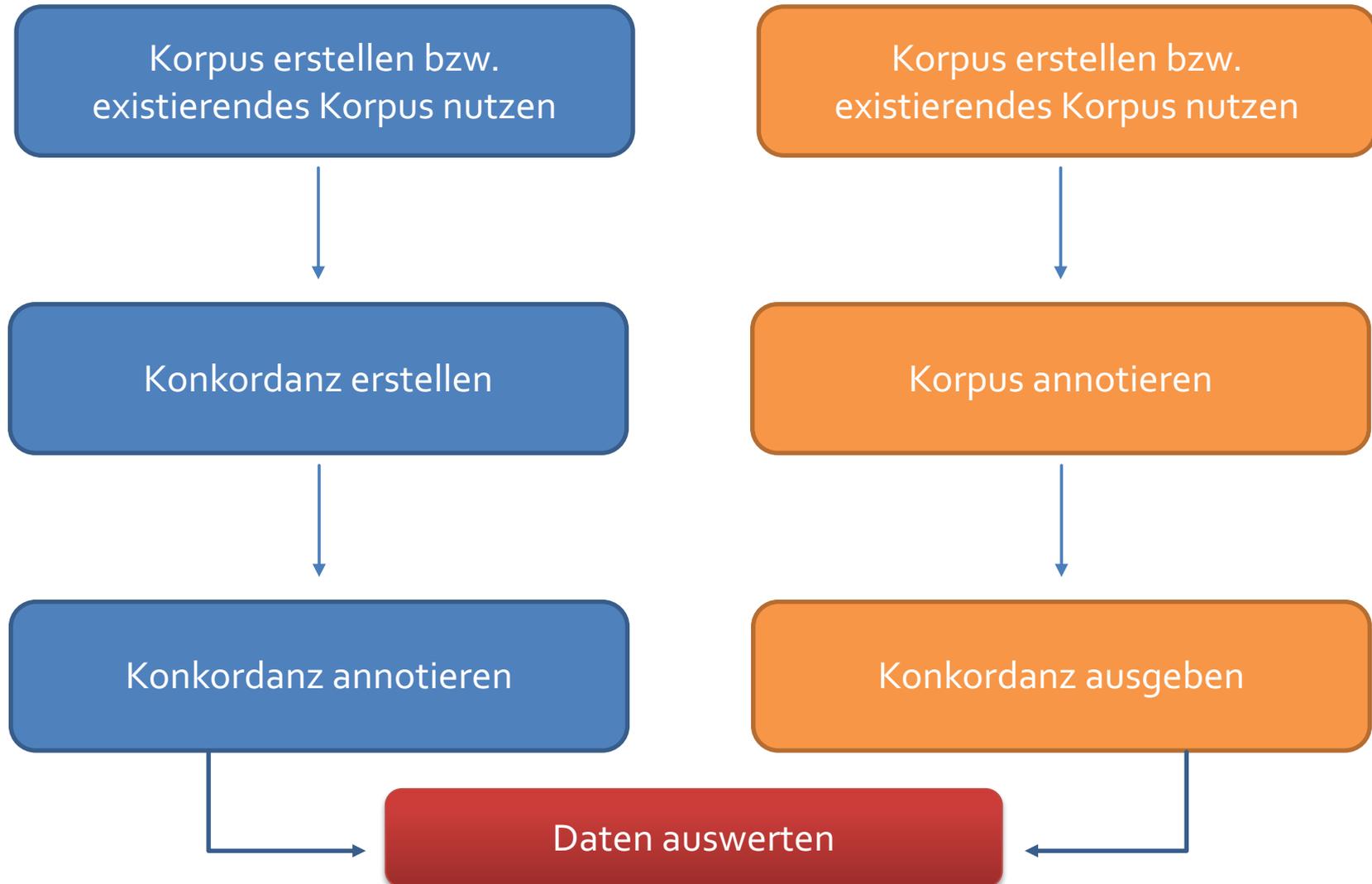
Berlin (dpo) - Seit bekannt wurde, dass die Bundesregierung entsprechende Pläne erwägt, diskutiert Deutschland über kostenlosen Nahverkehr. Nun schaltet sich der Berufsverband Deutscher Psychologinnen und Psychologen (BDP) in die Debatte ein und schlägt Alarm. Eine Abschaffung

der Fahrpreise würde tausende sadistisch veranlagte Ex-Fahrscheinkontrolleure auf die Allgemeinheit loslassen, so die Befürchtung. **mehr...**

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



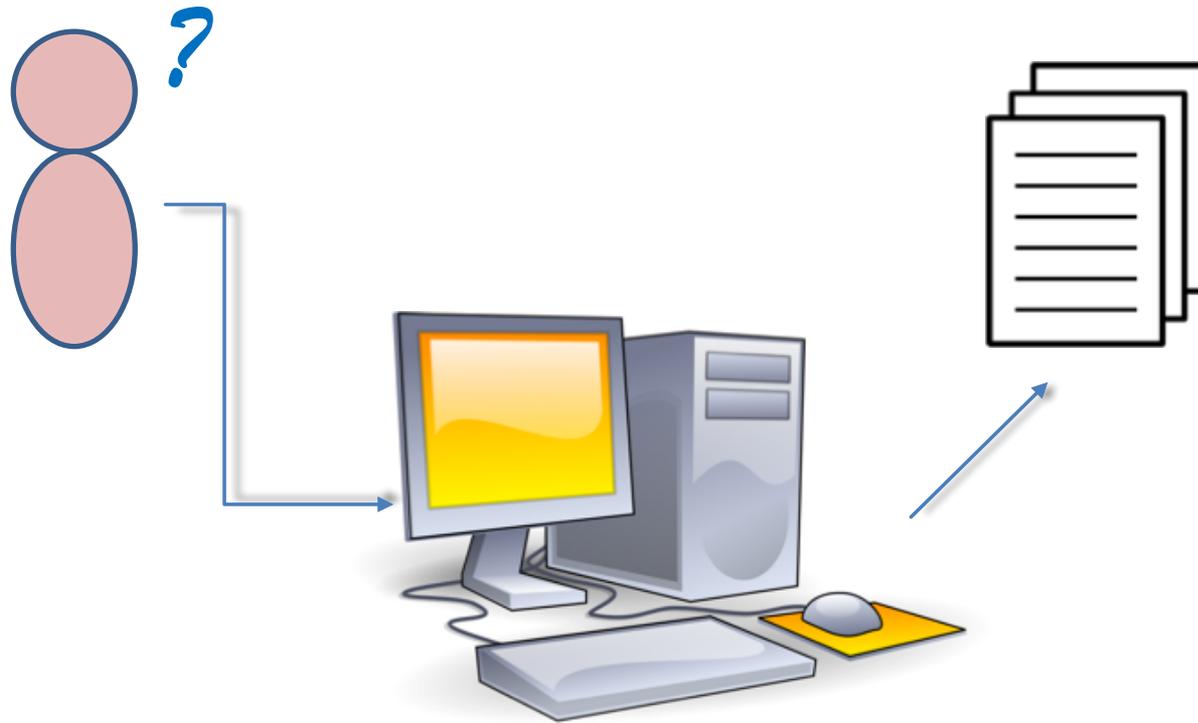
Korpus- oder Konkordanzannotation?

- **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



Fragestellung

Suchabfrage

Konkordanz

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/
l	NN	l

- 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



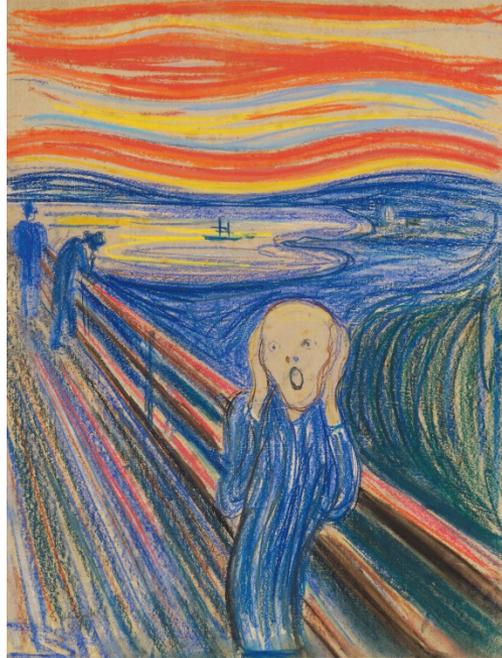
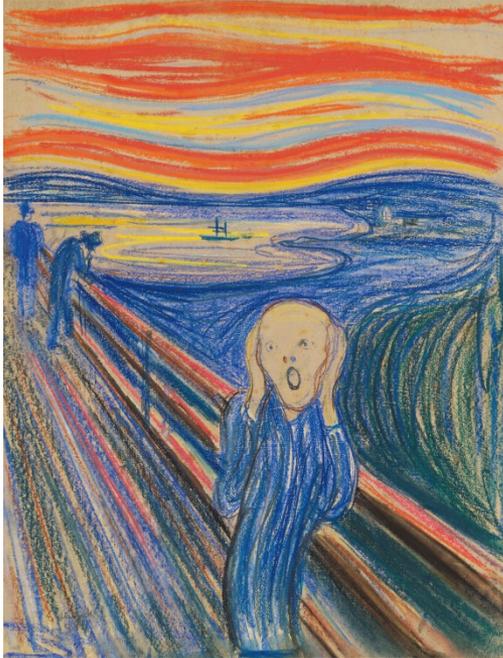
Types vs. Tokens



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. 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.

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*

Qualitativ ^{und} 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

Qualitativ ^{und} 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!