Empirisches Arbeiten in der Morphologie

Empirisches Arbeiten

- Empirische Forschung = "auf Erfahrung beruhende Forschung" (Albert & Marx 2010: 12)
- eine **systematisch zu erfassende** Erfahrung bildet die Grundlage, um die jeweilige Fragestellung zu beantworten.

Empirisches Arbeiten

- Welche Arten des empirischen Arbeitens kennen Sie?

Empirisches Arbeiten

Empirische Methoden umfassen u.a.

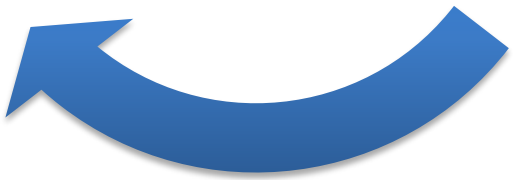
- Fragebogenstudien
- Experimente
- Arbeit mit authentischen Daten
  - → Korpuslinguistik

Deduktive vs. induktive Methode

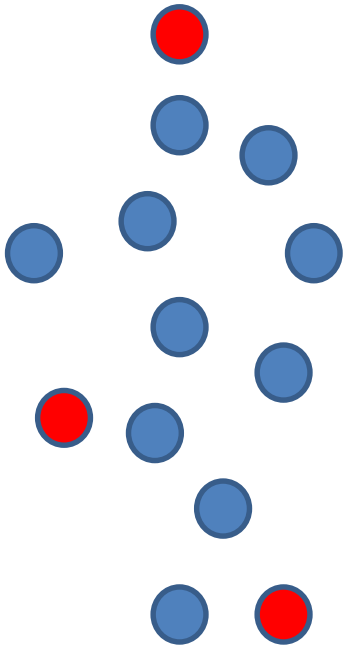
Operationalisierung

Hypothese

Test



Neue / modifizierte /
zusätzliche Hypothesen



Mustererkennung
aus den Daten

(nach Stefanowitsch 2017)

Was ist ein Korpus
und wie kann man es benutzen?

Korpuslinguistik

- Korpus: strukturierte Sammlung authentischer Sprachdaten

Wozu Korpuslinguistik?

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 die eigene Intuition allein reicht oft nicht!

Anwendungsbereiche

Welche der folgenden Fragestellungen lassen sich mit Hilfe von Korpusdaten angehen?

- Hat sich die Frequenz von *backte* vs. *buk* diachron verändert?
- Heißt es *die Wagen* oder *die Wägen*?
- In welchem Bereich / welchen Bereichen des Gehirns wird Sprache verarbeitet?
- Ist die Form *ging* stärker kognitiv verankert als die Form *Schnellfahrstreckenneubaugenehmigung*?

Anwendungsbereiche…

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

Was ist Korpuslinguistik?

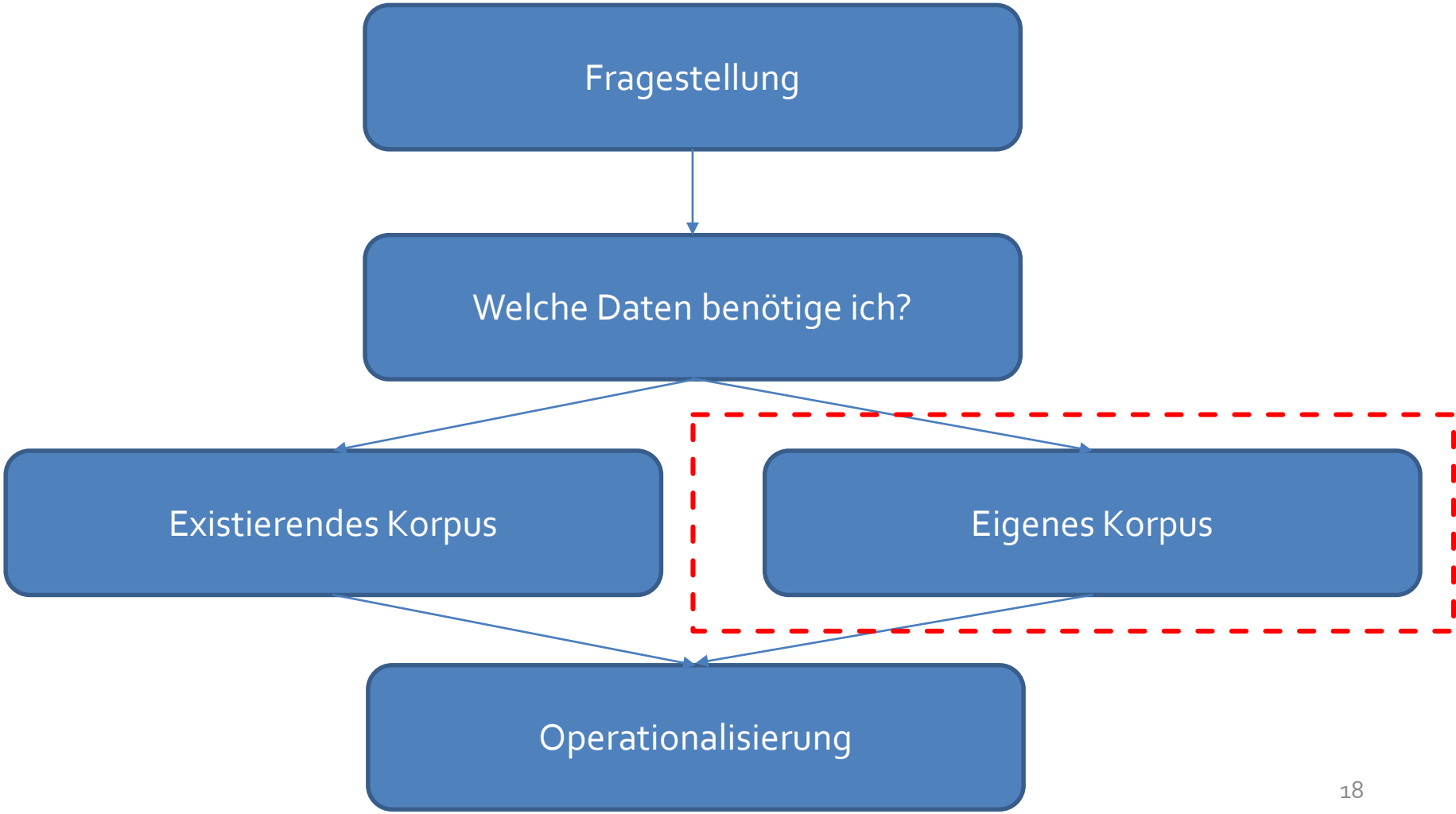
Korpusbasierte vs. korpus-illustrierte Ansätze

- "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 Bamberg sprechen.

Wie gehen Sie vor?

Korpusdesign

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

bei transliterierten Texten:

- Qualität der Transliteration

Ausgewogenheit und Repräsentativität

- " […] arguments that a particular corpus is representative, or balanced, are inevitably circular, in that the categories we are invited to observe are artifacts of the design procedure" (Hunston 2008)

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), Corpus linguistics. An international handbook, vol. 1, 154–168. Berlin: Walter de Gruyter

Korpusdesign

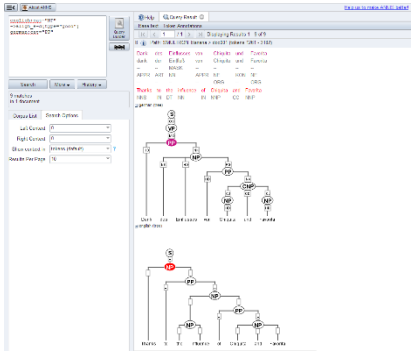
Grundsätzliche Fragen:

- 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.

Korpuserstellung

Text

Aufbereitung, Annotation



Zugänglich-machung in durchsuchbarem Format

Tools:

- automatische Tagger
- Programme zur manuellen Annotation

Korpusauswertung

- In dieser Sitzung beschränken wir uns auf die Arbeit mit **existierenden** Korpora.

- Kennen Sie bereits Korpora? Wenn ja, welche?

Korpora des Deutschen

Gegenwartssprache:

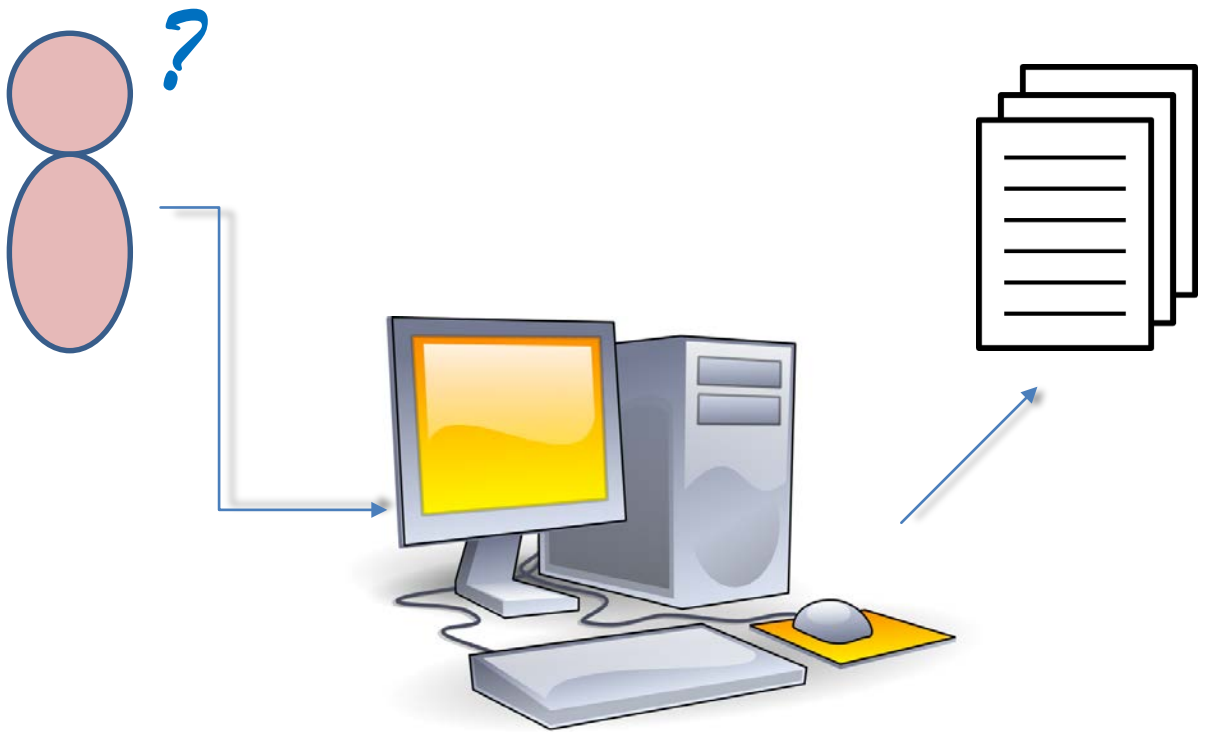
- Deutsches Referenzkorpus (IDS Mannheim, zugänglich über COSMAS II)
- Digitales Wörterbuch der Deutschen Sprache (DWDS), Kernkorpora 20 und 21 (Berlin-Brandenburgische Akademie der Wissenschaften, dwds.de)

Korpora des Deutschen

Historische Korpora:

- Ahd.: Referenzkorpus Altdeutsch (REA)
- Mhd.: Referenzkorpus Mittelhochdeutsch (REM)
- Fnhd.: Bonner Frühneuhochdeutschkorpus (demnächst auch: Referenzkorpus Fnhd.)
- Älteres Nhd.: Deutsches Textarchiv (auch zugänglich über DWDS)

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 - Institut für Maschinelle Sprachverarbeitung, Stuttgart): http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

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

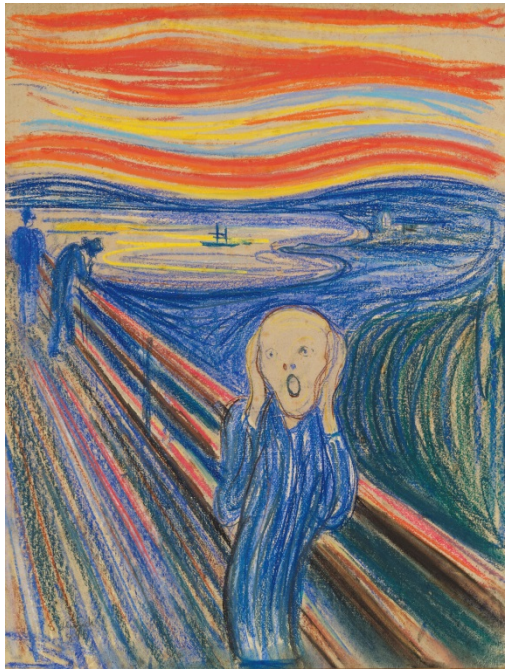


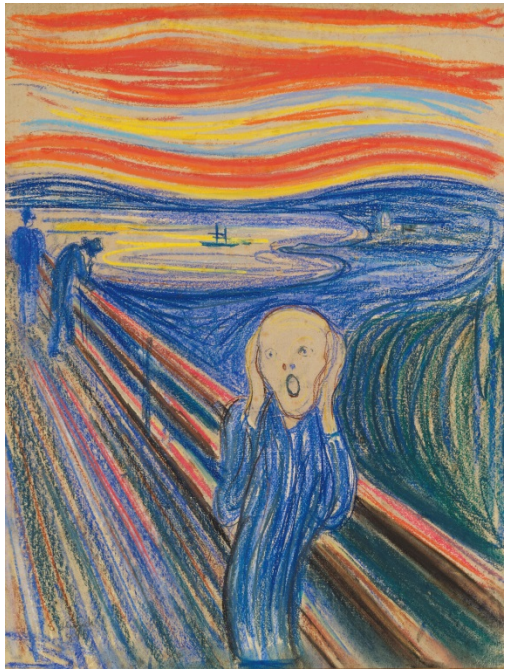



Types vs. Tokens

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)

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 ~~vs.~~ *und* 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

Von der Konkordanz zur Analyse

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

Literaturempfehlungen, Ressourcen, einschlägige Software

Literaturempfehlungen

- Scherer, Carmen. 2006. *Korpuslinguistik*. (Kurze Einführungen in Die Germanistische Linguistik 2). Heidelberg: Winter.
- Lemnitzer, Lothar & Heike Zinsmeister. 2015. *Korpuslinguistik. Eine Einführung*. 3rd ed. Tübingen: Narr.
- Stefanowitsch, Anatol. im Ersch. *Corpus linguistics. A guide to the methodology*. Berlin: Language Science Press.

Ressourcen und einschlägige Software

- für einfache Korpusrecherchen: AntConc
- für komplexere Korpusabfragen: CQP
- für Korpusannotation: GATE
- für Annotation und Auswertung von Konkordanzen: Tabellenkalkulationsprogramm (Excel, Calc)
- für alles mögliche: R und RStudio
- für Arbeit mit großen Datenmengen: Python und/oder Perl

Ressourcen und einschlägige Software

- Texteditor: Notepad++ (für Windows), für Mac z.B. TextWrangler

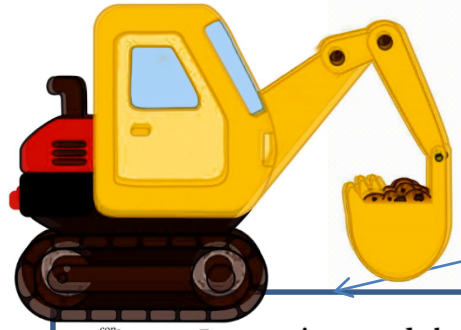
- zur Arbeit mit Konkordanzen: Spreadsheet-Programme wie Excel oder LibreOffice Calc

AntConc

- Programm für die Erstellung einfacher Konkordanzen aus Rohtextdateien
- laurenceanthony.net/software

Abfragesysteme und Abfragesyntax

Korpus und Abfragesystem



Abfragesystem

Korpusabfragesysteme

CWB
The IMS Open Corpus Workbench

ANNIS
A web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation.

NoSketch Engine

Sketch Engine

Korpusabfragesysteme

- Einige Abfragesysteme sind desktop-basiert, andere web-basiert
- Von den meisten desktop-basierten Systemen sind web-basierte Versionen verfügbar
- In die web-basierten Versionen lassen sich i.d.R. jedoch keine eigenen Korpora einspeisen.

Wie finde ich, was ich suche?

Reguläre Ausdrücke

- sind Ihnen ggf. aus Internet-Suchmaschinen bekannt
- am bekanntesten wohl: * als Platzhalterzeichen
- Reguläre Ausdrücke können aber viel mehr!

Die wichtigsten regulären Ausdrücke…

Gruppierung durch Klammern

- ( ) Runde Klammern definieren eine **Erfassungsgruppe** (*capturing group*)
- [ ] Eckige Klammern definieren eine **Zeichenklasse** (*character class*), z.B. [abc] = irgendein Zeichen aus dem Inventar a,b,c, [asdf] irgendein Zeichen aus dem Inventar a,s,d,f.
- [^] quasi das negative "Gegenstück" zu [ ]: irgendein Zeichen, das **nicht** in dem Inventar an Zeichen enthalten ist, das in den eckigen Klammern definiert wird, z.B. [^abc]: irgendein Zeichen, das nicht a, b oder c ist.

- (Wichtig: In anderen Kontexten bedeutet ^ etwas anderes!)

DWDS

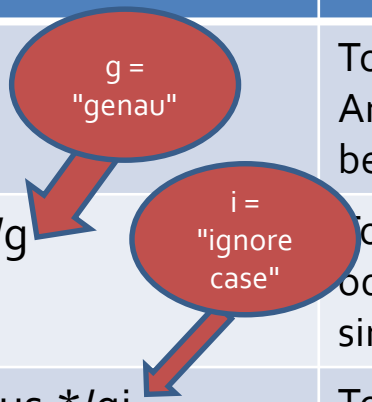
DWDS

- Digitales Wörterbuch der Deutschen Sprache
- zahlreiche Korpora über DWDS verfügbar, darunter Deutsches Textarchiv und DWDS-Kernkorpora des 20. und 21. Jahrhunderts

Suche im DWDS

- DWDS hat eine eigene Suchabfragesyntax (DDC), unterstützt aber (mit Einschränkungen) auch "normale" reguläre Ausdrücke

DWDS-Syntax	Suche mit "normalem" regulärem Ausdruck	findet
$p=N*	$p=/N.*/	Tokens, deren POS-Annotation mit *N* beginnt
Haus || Hof	$l=/Haus|Hof/g	Tokens, die als *Haus* oder *Hof* lemmatisiert sind
haus	$l=/.*haus|haus.*/gi	Tokens, deren Lemma mit *Haus* beginnt oder endet (*hausfraulich*, *Bauhaus*)



Weitere Infos: https://www.dwds.de/d/korpussuche

Beispiel: *-heit/-keit*

- Wir suchen im DWDS (www.dwds.de) nach Wortbildungsprodukten auf *-heit* und *-keit*.

- Bitte nutzen Sie die **Exportfunktion** und öffnen Sie die Konkordanz in einem Tabellenkalkulationsprogramm (z.B. Excel, Calc)

Beispiel: Nomina auf *-ung*

- Wir suchen im DWDS (www.dwds.de) nach Wortbildungsprodukten auf -ung.
- Worauf müssen wir achten?

Kleine Hausaufgabe

- Bitte suchen Sie nach im DWDS-Kernkorpus 21 nach Wörtern auf *-gate* (z.B. *Watergate, Dirndlgate, Dieselgate*).
- Exportieren Sie die Ergebnisse und versuchen Sie, eine Pivot-Tabelle mit den Ergebnissen zu erstellen.
- (Hinweise dazu hier: https://empirical-linguistics.github.io/korpus-schnelleinstieg/ - Kapitel 3.2)