Korpuslinguistik

Materialien

Github:

- hartmast.github.io/korpling-siegen/

Dropbox:

- tinyurl.com/korpling-siegen2018

Vorbemerkungen

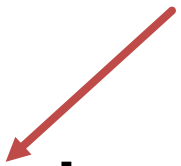
- Keine Angst vor Computern!
- Keine Angst vorm Programmieren!

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:



stefan1.hartmann@uni-bamberg.de

Was ist ein Korpus
und wie kann man es benutzen?

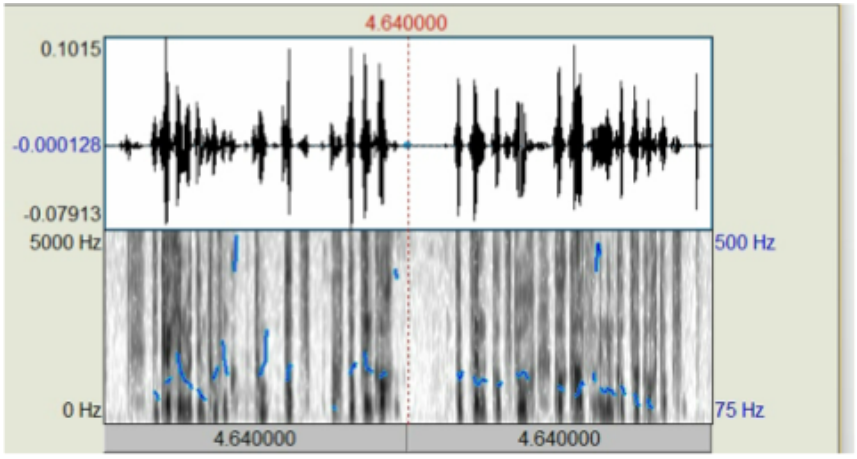
Zeitmaschinen…

Zeitmaschinen…

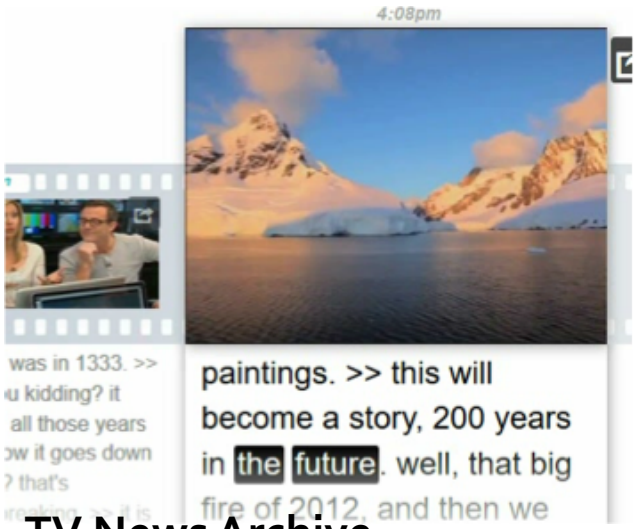
Zeitmaschinen…

Deutsches Textarchiv

British National Corpus



Alcohol Language Corpus



TV News Archive

1. Korpora sind Zeitmaschinen

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 gifftige 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.

1. Korpora sind Zeitmaschinen

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 gifftige 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.

1. Korpora sind Zeitmaschinen

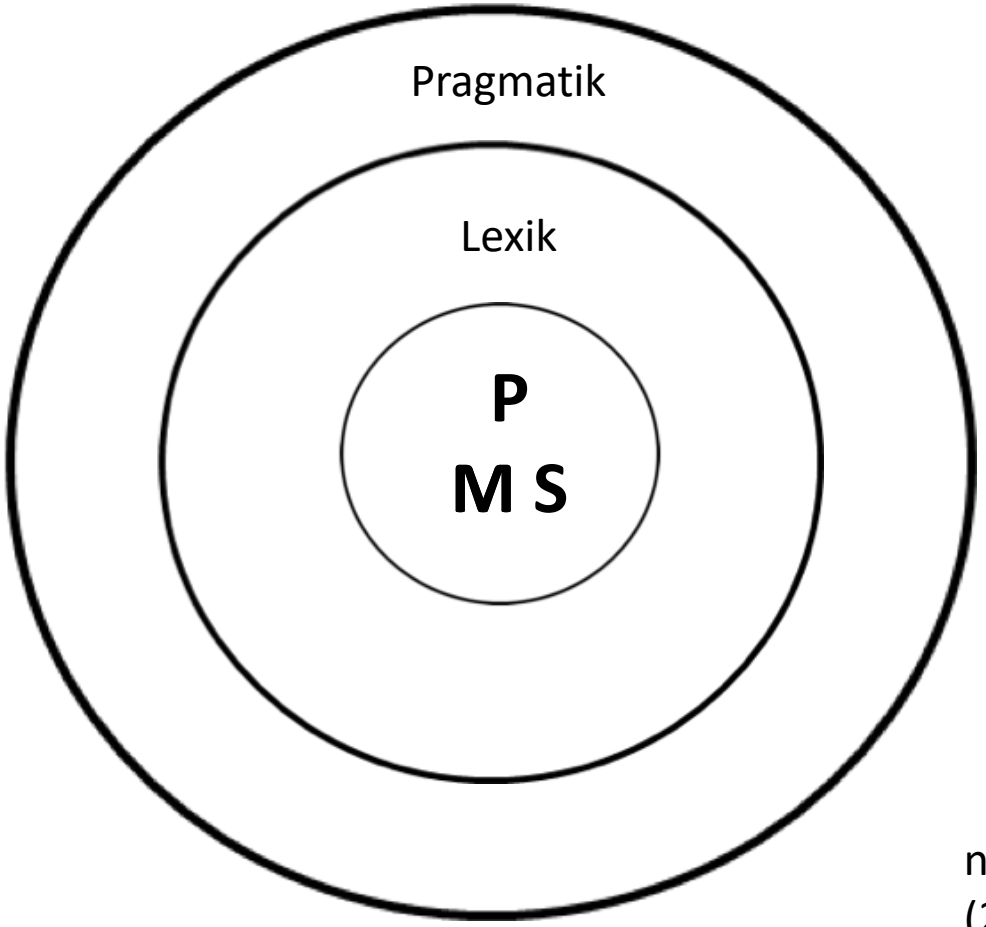
morphologischer Wandel:
feuchtiger Natur, keine Baert

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)

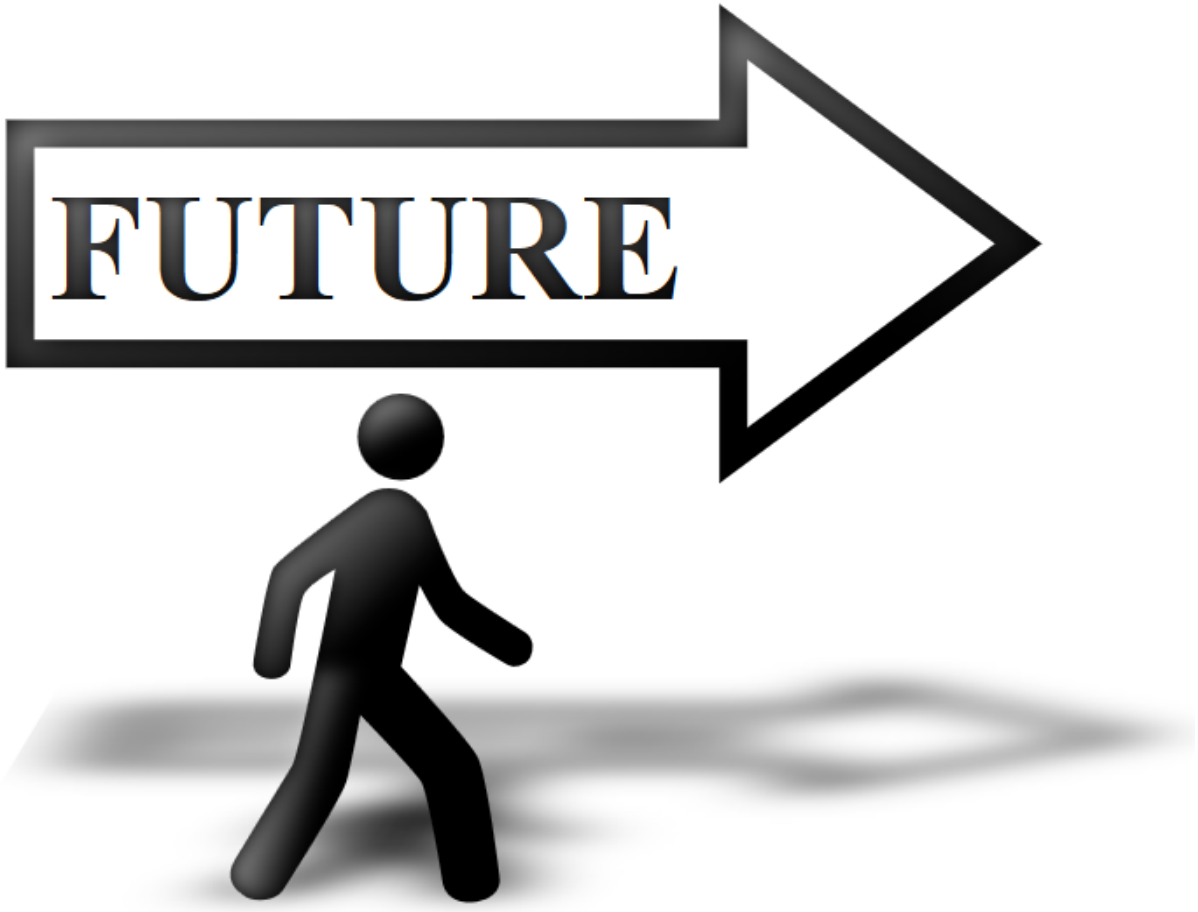
2. Die Zukunft beginnt heute…

- 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…

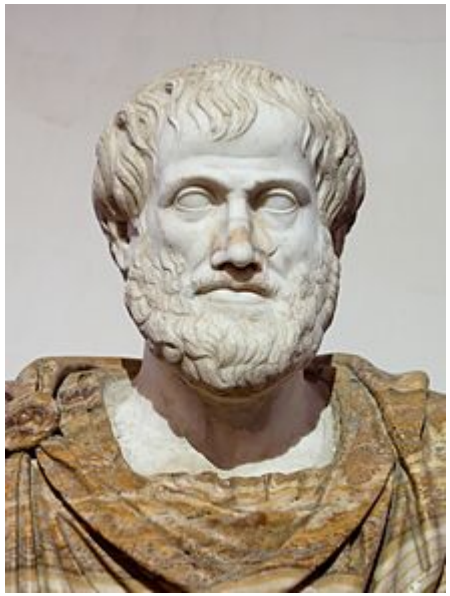
- erfahren, wie ein gutes Korpus erstellt wird
- **bestehende Korpora kennenlernen**
- mit Korpora umgehen lernen
- grundlegende statistische Methoden kennenlernen


FUTURE

Wieso, weshalb, warum? Empirische Methoden und Wissenschaftstheorie

(nach Maxwell & Delaney 1999)

Aristoteles



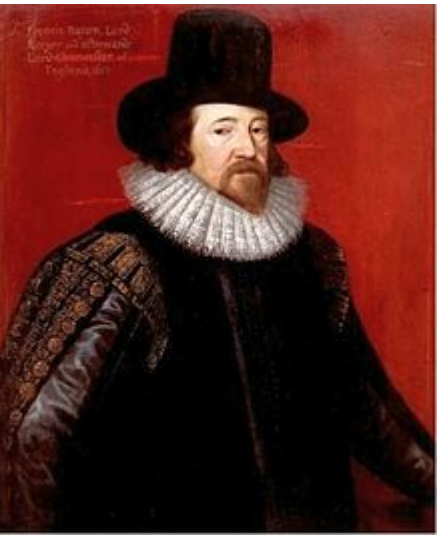
- **deduktive** (ableitende) Methode
- Ideal: Syllogismus

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

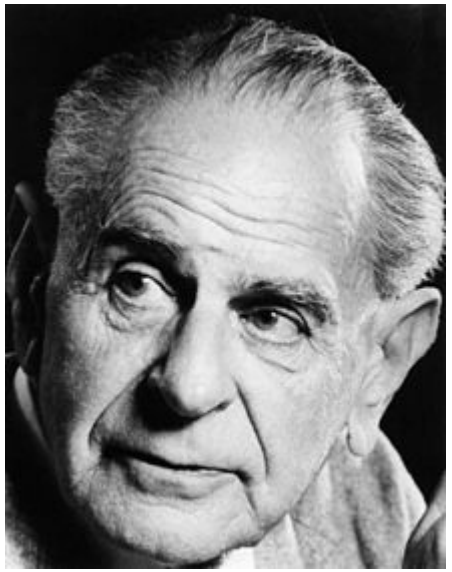
Auguste Comte

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…

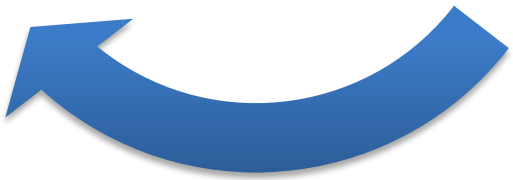
- Nicht alle (korpus-)linguistischen Methoden sind falsifikationistisch.
- Verbreitet ist auch **exploratives** Arbeiten: Muster in den Daten erkennen.
- Auch die derzeit sehr gängigen **Bayesschen Ansätze** sind nicht, oder zumindest nicht immer, falsifikationalistisch orientiert.

Deduktive vs. induktive Methode

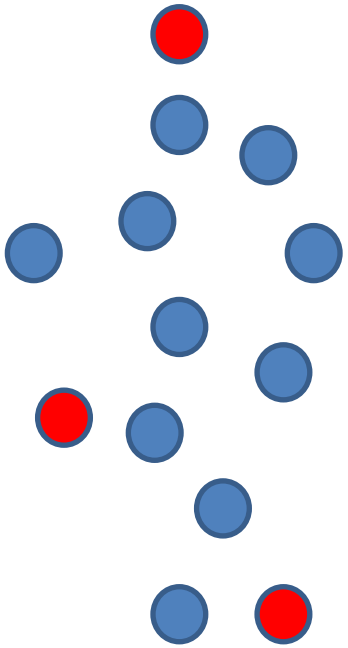
Operationalisierung

Hypothese

Test



Neue / modifizierte / zusätzliche Hypothesen

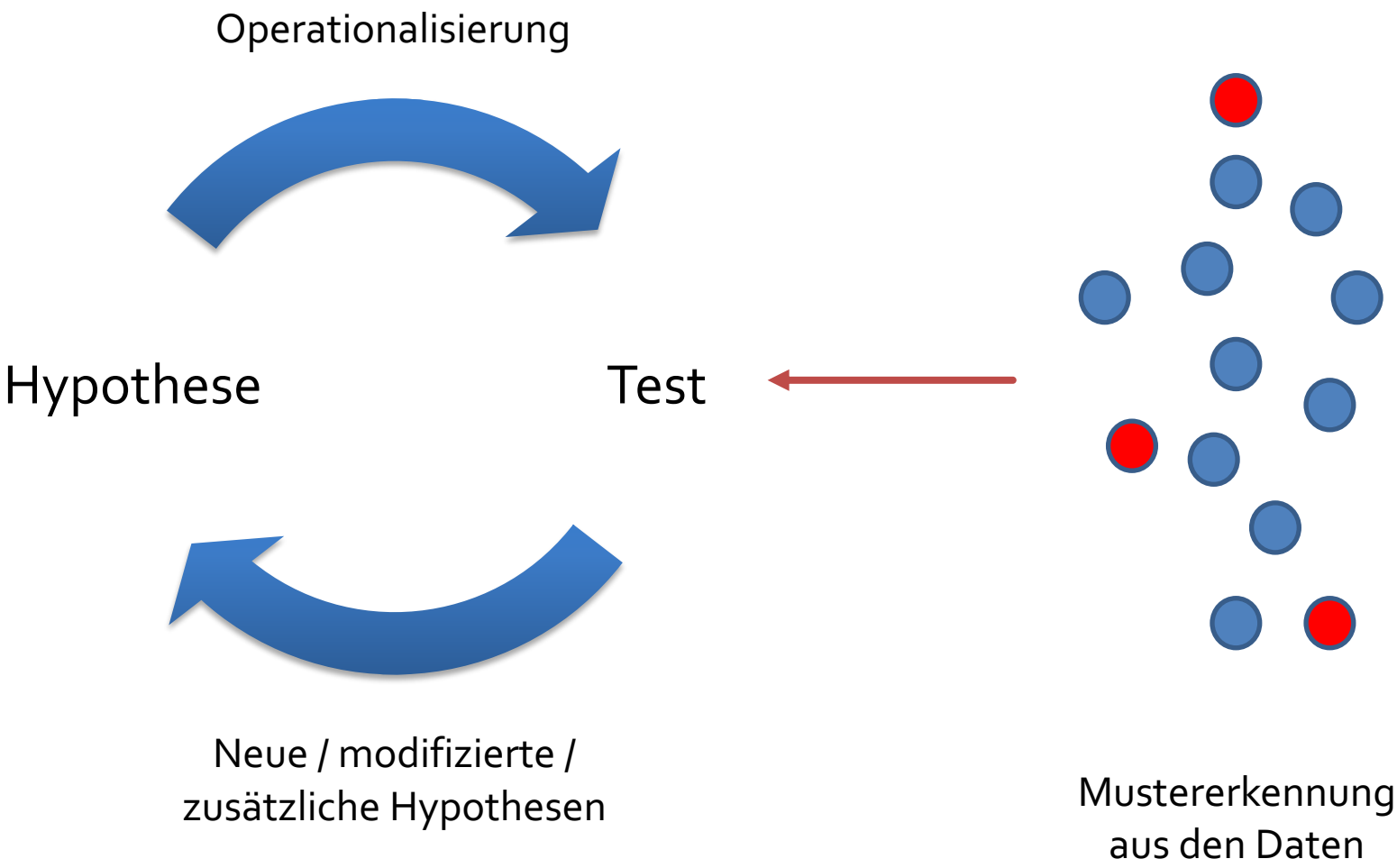


Mustererkennung aus den Daten

(nach Stefanowitsch 2017)


WOLVERINE?....
OR 2 BAT MEN?

Deduktive vs. induktive Methode



(nach Stefanowitsch 2017)

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)

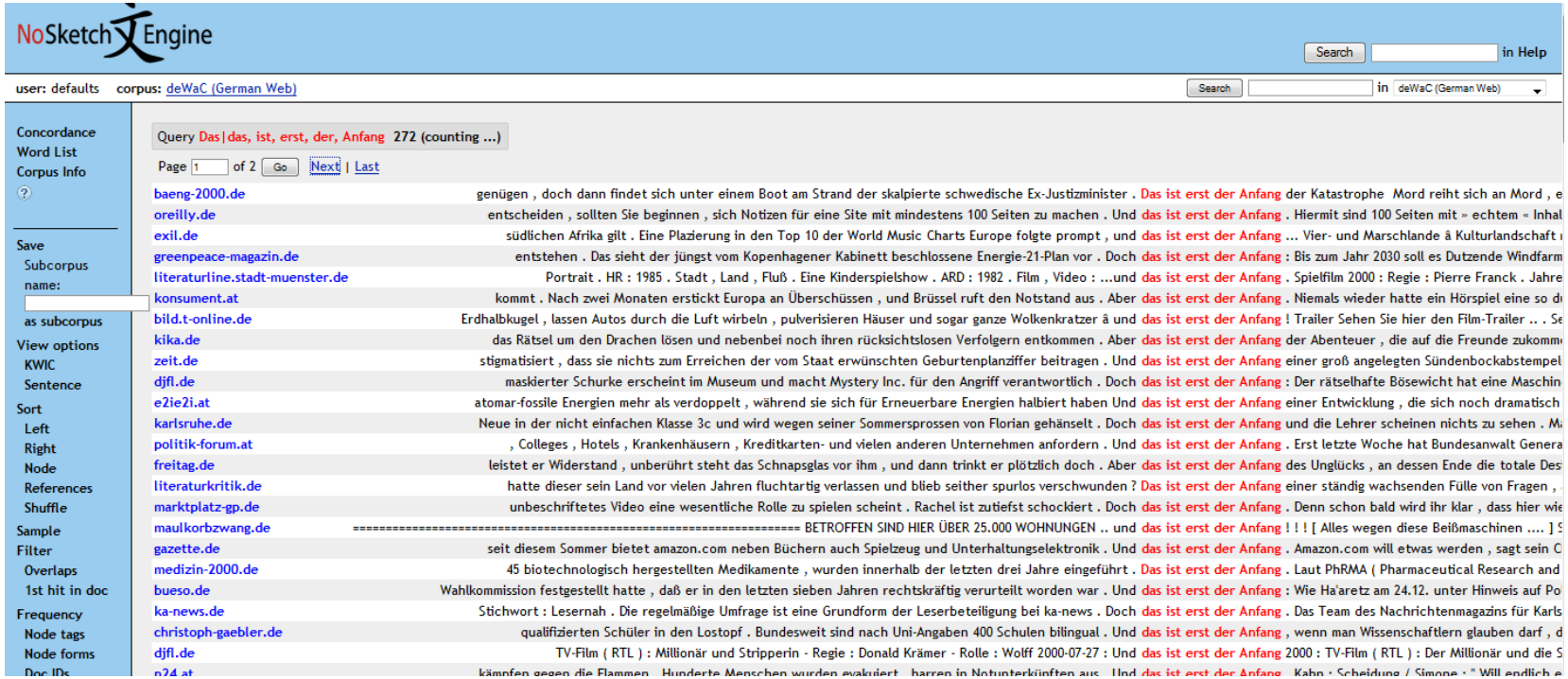
- Demnach ließe sich also leicht die Feststellung treffen, ein kongenialeres und **gleichsam** spannungsreicheres Duo als Joshua Redman und Brad Mehldau ließe sich nur schwerlich im 21. Jahrhundert auf einer Jazzbühne vereinigen. (http://bit.ly/1PQ8m8U)

Können wir unseren Intuitionen trauen?

- Intuiton 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*



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?

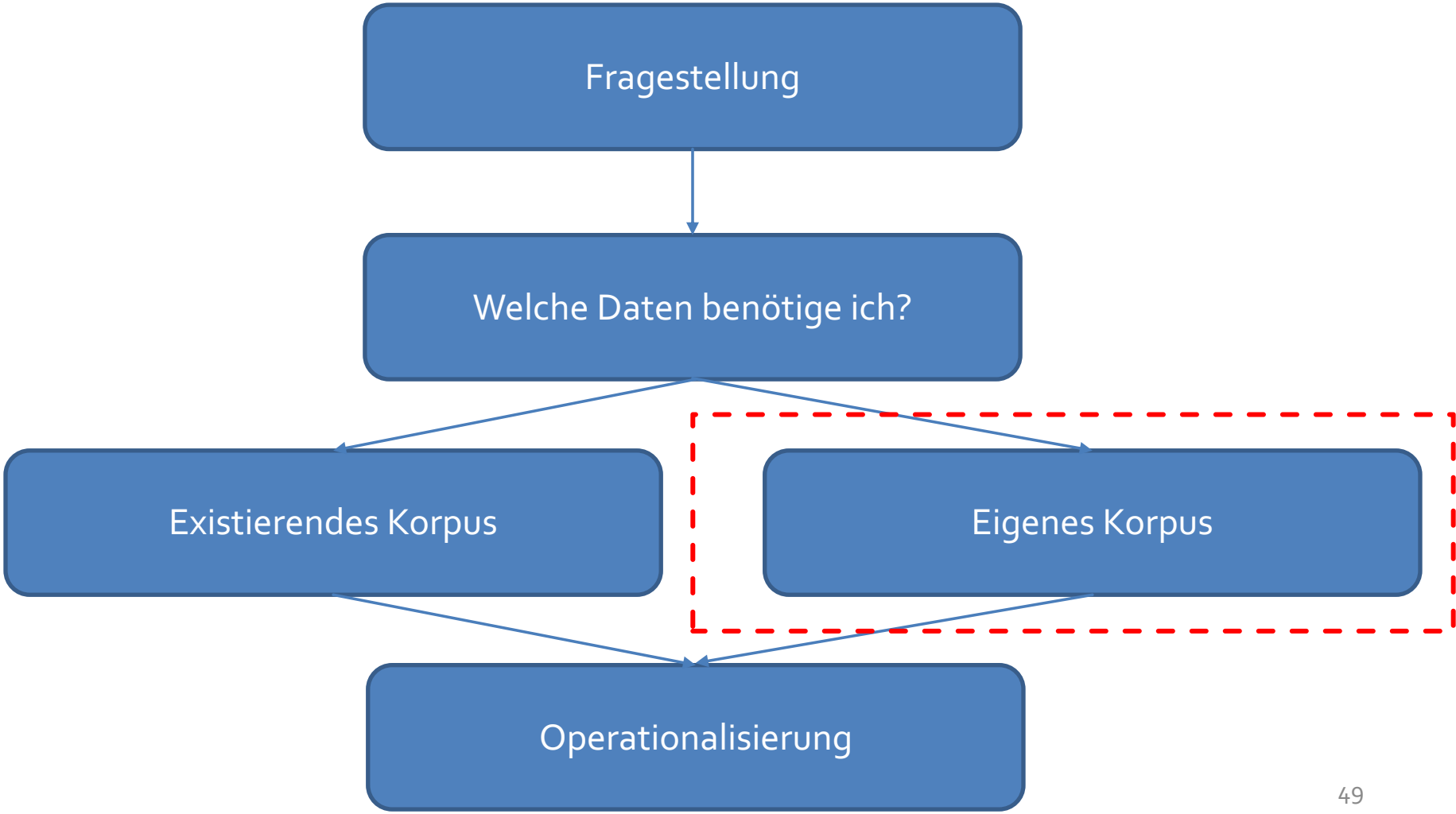
Korpusbasierte vs. korpus-illustrierte Ansätze

- "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

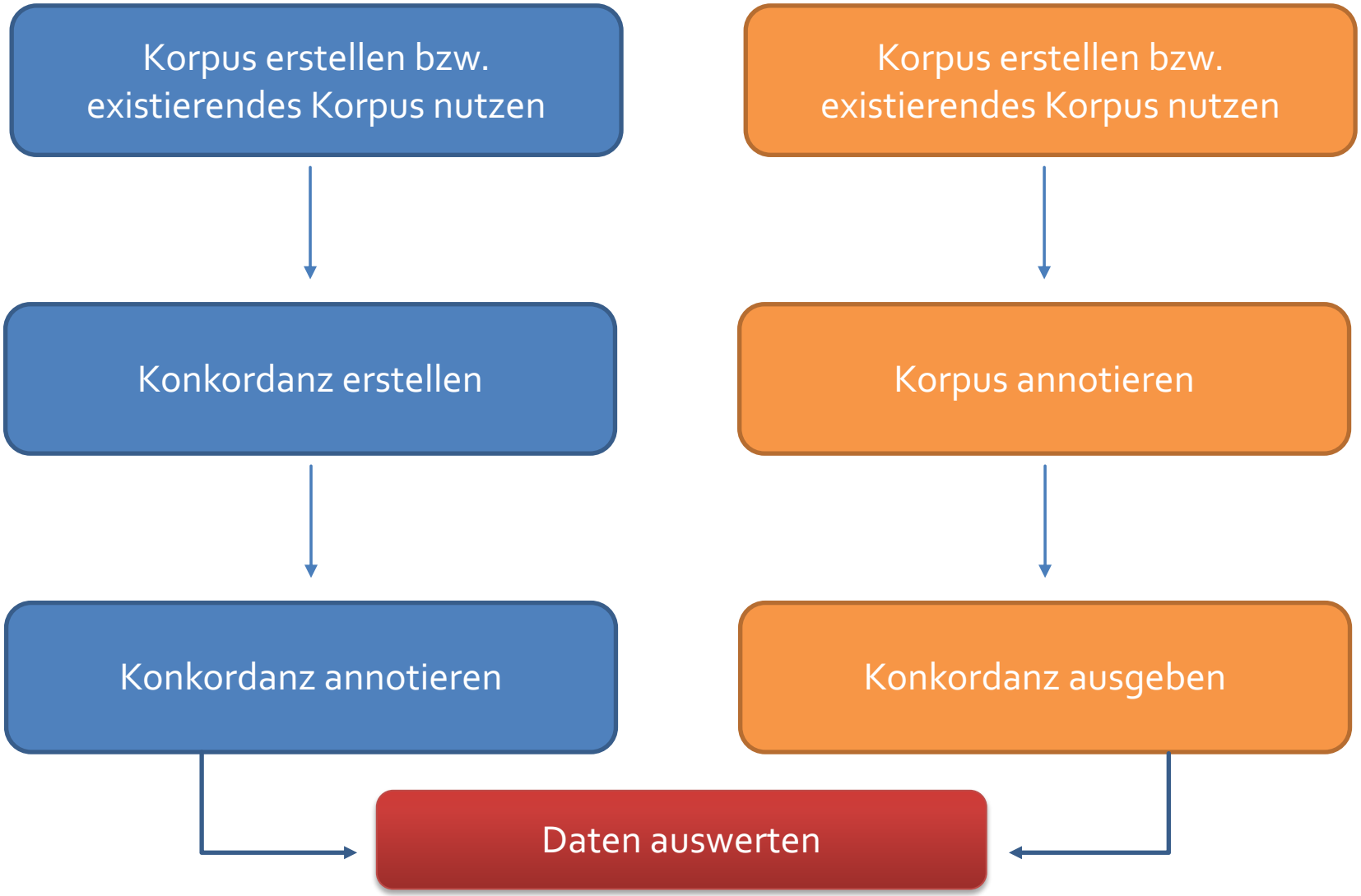
Psychologen warnen: Kostenloser Nahverkehr würde tausende sadistisch veranlagte Ticketkontrolleure auf Gesellschaft loslassen

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

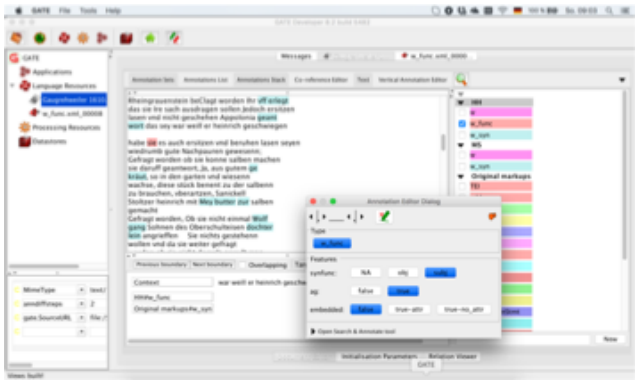
- Vorteil: Postillon-Texte unter Creative-Commons-Lizenz
- Text lässt sich recht einfach extrahieren
- Man muss nur den Seitenquelltext analysieren…

Mögliche Workflows



Annotation: Zwei Wege

Annotation
direkt im Korpus



Export

Export

Annotation
außerhalb des Korpus

Korpus- oder Konkordanzannotation?

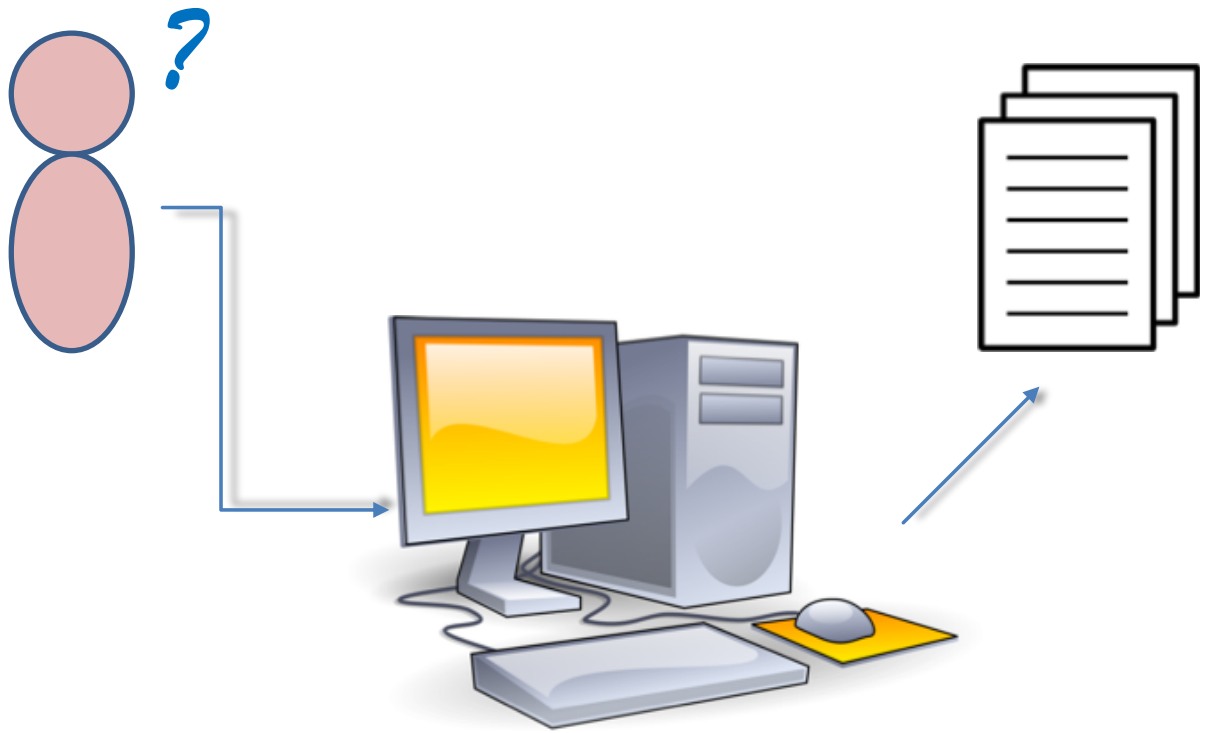
- **Korpusannotation**: Die Annotationen werden für das gesamte Korpus direkt in den Korpusdateien vorgenommen.
- **Konkordanzannotation**: Die Annotation erfolgt nach Ausgabe der Konkordanz in einem Tabellenkalkulationsprogramm.

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/

- 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

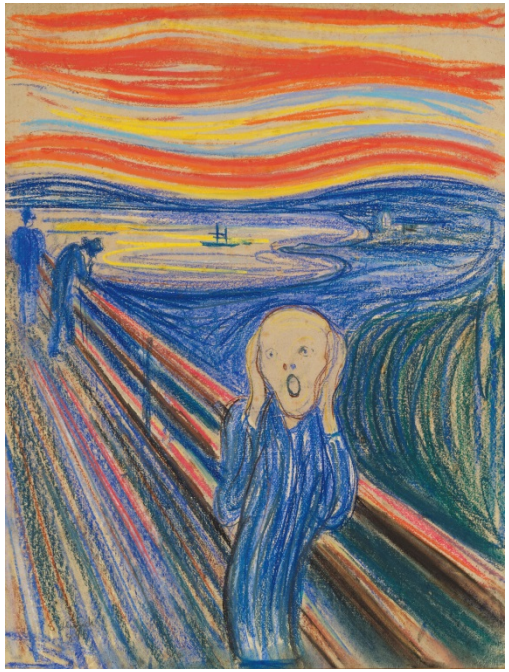


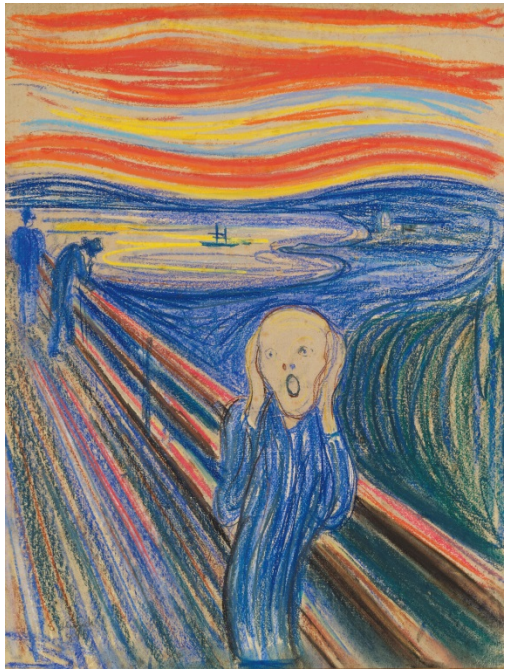



Types vs. Tokens

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 gifftige 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*

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 Sohnen 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!