Corpus linguistics

Materials

- https://t1p.de/cl_turin2024

Goals

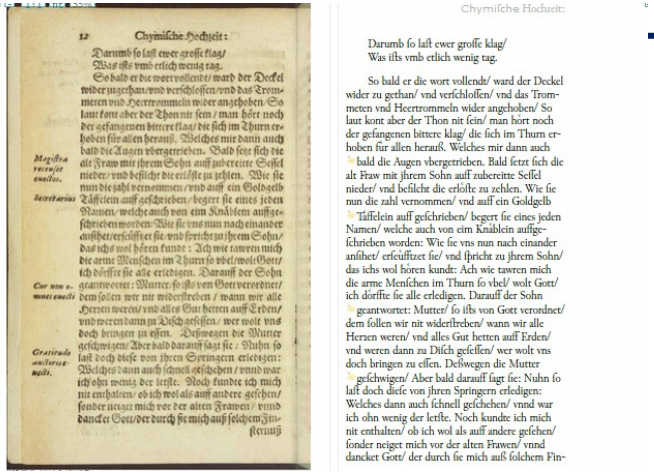
- What is corpus linguistics and why do we need it?
- learn how to deal with (German) corpora
- learn how to address research questions relating to language history from a corpus-linguistic point of view
- get to know "best practices" for corpus-linguistic studies

What is a corpus and how can we use it?

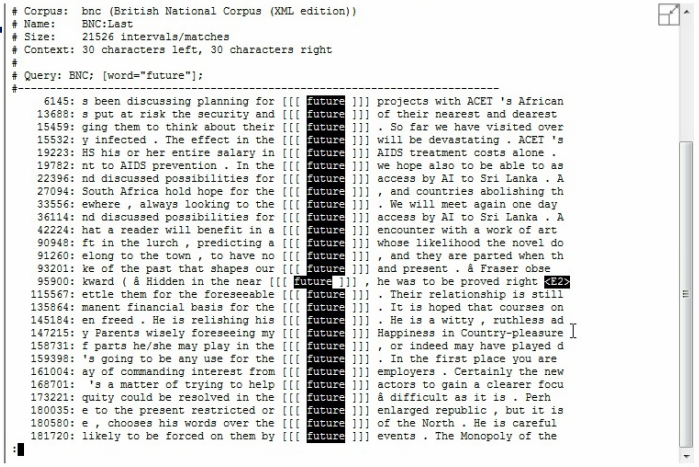
Time machines…

Time machines…

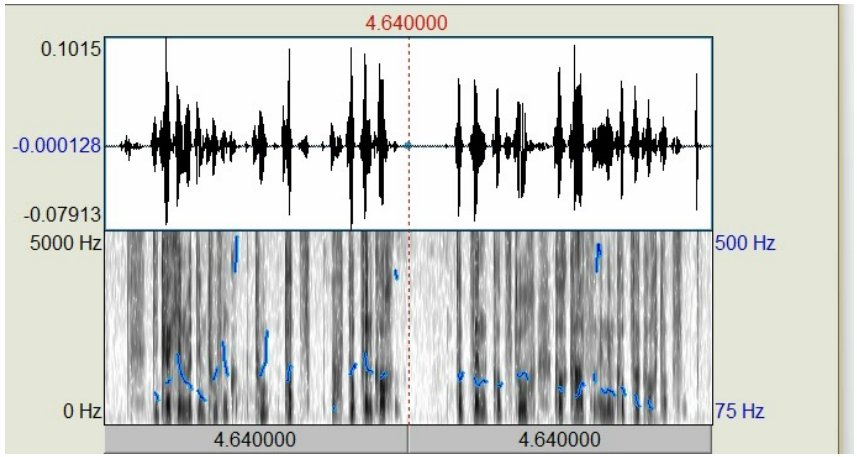
Time machines…



Deutsches Textarchiv



British National Corpus



Alcohol Language Corpus

TV News Archive

Corpora as time machines

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.

Corpora as time machines

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.

Corpora as time machines

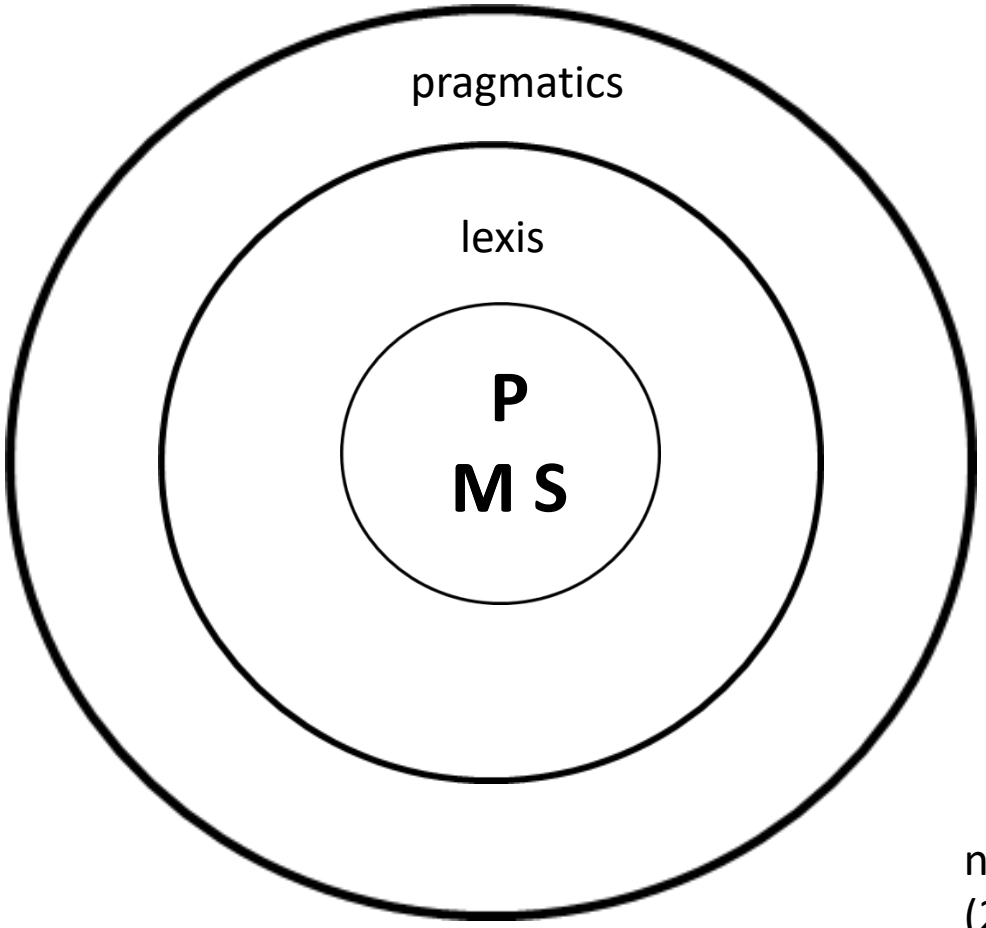
morphological change:
feuchtiger Natur > feuchterer Natur, keine Baert > keine Bärte

graphemic change:
mehr Saamens > mehr Samen; / (virgule)

syntactic change:
sind auch schnupffiger vnd fluessiger/ > auch … sind
Sagt auch … > Dies sagt auch

semantic change:
vnd darmit wird auch die Materi des Hirns ueberfluessiger
(*überflüssig* 'overflowing' > 'superfluous')

The "linguistic onion"



nach Nübling et al. (2006: 2f.)

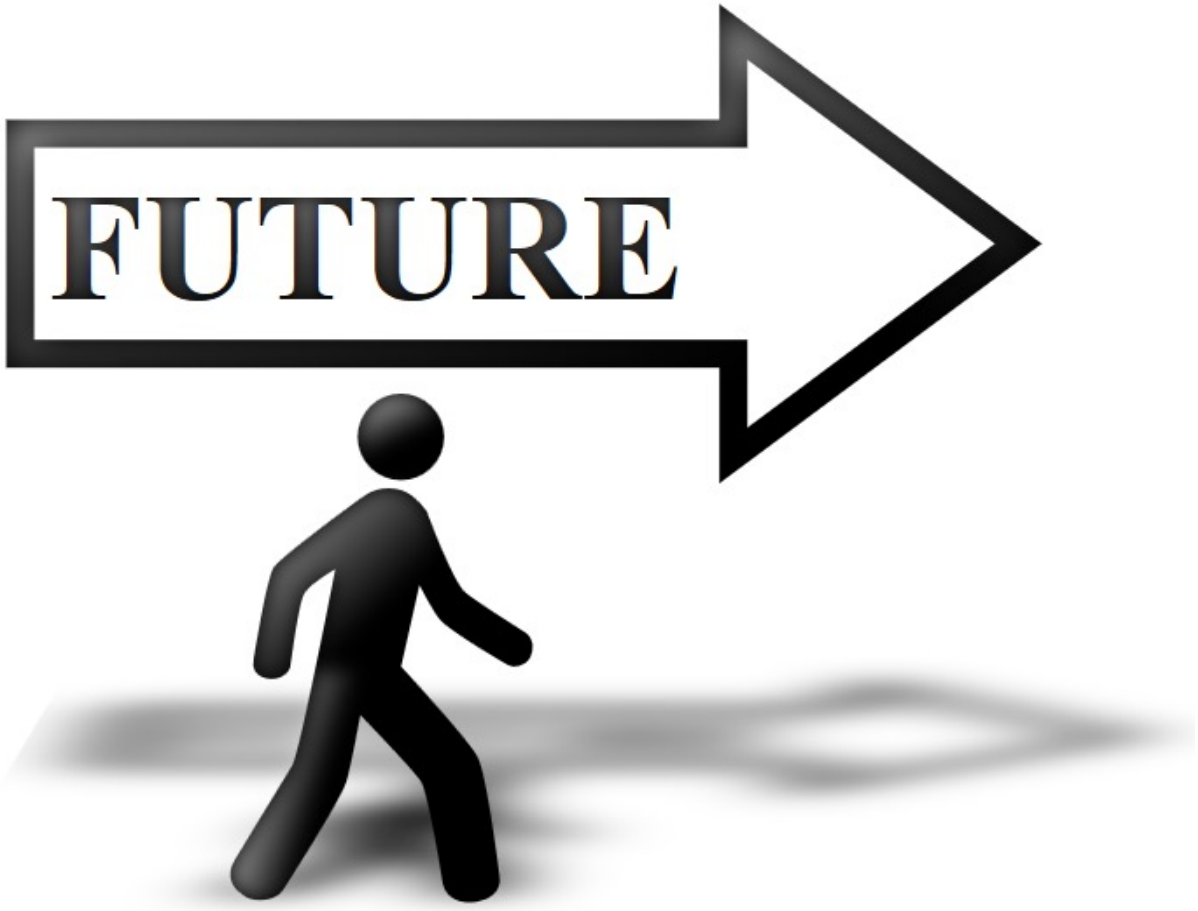
Corpora as time machines

- Corpora allow us to study language use in the past
- Arguably they also allow for some predictions about future developments

What we are going to do

In the next few weeks we want to…

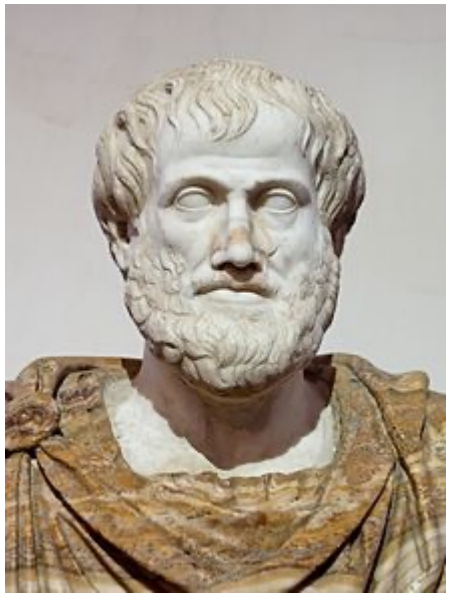
- learn how one can compile a good corpus
- **get to know existing corpora**
- learn how to analyze corpora
- perhaps get to know some very basic statistical methods


FUTURE

Why all this?
The scientific method and scientific theory

(based on Maxwell & Delaney 1999)

Aristotle



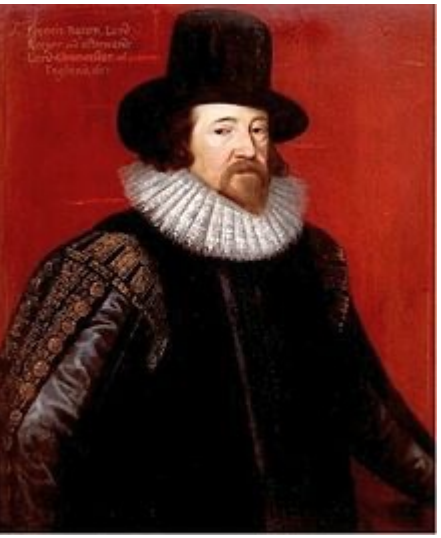
- **deductive** method
- Ideal: Syllogism

All humans are mortal.

All greeks are humans.

→ *All greeks are mortal*

Sir Francis Bacon



Novum Organon (1620):

- **Inductive** Method
- observe natural phenomena and draw general deductions
- methods: Experiments!
- Ideal: Researcher is objective and rational
- Explorative approach: Experiments are not hypothesis-driven

Pre-assumptions

- Science is never free of preassumptions
- The most important ones among them:
  - Uniformity of nature
  - Finite causality

Uniformity of nature

Uniformity of nature

- Nature follows certain regularities
- Hence, we can draw generalizations

Finite causality

- The causal chain leading to an effect is finite.
- Hence, the effect is **replicable.**

Positivism

- Predecessor: David Hume → inference of a causal relation between unobserved events is never justified.
- Comte: Positvism as ("ultimate") religion

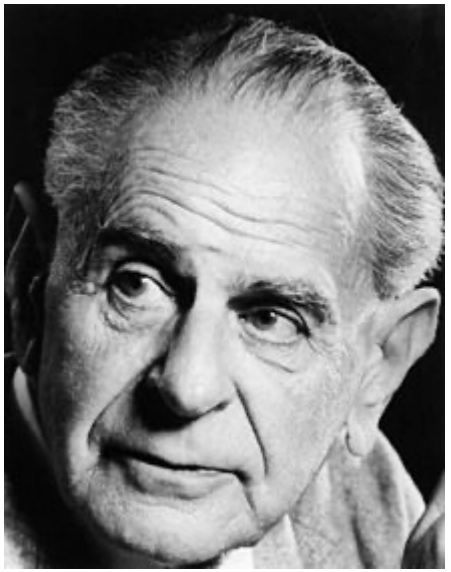
Auguste Comte

Logical positivism

- Vienna circle (e.g. Rudolf Carnap, Herbert Feigl)
- Symbolic logic as most important analytic tool
- verificationism: A proposition only makes sense when there is an empirical method to decide whether it is true or false.
- However: Not all scientific questions can be phrased as universally valid propositions.

Falsificationism

- Karl Popper: Scientific progress through **falsification** of theories
- Back to deduction, but on empirical grounds.



Syllogism of confirmation:
If my theory is true, my data will follow the predicted pattern.
My data do follow the predicted pattern.
Hence, my theory is true.

Syllogism of falsification:
If my theory is true, the data will follow the predicted pattern.
My data do *not* follow the predicted pattern.
Hence, my theory is wrong.

Occam's razor

- "*Entia non sunt multiplicanda praeter necessitatem*" (Johannes Clauberg, 17th century)
- named after William of Occam (13th century.)

Relation to corpus linguistics

- In principle, each corpus study is an experiment.
- Especially **quantitative** corpus linguistics is all about hypothesis testing.
- It usually follows a falsificationist approach:
  - I formulate a hypothesis…
  - …but I test the **null hypothesis**

→ null hypothesis significance testing (NHST)

But…

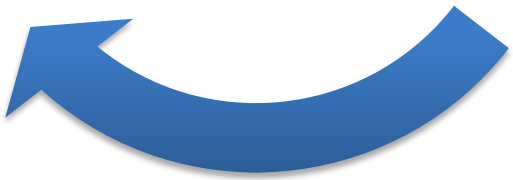
- Not all (corpus-)linguistic methods are falsificationist.
- Many corpus-linguistic studies follow an **explorative** approach: finding patterns in the data.
- Also, the Bayesian approaches that have become popular in recent years are usually not falsificationist (but instead try to quantify uncertainty).

Deductive vs. inductive method

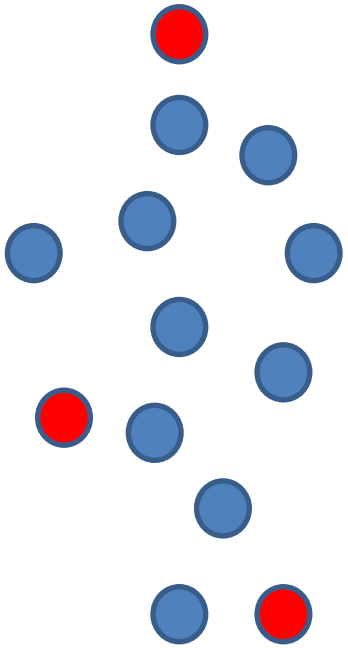
Operationalization

Hypothesis

Test

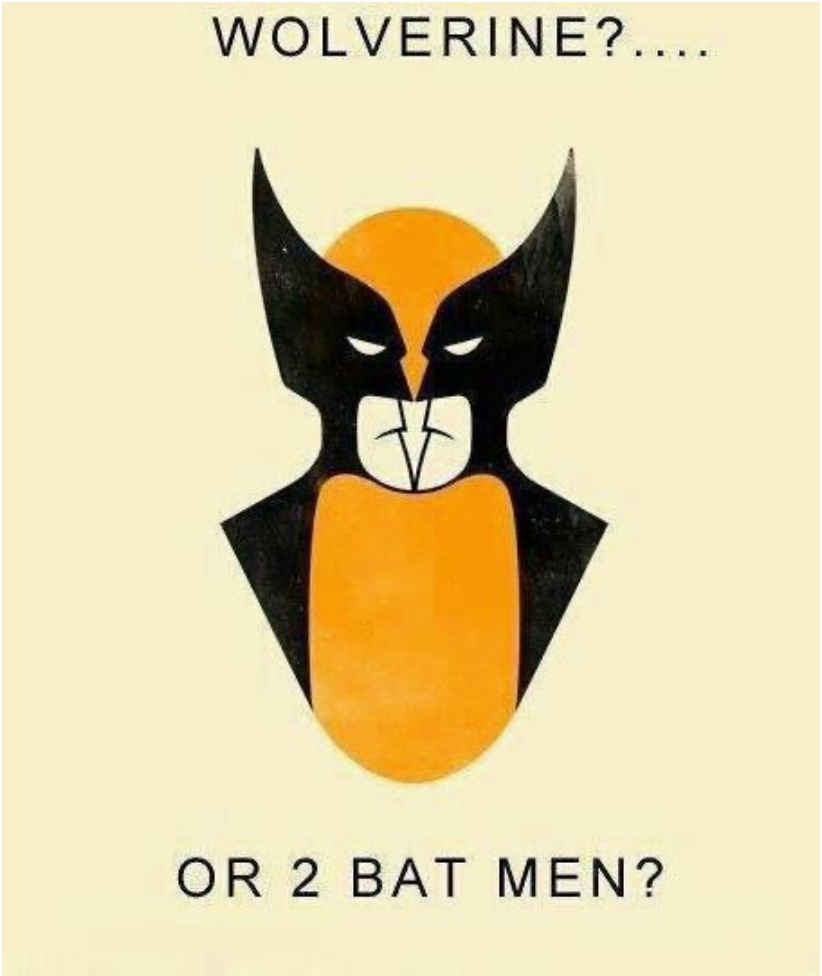


New / modified / additional hypotheses

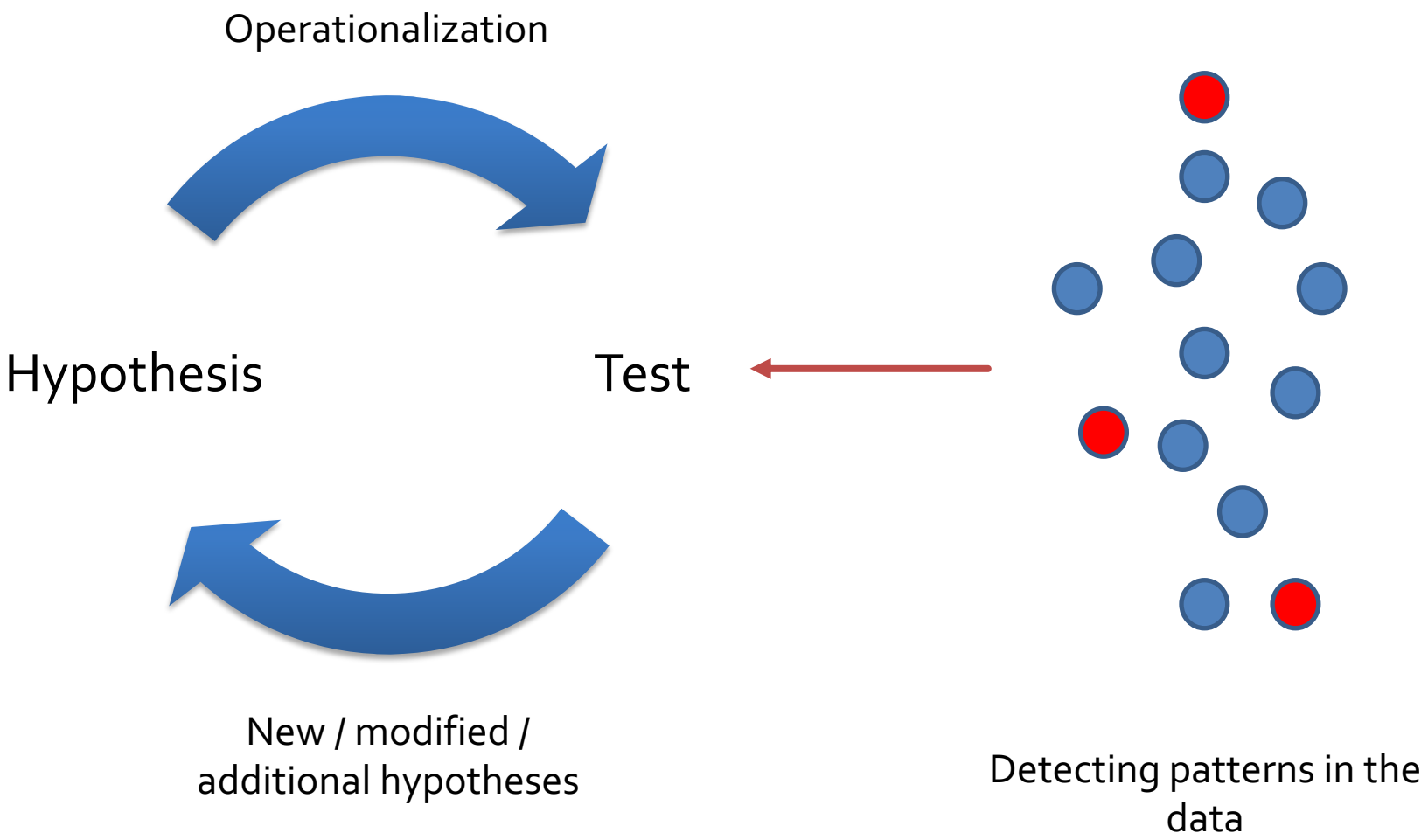


Detecting patterns in the data

(based on Stefanowitsch 2017)


WOLVERINE?....
OR 2 BAT MEN?

Deductive vs. inductive method



(based on Stefanowitsch 2017)

Why corpus linguistics?

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

Why?

I speak English (German, Italian, Russian, Persian….) – why do I need corpus linguistics?!?

beizeiten, mitunter

from the corpus of Wikipedia discussions (available via 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

from the corpus of Wikipedia discussions (available via 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

from the corpus of Wikipedia discussions (available via 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

web attestations:

- [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)

Can we trust our intuitions?

- Intuition is a necessary first step…
- …but it can only be the beginning!
- …or just the beginning?!?

```
BNC> A = "only|just" "the" "beginning"
BNC> count A by hw on match[0]
63        only  [#22-#84]
22        just  [#0-#21]
```

How to apply corpus linguistics

- doubtful cases
- historical change
- langage variation, dialectology
- graphemic variation
- multimodality
- phonetics
- .....

What is corpus linguistics?

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)

What is corpus linguistics?

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)

the definition follows from the prerequisites of a scientific questions when taking a falsificationist approach – valid questions e.g.:

- The genitive appears **more often** in older than in newer texts.
- Older speakers use loan words **less often** than younger ones.
- Women use **more discourse particles** than men.
- The term *parkieren* is **only** used in Swiss German.

- but **not**: Anglicisms are used **often** in German.

What is corpus linguistics?

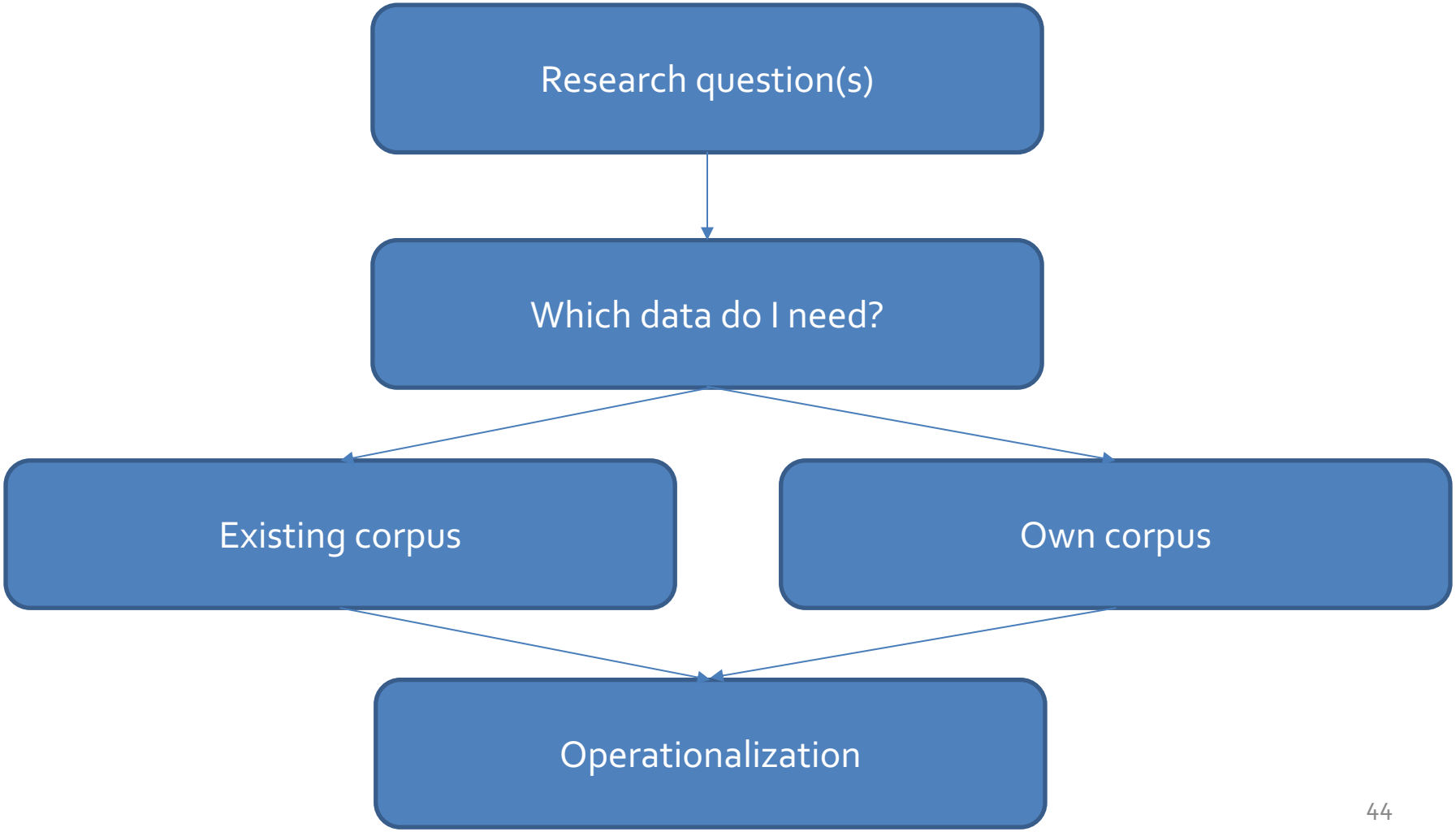
Corpus-based vs. corpus-illustrated approaches

- "corpus-illustrated" approaches are qualitative, but use selectively chosen attestions
  corpus-based approaches can be purely quantitative, but usually they combine quantitative and qualitative approaches (Lemnitzer & Zinsmeister 2015)

What is corpus linguistics?

- Purely quantitative approaches rely only on corpus data (e.g. n-grams, distributional semantics…)
- Quantitative-qualitative approaches rely on analysis and interpretation of the individual attestations (annotation!)

Corpus-linguistic workflow



Corpus design

Task:

A scientist from Mars asks you to build a corpus that represents as accurately as possible how people in Turin speak. How would you do that?

Corpus design

- Representativity
- Balance
- Size
- Adequacy for the individual research question

in the case of transcribed/transliterated data:

- Quality of transcription/transliteration

Balance and representativity

- " […] 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

Corpus design

Fundamental questions:

- What exactly do I want to investigate?
- What type of data do I need for that?
- Does such a corpus already exist?
- In what respect does the corpus have to be particularly accurate?
  - e.g. in the case of graphemic investigations: represent graphemic variants of the original as closely as possible

Corpus design

If I compile my own corpus:

- How can I obtain data?
- Are there any concerns regarding copyright?
- Are there any moral or ethical concerns?

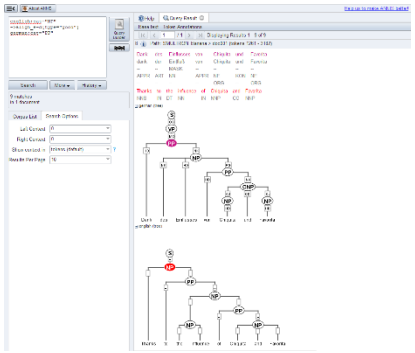
Creating a corpus

- Collecting and preprocessing data
- Tokenization
- Lemmatization and POS-tagging
- further annotation

Creating a corpus

Text

Preprocessing, Annotation



(idelly): make available publicly

Tools:

- automatic taggers
- Software for manual annotation

Let's build a corpus!

NEWSTICKER

Mittwoch, 15. August 2012

Neue Zeitform Futur III eingeführt, um Gespräche über Flughafen BER zu ermöglichen

Berlin (dpo) - "Ich werde nächstes Jahr im Sommer nach Mallorca in den Urlaub

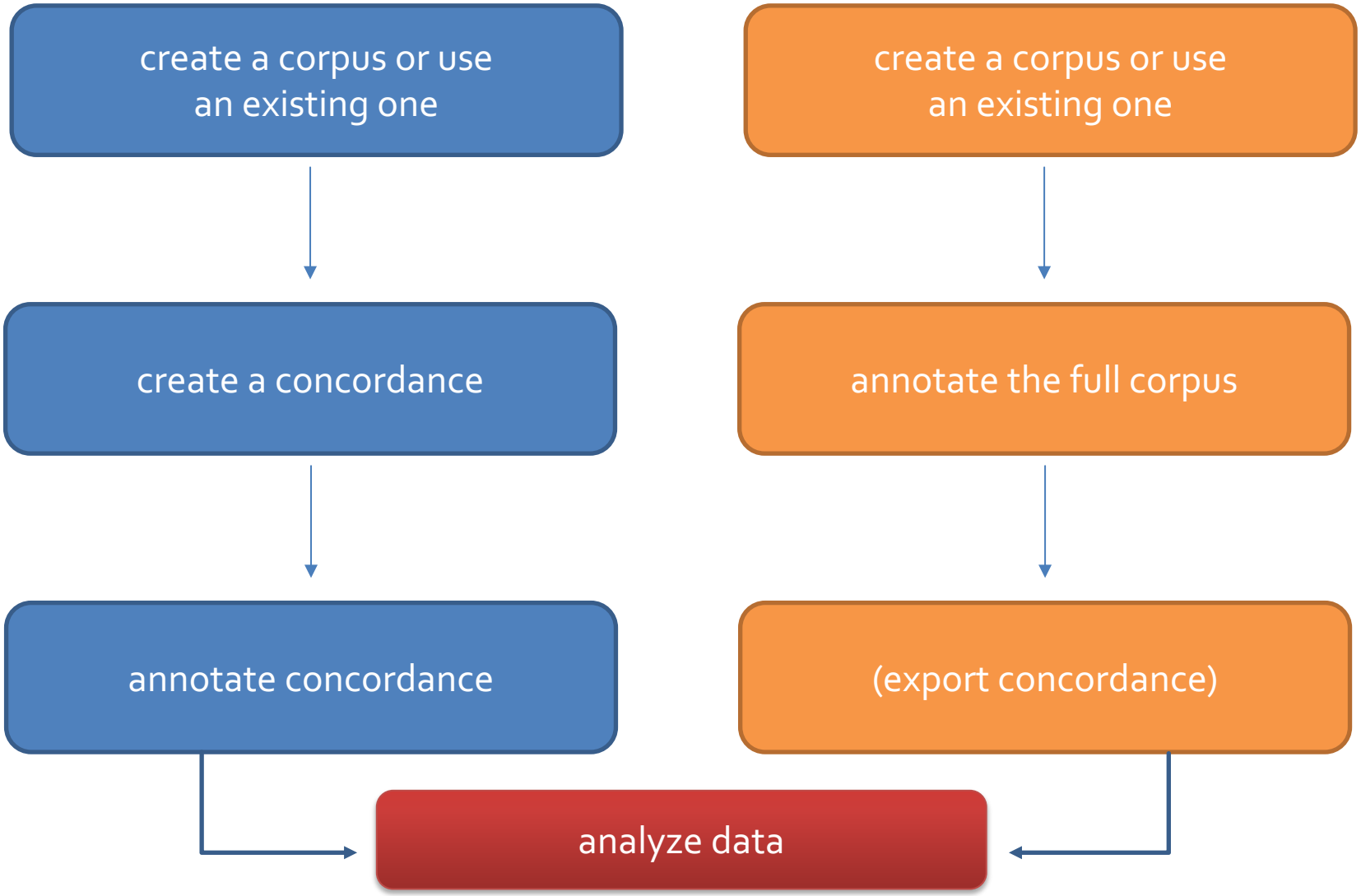
AUS DEM ARCHIV

Kampf gegen illegale Einwanderung: Trump will Indianer nach Indien abschieben

REKLAME

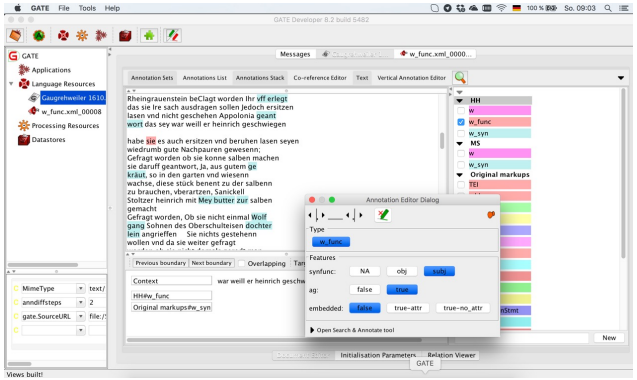
Nur für Postillon-Minus-Abonnenten.

Possible workflows



Annotation: Two ways

Annotation
directly in the corpus



(Export)

Export

Annotation
in concordance files

URL	Left	Key	Right	Lemma	POS	quasispont	Tokens
http://www.alemannRande der Stadt , nicht zugänglich für die Öffentlichkeit und eingeschlossen von einer hohen , efeuberankten			Mauer , ist der Grund dafür , dass der Friedhof fast vergessen	(unknown)	ADJA	2	
http://www.lleinthes	Ein bärtiger Flaum - wahlweise auch zum Schnurrbart verkommen - unter den		Augen , dazu noch ein Hütchen zwecks Beschattung der Denke	bebrillt	ADJA	3	
http://spirituelle-ded			In welchem Teil sind wir also jetzt	beheimaten	VVPP	4	
http://www.wasmita	lich sind in Deutschland mehrere führende Unternehmen aus dem Bereich der Agrartechnik		beheimatet	, die ihre Maschinen in aller Welt verkaufen .	beheimaten	VVPP	3
http://www.wissen-i			Das Bikini-Atoll und das	benachbarte	Eniwetok-Atoll im Pazifik wurden als Testgebiete gewählt , we	benachbart	ADJA	1
http://www.waldorfs			Den Namen hat das Haus von dem	benachbarten	Naturschutzgebiet " Methorst-Teich " , einem Wa	benachbart	ADJA	3
http://www.awo-owl			Ein so genannter " Zwergenpfad " führt von der	benachbarten	Integrativen Kindertagesstätte zum Spielgelände .	benachbart	ADJA	3
http://www.wbs-law.			Beiden	unbemannten	Luftfahrzeugen gleich ist , dass der Betrieb innerhalb der Sicht	unbemannt	ADJA	1
http://deutschland-irrowohl der Name an die hawaiianische Bezeichnung für " Wal " erinnert , steht die				unbemannte	Station dort , wo es am Kältesten auf der Welt ist .	unbemannt	ADJA	4
http://www.vogelwac	, und da zu allem Überfluss der Konzertsaal lediglich drei Gehminuten von meiner Wohnung			beheimatet	ist , wohnte ich doch ganz spontan mal jener Aufführung bei .	beheimaten	VVPP	1
http://www.bang-yo			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.welpen.cer	, sondern versteckt sich hinter jedem seiner Stoppelhaare , in seinem Bart , seinen stopplig			behaarten	Kopf und seinen stets lustigen kohlrabenschwarzen Augen , di	behaart	ADJA	1
http://www.nexgam.	Entwickler durchaus neue Gameplay-Kniffe abseits des Genre-Standards einfallen lassen : Ein			beherzter	Druck blendet mehrere " Gates " im Head-Up-Disp	beherzt	ADJA	304
http://www.deutsche			Auch in	benachbarten	Politikfeldern ist Brüssel aktiv .	benachbart	ADJA	2
http://www.bruderra			So richtig ins Schwitzen kamen die mehr oder weniger	beleibten	Marschierer bei der Parade .	beleibt	ADJA	3
http://www.gilmoreg			Auf der leicht	bemehlten	Arbeitsfläche den Teig 1 cm dick rechteckig ausrollen .	bemehlt	ADJA	697
http://www.tv-jahn-loedeutet mir sehr viel " , sagt der Springer , der seit fast zehn Jahren in der Deisterstadt				beheimateten	ist .	beheimaten	VVPP	5
http://steine.helga-in			Im 15. Jahrhundert errichteten die Fugger eine Bleihütte (Fuggerau) 1797 entstand im	benachbarten	Gailitz eine Glättefabrik , die später Rostschutzfarbe und von 1	benachbart	ADJA	2
http://eco.psy.ruhr-u - 27 (links unten) symbolisiert ein Punkt den Mittelwert der Helligkeitsunterschiede von vier				benachbarten	Bildpunkten .	benachbart	ADJA	201
http://www.tabularas			Im Kosovo und dem	benachbarten	Bosnien gärt und brodelt es längst , und auch das instabile Ma	benachbart	ADJA	1
http://www.eisenbahˈuot; Vitosha Express " vermarkteten Touristik-Sonderzuges handelt , welcher in Plovdiv				beheimatet	ist .	beheimaten	VVPP	2
http://www.sg-lobba			Vom Coach des Gegners erhielten unsere Jungs mächtig Lob für ihr	beherzten	und jederzeit faires Spiel .	beherzt	ADJA	225
http://www.maratho			In einer Sonderfahrt für	bemützte	Sonderlinge , schlabbe 50 Minutos .	(unknown)	ADJA	807
http://www.leidartik	So ist ja auch der (langweiligere) Teil der Fachhochschule Bonn-Rhein-Sieg in Rheinbach			beheimatet	.	beheimaten	VVPP	206
http://hall-of-fi3 30 000 Flüchtlinge in das nicht mehr genutzte Sportstadion « Astrodome » nach Houston im				benachbarten	Bundesstaat Texas evakuiert werden .	benachbart	ADJA	793
http://www.wocamp			In Brandenburg , Berlin und auch nördlich davon ist eher ein	beherztes	Wort gefragt .	beherzt	ADJA	397
http://www.keller-fa			Kleine	bewaldete	180 m hohe Erhebung über dem Katzbachtal an der Riemberg	bewaldet	ADJA	2

Annotation in corpora or concordances?

- **Corpus annotations**: The annotations are added to the data in the corpus itself.
- **Annotation of concordances**: We first export concordances, i.e. lists of attestations, and then add annotations

Workflow for concordance-based analyses

Research question

corpus query

concordance

Existing corpora

- Which corpora do you know already?

Basic notions of corpus linguistics

POS-Tagging & Lemmatization

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/

- often automatically, e.g. with TreeTagger or Spacy
- Pro: extremely fast and efficient
- Con: often error-prone
- we do not always have existing taggers for historical data
- however, there are some options for historical German:
  - TreeTagger parameter file for Middle High German
  - CAB (DWDS/DTA)

Tagsets

- different Tagsets for POS
- very common for German data: Stuttgart-Tübingen Tagset (STTS)
- for historical German data, there's an adaption called HiTS
- Overview of STTS: http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

Types and Tokens

Word	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



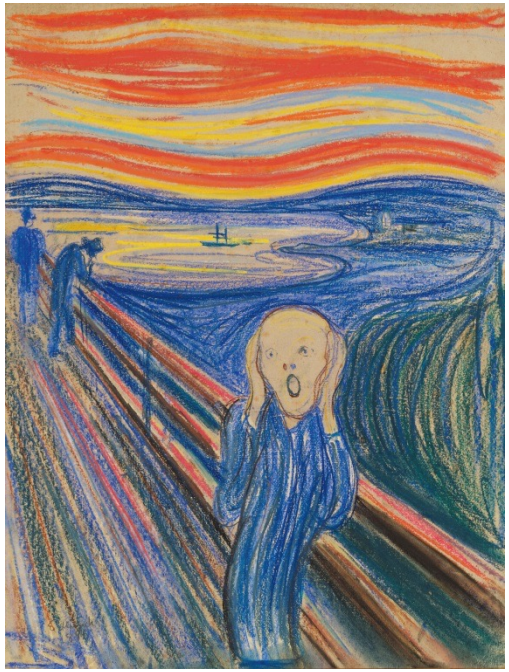


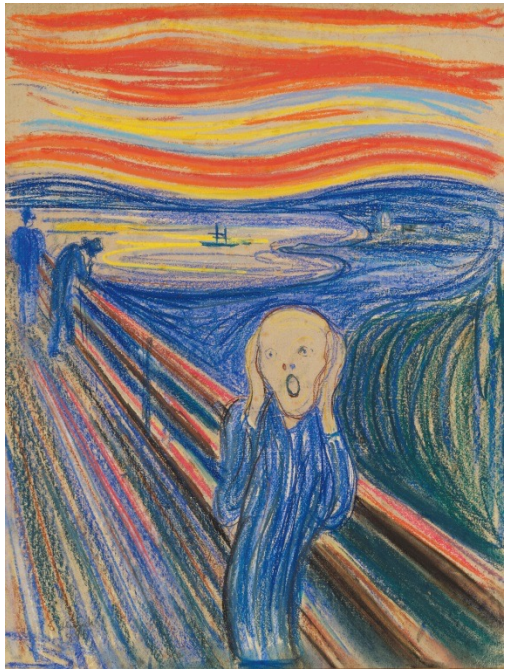




Types vs. Tokens

Types vs. Tokens





Types vs. Tokens






How many types?

It depends….

Types and Tokens

Wenn Fliegen neben Fliegen fliegen, fliegen Fliegen neben Fliegen.

Lemma	**Tokens**
Fliege	4
fliegen	2
wenn	1
neben	2

Quantitative and qualitative methods

Corpus analysis

qualitative analysis:

- Observations based on individual attestations
- can relate to any aspect from semantics via morphology to syntax
- ideal for semantic and pragmatic analyses

Corpus analysis

quantitative analysis:

- taking into account many attestations at once instead of individual ones
- quantification via e.g.
  - counting words, parts-of-speech, grammatical patterns etc.
  - statistical methods (e.g. collocation measures)

Corpus analysis

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.

How can we investgate this text?

Qualitative vs. quantitative

Which advantages and disadvantages do quantitative and qualitative approaches have?

Which approach would you choose for investigating…

1. …changes in the position of the genitive (*des Vaters Haus* > *das Haus des Vaters*)
2. Racism in letters to the editor
3. Semantic change of *geil*

Qualitative ~~vs.~~ *and* quantitative

- Most corpus-linguistic approaches are both qualitative and quantitative
- operationalization of individual variables (e.g. semantic ones) usually requires qualitative interpretation of individual attestations
- Example: Annotation of animacy!

Qualitative ~~vs.~~ *and* quantitative

- Even syntactic annotation often requires an in-depth qualitative interpretation of the data

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)

From concordance to analysis

- Operationalizing hypotheses
- → clear annotation guidelines!

Literature, software, resources

Literature recommendations

- Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R: introduction to quantitative methods in linguistics*. New York, NY: Springer.
- Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus linguistics and linguistically annotated corpora*. London: Bloomsbury.
- Stefanowitsch, Anatol. 2020. *Corpus linguistics. A guide to the methodology*. Berlin: Language Science Press.

Resources & software

- for simple corpus analyses: AntCont
- for working with raw texts: Notepad++ (for Windows), BBEdit (for Mac); alternatively e.g. VS Code
- for more complex querying: CWB
- for corpus annotation: e.g. GATE, INCEpTION, Hexatomic
- for annotating and analyzing concordances: spreadsheet software (Excel, Calc)
- for basically everything: R/RStudio and/or Python

AntConc

- laurenceanthony.net/software

Resources & software

- for converting export files in KWIC format:

  github.com/hartmast/concordances

- Some tutorials on

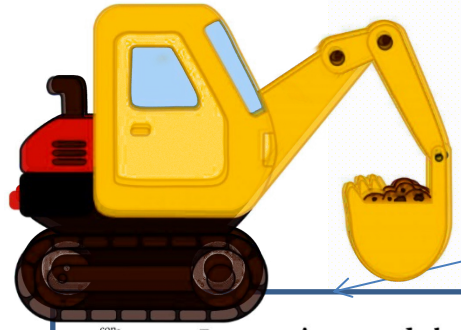
  https://empirical-linguistics.github.io/

R

- Stats software and programming language
- useful IDE: RStudio

Corpus query systems and query syntax

Corpus and query system



Abfragesystem

Corpus query systems

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

Anatomy of a corpus

Halo i bims ein Beispiel-Korpustext

Ich bin ein Korpustext. Jedes Korpus beginnt mit Texten, und manche enden hier auch: Einige Korpora bestehen lediglich aus Rohtexten. Andere Korpora sind mit zusätzlichen Informationen angereichert, sogenannten Annotationen. So sind in vielen Korpora die Wortarten annotiert. Auch sind viele Korpora auf ihre Grundform (Lemma) hin getaggt.

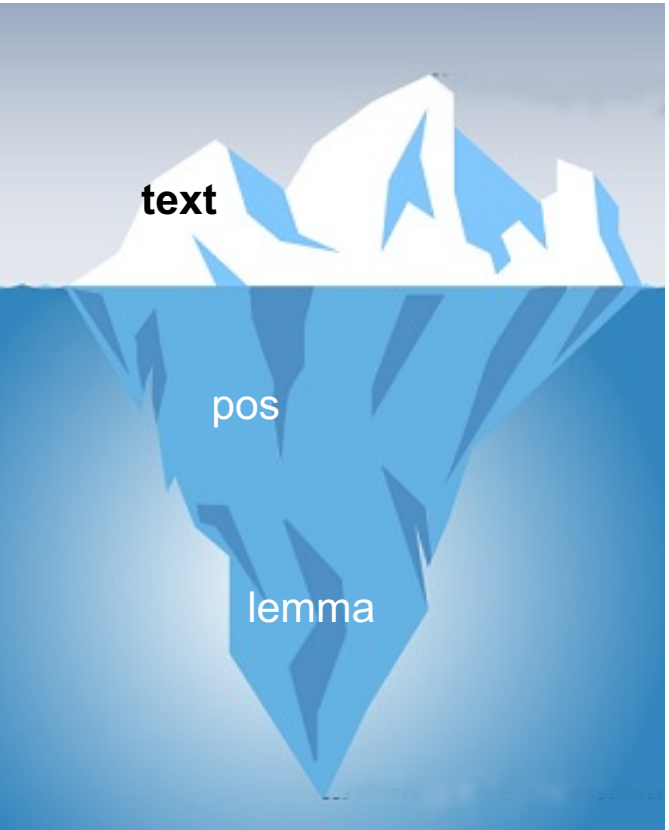
Anatomy of a corpus

<header>Halo/NN/<unknown> i/FM/i
bims/VVIMP/bimsen ein/ART/eine Beispiel-
Korpustext/NN/<unknown> </header>

Ich/PPER/ich bin/VAFIN/sein ein/ART/eine
Korpustext/NN/<unknown> ./$./. Jedes/PIAT/jede
Korpus/NN/Korpus beginnt/VVFIN/beginnen
mit/APPR/mit Texten/NN/Text|Texten ,/$,/,
und/KON/und manche/PIS/manche
enden/VVFIN/enden hier/ADV/hier auch/ADV/auch

Corpus and query system

- In some corpus query systems, one can search on all levels but only gets the text on the main level as a result...



Halo	i	bims	ein	Korpustext.
ITJ	PPER	VVFIN	ART	NN
hallo	ich	bins	ein	Korpustext.

Corpus query systems

- standalone or web-based
- most standalone system also have web-based counterparts
- in the web-based versions you often cannot upload your own corpora (exception e.g. some instances of CQPweb)

Query syntax

- Corpus query systems often have their own **syntax**.
- Corpus query syntax is a bit like learning the syntax of foreign languages:

- ➢ You can use a foreign language even if you have only mastered the basics of its syntax…
- ➢ … but to make full use of its expressive power, you need a good knowledge of its syntax.

Query syntax

Which options do we need when searching a corpus?

Example 1:

We search for all attestations of the verb *legen* 'lie' (as in 'lie in bed') in a corpus that is neither POS-tagged nor lemmatized.

Query syntax

Which options do we need when searching a corpus?

Example 2:

We search for all attestations of the verbs *setzen*, *stellen*, *legen* (all roughly 'put' in a **lemmatized** corpus)

Query syntax

Which options do we need when searching a corpus?

Example 3:

We search attestations for the pattern *je* X-*er* *desto* Y-*er* 'the x-er the y-er' in a POS-tagged corpus.

Query syntax

Which options do we need when searching a corpus?

Example 4:

We search attestations for *weil* + noun and *weil* + adjective in a pos-tagged corpus:

- *Ich kann heute nicht ins Kino, weil Seminar.*
- *Ich will heute nicht ins Kino, weil müde.*
- but not: *…weil Seminar ist.*

Query syntax

Which options do we need when searching a corpus?

Example 5:

We search queries for V-*en gehen* in a tagged corpus.

- *Ich gehe heute schwimmen.* 'I go swimming today'
- *Ich will heute schwimmen gehen.* 'I want to go swimming today'
- *Ich gehe heute mit meinem Freund schwimmen.* 'I go swimming with my (boy)friend today'

So what do we need?

Logical operators

- AND
- OR
- NOT

Wildcards

- leg*, setz*, stell*

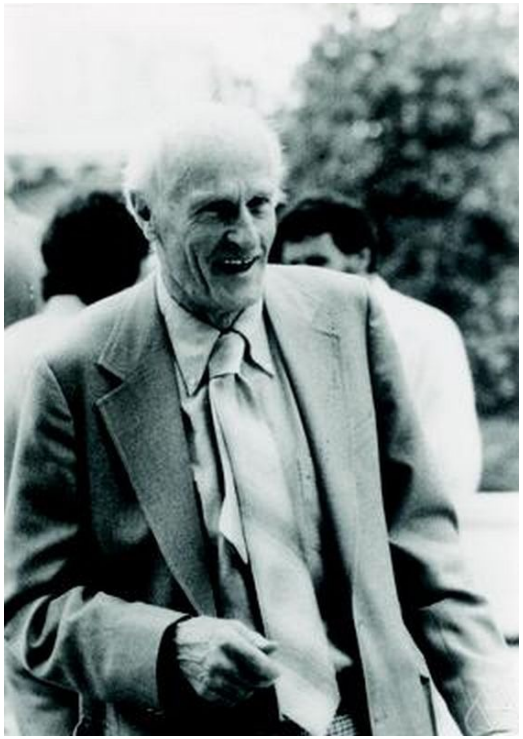
Distance operators

- Ich gehe {0-5} schwimmen

How can I find what I'm looking for?

Regular expressions

- sequence definining a search pattern
- comes from mathematics and computer science



Stephen Kleene

The most important regular expressions

Grouping

- ( ) ***capturing group,*** e.g. *urg(ing|ent)*
- [ ] ***character class***, e.g. [abc] = any one character out of the set a,b,c; [asdf] any one charater out of the set a,s,d,f.
- [^] basically the negative "counterpart" to []: any character not included in the set of characters defined in the square brackets, e.g. [^abc]: any one character that is **not** a, b, or c.
- (Note that in other contexts, ^ has a different meaning!)

The most important regular expressions

Wildcards and repetition operators

- . any character
- ? the immediately preceding character occurs 0 times or once
- * the immediately preceding character occurs 0 to any number of times (in direct sequence)
- + the immediately preceding character occurs 1 to any number of times (in direct sequnce)
- {n} the immediately preceding character occurs n times (in direct sequnce)
- {x,} the immediately preceding character occurs at least x times (in direct sequnce)
- {x,y} the immediately preceding character occurs between x and y times (in direct sequnce)

The most important regular expressions

More operators

- | OR operator
- \ escape string, e.g. for finding "real" question marks
- ^ start position
- $ end position

Using regular expressions

- Some differences in "dialects" of regular expressions and their implementation in query systems
- Hence: Like real languages, corpus query languages have to be learned to some extent…

Next level

Bracket expressions

- [:alnum:] alphanumeric (a, b, 1, 2)
- [:alpha:] alphabetic (a, b, c, nicht 1, 2)
- [:digit:] digita (1, 2, 3, … nicht a, b, c)
- [:blank:] whitespace, tabstop
- [:punct:] punctuation

Lookaround / Lookahead

(?=foo)	Lookahead	The string immediately following the position we search for is *foo*
(?<=foo)	Lookbehind	The string immediately preceding the position we search for is *foo*
(?!foo)	Negative Lookahead	The string immediately following the position we search for is not *foo*
(?<!foo)	Negative Lookbehind	The string immediately preceding the position we search for is not *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	The string immediately following the position we search for is *foo*
(?<=foo)	Lookbehind	The string immediately preceding the position we search for is *foo*
(?!foo)	Negative Lookahead	The string immediately following the position we search for is not *foo*
(?<!foo)	Negative Lookbehind	The string immediately preceding the position we search for is not *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	The string immediately following the position we search for is *foo*
(?<=foo)	Lookbehind	The string immediately preceding the position we search for is *foo*
(?!foo)	Negative Lookahead	The string immediately following the position we search for is not *foo*
(?<!foo)	Negative Lookbehind	The string immediately preceding the position we search for is not *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	The string immediately following the position we search for is *foo*
(?<=foo)	Lookbehind	The string immediately preceding the position we search for is *foo*
(?!foo)	Negative Lookahead	The string immediately following the position we search for is not *foo*
(?<!foo)	Negative Lookbehind	The string immediately preceding the position we search for is not *foo*

Hands-on example

- Please write in a text editor:

```
Dies ist ein Test, der testen soll, wie
Lookahead und Lookbehind funktionieren.
```

- Use the search function and lookahead to insert the word *schöner* before *Test*, so that the end result is:
- `Dies ist ein Test, der testen soll, wie Lookahead und Lookbehind funktionieren.`

- Use lookahead to insert a tabstop (`\t`) before each punctuation mark (`[[:punct:]]`).

Using regular expressions

- Let's try to use htpps://www.dwds.de/

  We search for

  – all compounds that end with *-papst*
  – …all verbs with the prefix *be-*
  – all proper nouns (NE) and common nouns (NN)
  – the construction *je* X-*er desto* Y-*er*
  – variants of *je* X-*er desto* Y-*er* as in e.g. *je* X-*er umso* Y-*er* oder auch *je* X-*er je* Y-*er, umso* Y-*er umso* Y-*er*

Using regular expressions

We search for…

– fronted particle verbs, e.g. *an hat sie das Licht gemacht, nicht aus*

– genitives in pre- and postposition

Encoding Hell

Encöding h≈ll

- Use Unicode
- Use Unicode
- Use Unicode

(Quelle: http://pt.slideshare.net/MapRTechnologies/data-breaking-bad/19?smtNoRedir=1)

Encoding Hell

- Windows uses Windows-1252 by deafult (similar to ISO/IEC 8859-1 / Latin-1 / ASCII)
- Unix and Linux use UTF-8 by default (Unicode-based)
- Most German-language historical corpora are coded in UTF-8 (because of special characters)
- Windows can handle UTF-8 but sometimes only does so relucantly…
- When working with Windows (and Microsoft apps in general) always double-check that no special characters get lost!

Behind the scenes of a corpus

Why looking behind the scenes?

- Corpus files often look harrowingly complex…
- … but they are highly structured!
- This has many advantages when doing queries that go beyond the capacities of specific corpus query languages.

Why looking behind the scenes?

- Example: The reference corpus of Middle High German is annotated for sentence boundaries…
- … but currently it's not possible to tell ANNIS to search within a specific sentence.
- But if we search the corpus with our own scripts, this is no problem.

Example: DWDS/DTA

```
<TextCorpus xmlns="http://www.dspin.de/data/textcorpus" lang="de">
  <tokens>
    <token ID="w1">Herrn</token>
    <token ID="w2">Hannß</token>
    <token ID="w3">Aßmanns</token>
    <token ID="w4">Freyherrn</token>
    <token ID="w5">von</token>
    <token ID="w6">Ab&#x017F;chatz</token>
    <token ID="w7">/</token>
    <token ID="w8">Weyl</token>
    <token ID="w9">.</token>
    <token ID="wa">gewe&#x017F;enen</token>
    <token ID="wb">Landes-Be&#x017F;tellten</token>
    <token ID="wc">im</token>
    <token ID="wd">Fu&#x0364;r&#x017F;tenthum</token>
    <token ID="we">Lignitz</token>
    <token ID="wf">/</token>
    <token ID="w10">und</token>
    <token ID="w11">bey</token>
    <token ID="w12">den</token>
    <token ID="w13">Publ</token>
    <token ID="w14">.</token>
    <token ID="w15">Conventibus</token>
    <token ID="w16">in</token>
    <token ID="w17">Breßlau</token>
    <token ID="w18">Hochan&#x017F;ehnl</token>
    <token ID="w19">.</token>
```

Example: REM

```
<token id="t12" trans="in||handon(.)" type="token">
  <tok_dipl id="t12_d1" trans="inhandon" utf="inhandon"/>
  <tok_anno ascii="in" id="t12_m1" trans="in|" utf="in">
    <norm tag="in"/>
    <token_type tag="MS1"/>
    <lemma tag="in"/>
    <lemma_gen tag="in"/>
    <lemma_idmwb tag="81741000"/>
    <pos tag="APPR"/>
    <pos_gen tag="AP"/>
    <infl tag="c.D"/>
    <inflClass tag="--"/>
    <inflClass_gen tag="--"/>
  </tok_anno>
  <tok_anno ascii="handon" id="t12_m2" trans="|handon" utf="handon">
    <norm tag="handen"/>
    <token_type tag="MS2"/>
    <lemma tag="hant"/>
    <lemma_gen tag="hant"/>
    <lemma_idmwb tag="68277000"/>
    <pos tag="NA"/>
    <pos_gen tag="NA"/>
    <infl tag="Dat.Pl"/>
    <inflClass tag="st(u).Fem"/>
    <inflClass_gen tag="st(u).Fem"/>
    <punc tag="DE"/>
  </tok_anno>
</token>
```

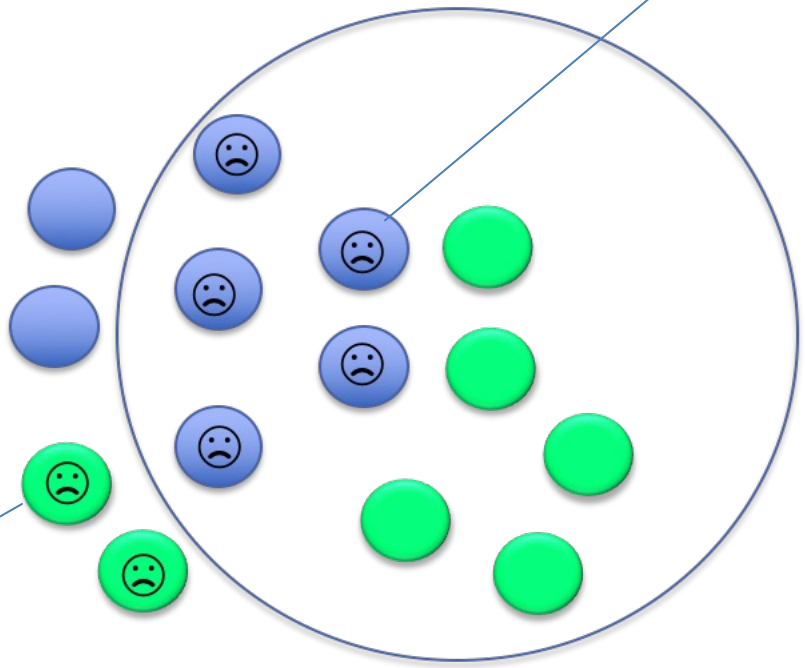
Precision and Recall

Am I actually finding what I'm looking for?

- When doing a corpus query we usually want to find all relevant attestations.
- At the same time, we would like to keep the number of false hits as low as possible.
- These two aspects are also called precision and recall.

Precision and Recall

- We're looking for **green**...



Precision and Recall

- $Precision = \frac{True\ positives}{True\ positives + False\ positives}$

- $Recall = \frac{True\ positives}{True\ positives + false\ negatives}$

- Ideal: 100% Precision & 100 % Recall

- What's more important: precision or recall?

Precision and Recall

What do you think:

- Which factors can lead to false negatives (i.e. hits our query doesn't find)?

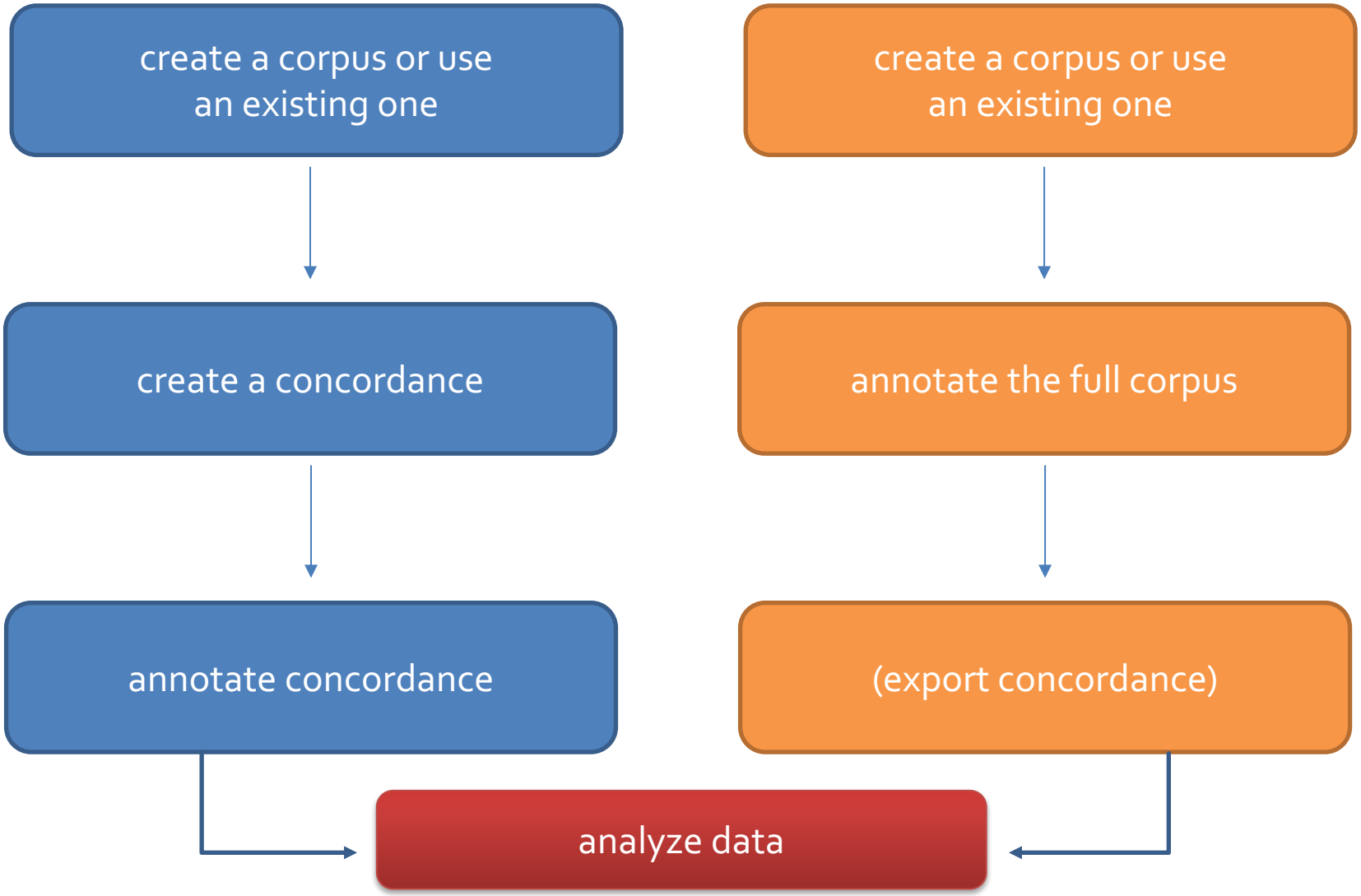
- How can we keep the number of false positives and false negatives as low as possible?

Annotation

Annotation

- Enriching language data with additional information
- can be added to the entire corpus or to concordances

Possible workflows



Annotation

- Key element of each annotation campaign: Annotation guidelines / annotation schema
- The annotation schema defines a set of clearly defined labels for a specific annotation unit (e.g. word, phrase, sentence) (see Ide 2017)
- Example: STTS tagset as an annotation schema

Ide, Nancy. 2017. Introduction. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 1–18. Dordrecht: Springer.

POS tagging with STTS

das	ist	ein	Beispiel
das	sein	ein	Beispiel
PDS	VVFIN	ART	NN

Lemma

Wortart (POS: Part Of Speech)

Annotation

- Annotation can be done automatically or manually
- in the case of automatic annotation, an algorithm is trained to apply the annotation schema
- in the case of manual annotation, the data are annotated by human coders

Example: Sentence annotation

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. (…) Es war kein Traum. Sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden. (Franz Kafka, Die Verwandlung)

Example: Sentence annotation

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. (…) Es war kein Traum. Sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden. (Franz Kafka, Die Verwandlung)

Example: Sentence annotation

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. (…) Es war kein Traum. Sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden. (Franz Kafka, Die Verwandlung)

she's shit at life she won't stay out and party mm she 's such a granny considering she 's only twenty-nine she 's such a granny (SpokenBNC2014)

Example: Sentence annotation

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. (…) Es war kein Traum. Sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden. (Franz Kafka, Die Verwandlung)

she's shit at life she won't stay out and party mm she 's such a granny considering she 's only twenty-nine she 's such a granny (SpokenBNC2014)

Example: Sentence annotation

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. (…) Es war kein Traum. Sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden. (Franz Kafka, Die Verwandlung)

she's shit at life she won't stay out and party mm she 's such a granny considering she 's only twenty-nine she 's such a granny (SpokenBNC2014)

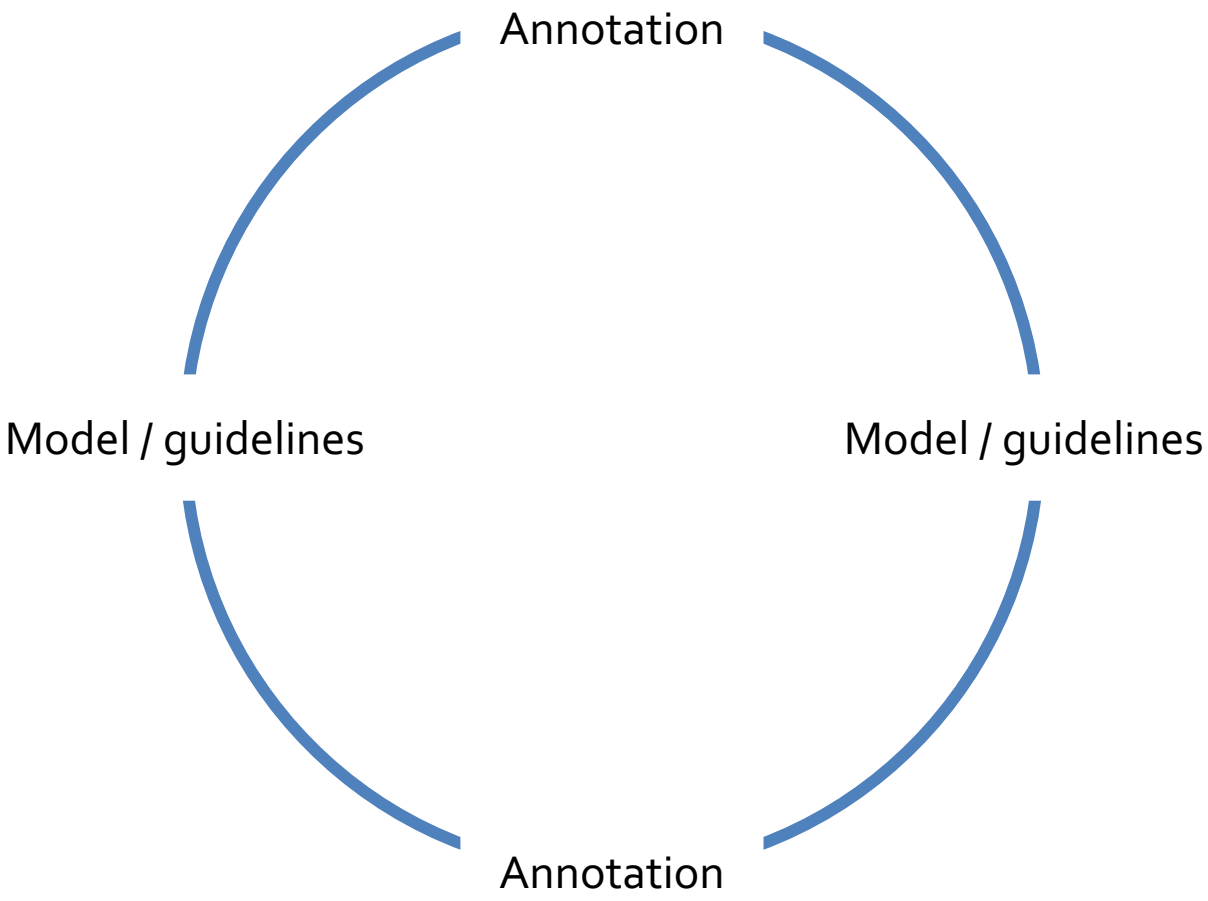
Example: Animacy

- different operationalizations, some more fine-grained, others more coares-grained
- Example: binary categorization animate/inanimate
  - But: Are bacteria and viruses animate or inanimate?
  - What about plants, plancton, …?
- → Annotation = Interpretation!
- Hence: maximal transparency about annotation criteria is important!

Annotation guidelines

- precise and transparent presentation of the annotation criteria
- Principle of reproducibility and replicability
- clear critera: when independent coders use the criteria, they should ideally arrive at the same results

MAMA cycle

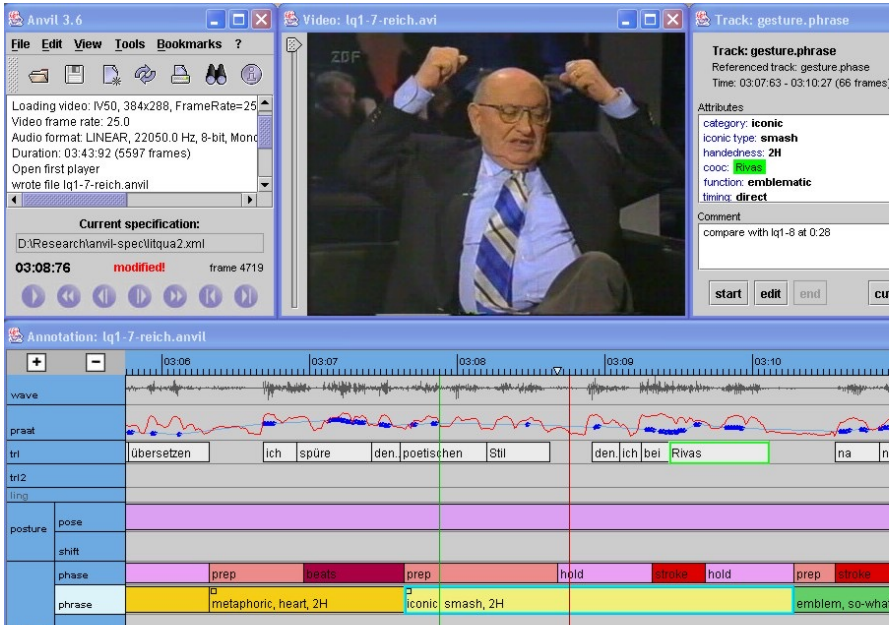

Annotation
Model / guidelines
Model / guidelines
Annotation

Annotation tools

- Multi-layer: e.g. GATE, INCEpTION (successor of WebAnno), MMAX2

- for multimedia data:
  - ELAN
  - EXmaralda

EXMARaLDA



Inter-Annotator Agreement

Annotator/in A

Annotator/in B	belebt	unbelebt	abstrakt	Summe
belebt	5	7	10	22
unbelebt	7	8	5	20
abstrakt	3	5	9	17
Summe	15	20	24	59

- The diagonal values show how many observations are allocated to the same category by the annotators
- By summing up the diagonal values and dviding by the total number of observations, we receive a simple measure of inter-annotator agreement:

$$p = \frac{\sum Digonal\ fields}{n} = \frac{(5 + 8 + 9)}{59} = 0{,}37$$

Inter-Annotator Agreement

- Which disadvantage does this measure have?

Inter-Annotator Agreement

- Which disadvantage does this measure have?

Agreement can be due to chance: Annotators can have chosen the same category by pure coincidence. This percentage is the higher the less categories are used.

→ Hence: The IAA measure should be corrected for potential chance agreements!

(vgl. Bortz, Jürgen & Nicola Döring. 2006. *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler.* 4. Aufl. Heidelberg: Springer.)

Inter-Annotator Agreement

- Example: Cohen's Kappa

$$\kappa = \frac{p - p_e}{1 - p_e}$$

$$p_e = \frac{1}{n^2} \sum_{j=1}^{k} row\ sum_j\ column\ sum_j$$

Inter-Annotator Agreement

	animate	inanimate	abstract	sum
animate	5	7	10	22
inanimate	7	8	5	20
abstract	3	5	9	17
Summe	15	20	24	59

p = 0,37, s.o.

$$p_e = \frac{1}{59^2} \cdot (15 \cdot 22 + 20 \cdot 20 + 24 \cdot 17) \approx 0{,}33$$

$$\kappa = \frac{p - p_e}{1 - p_e} = \frac{0{,}37 - 0{,}33}{1 - 0{,}33} = 0{,}07$$

Annotation: Wrap-up

- Annotation is "art" and "science"
- Annotation usually is qualitative and interpretative
- in the spirit of reproducibility, it's important to keep annotation guidelines as clear and transparent as possible.

From concordance to analysis: Practical examples for working with spreadsheet programmes

Spreadsheet programmes

- e.g. Excel and Calc
- GoogleSheets and other web-based applications for collaborative work

Example: *programmiert* vs. *vorprogrammiert*

- Bastian Sick: *vorprogrammiert* doesn't make sense because everything is programmed in advance….
- → Does this popular metalinguistic statement have an impact on actual language use?
- Method: Searching for *programmiert* and *vorprogrammiert* in DWDS

Step 1: Easy

- Please go to https://dwds.de and search for *programmiert* and *vorprogrammiert* in the Core Corpus of the 20th Century (DWDS-Kernkorpus des 20. Jahrhunderts)
- Import the data to a spreadsheet programme.

Step 2: Harder

- Create a pivot table showing the (absolute and/or relative) frequency of *programmiert* and *vorprogrammiert* by decade.
- Try to visualize the table.

Step 3: Not exactly trivial

- Create a new column containing the **last word** from the column with the left context
- Sort the table by the last word in the left context.
- Create a pivot table showing which words occur often before *programmiert* and *vorprogrammiert.*

Example: *Sinn machen* vs. *Sinn ergeben*

- Bastian Sick: *Sinn machen* doesn't make sense because sense is nothing that can be made.
- Again, we ask: Does this popular meta-linguistic statement have an impact on language use?
- Method: Query for *Sinn machen* vs. *Sinn ergeben* in DWDS