Korpuslinguistik

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

AntConc

- laurenceanthony.net/software

Ressourcen und einschlägige Software

- für Konvertierung der Ausgabedateien von Korpora ins KWIC-Format:

  github.com/hartmast/concordances

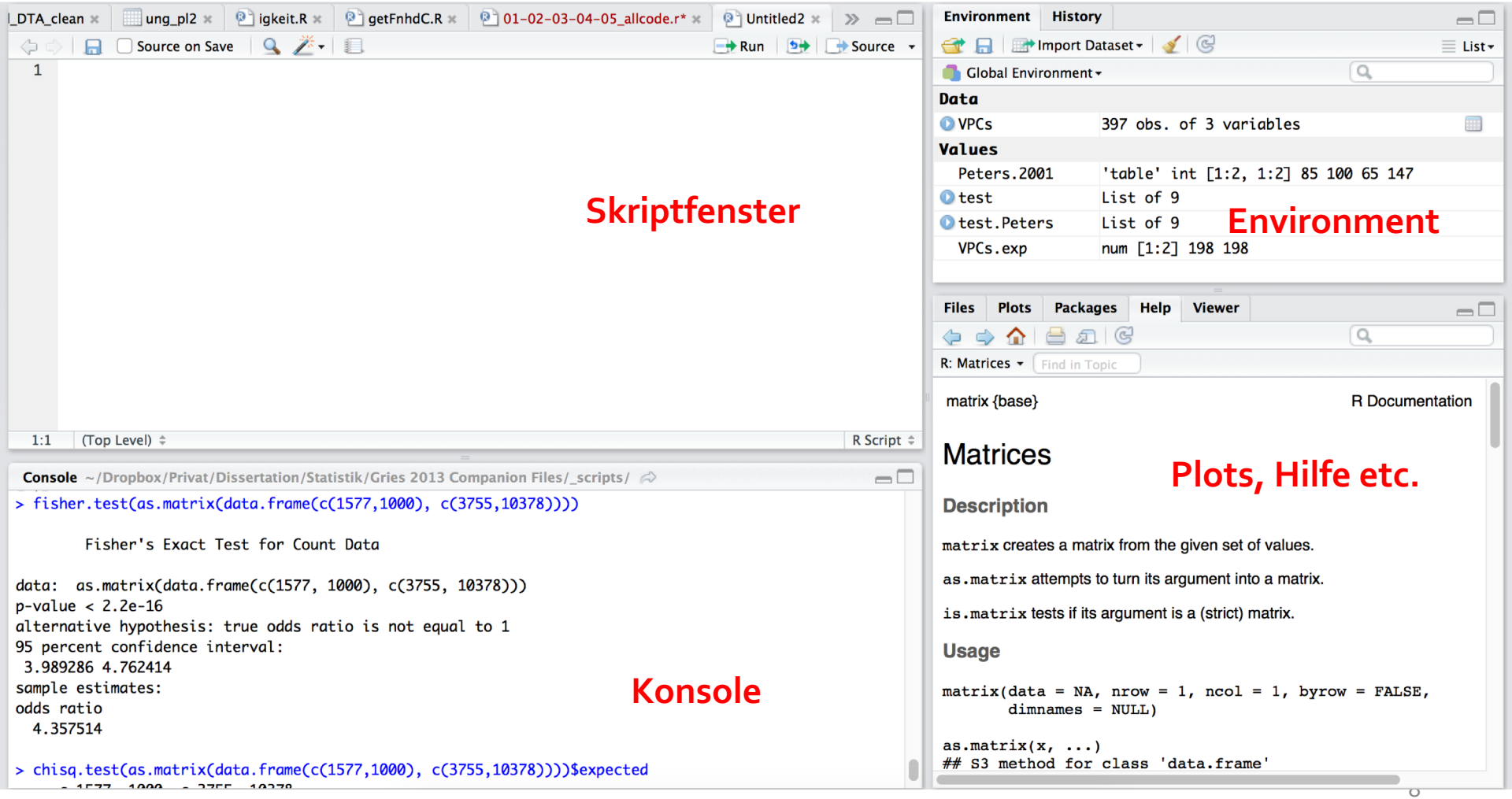
- Tutorials auf

  hartmast.github.io/sprachgeschichte

R

- Statistikprogramm und Programmiersprache
- kostenlos verfügbar unter www.r-project.org
- (geringfügig schnellere Variante: Revolution R – kann mehrere Prozessorkerne gleichzeitig benutzen)
- R ist ein Konsolenprogramm, eine gute Benutzeroberfläche bietet z.B. RStudio.

R Studio


_DTA_clean
ung_pl2
igkeit.R
getFnhdC.R
01-02-03-04-05_allcode.r*
Untitled2
Source on Save
Run
Source
1
Skriptfenster
1:1
(Top Level)
R Script
Console ~/Dropbox/Privat/Dissertation/Statistik/Gries 2013 Companion Files/_scripts/
> fisher.test(as.matrix(data.frame(c(1577,1000), c(3755,10378))))
Fisher's Exact Test for Count Data
data: as.matrix(data.frame(c(1577, 1000), c(3755, 10378)))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
3.989286 4.762414
sample estimates:
odds ratio
4.357514
> chisq.test(as.matrix(data.frame(c(1577,1000), c(3755,10378))))$expected
Konsole
Environment
History
Import Dataset
List
Global Environment
Data
VPCs
397 obs. of 3 variables
Values
Peters.2001
'table' int [1:2, 1:2] 85 100 65 147
test
List of 9
test.Peters
List of 9
VPCs.exp
num [1:2] 198 198
Environment
Files
Plots
Packages
Help
Viewer
R: Matrices
Find in Topic
matrix {base}
R Documentation
Matrices
Plots, Hilfe etc.
Description
matrix creates a matrix from the given set of values.
as.matrix attempts to turn its argument into a matrix.
is.matrix tests if its argument is a (strict) matrix.
Usage
matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,
dimnames = NULL)
as.matrix(x, ...)
S3 method for class 'data.frame'

R

- R-Paket *concordances* lässt sich weitgehend **ohne** jegliche R-Kenntnisse benutzen.
- Wichtig ist nur, dass die in den Tutorials dargestellten Schritte richtig ausgeführt werden.
- Fehler sind natürlich dennoch möglich…

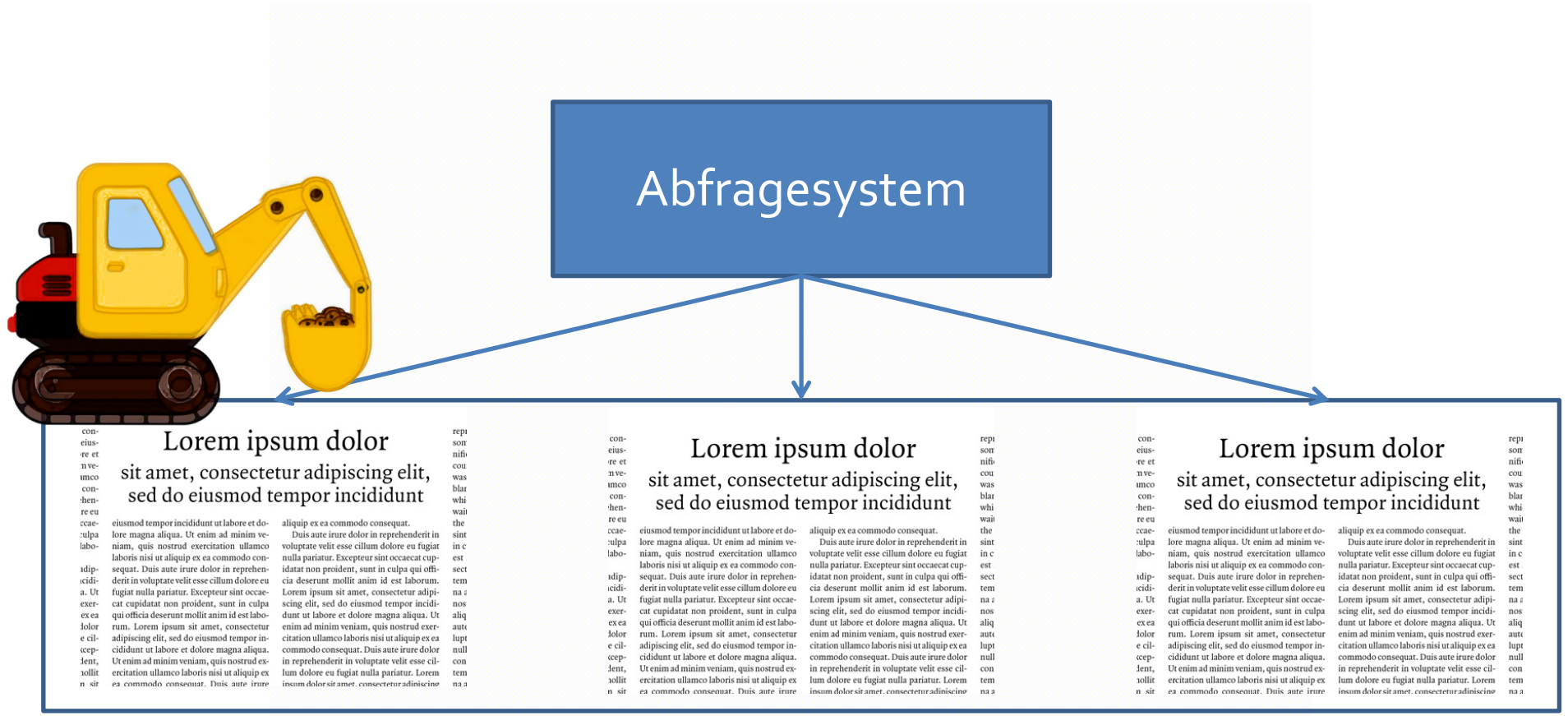
Ressourcen und einschlägige Software

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

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

Abfragesysteme und Abfragesyntax

Korpus und 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

Anatomie eines Korpus

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.

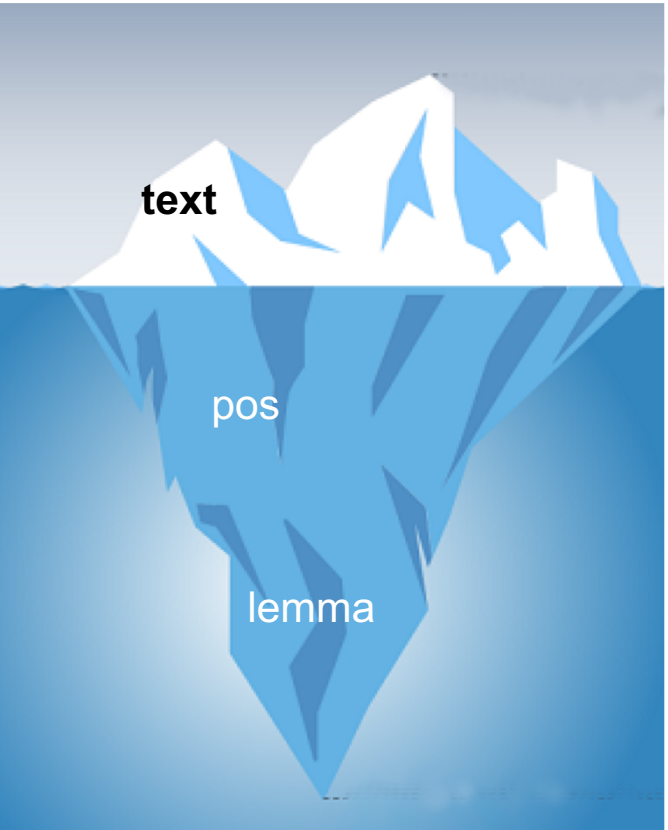
Anatomie eines Korpus

<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

Korpus und Abfragesystem

- In den meisten Korpora kann man auf allen Ebenen suchen, als Ergebnis bekommt man aber oft „nur“ den Text angezeigt.



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

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.

Abfragesyntax

- Korpusabfragesysteme haben oft eine eigene **Syntax**.
- Korpusabfrage-Syntax ist ein bisschen wie Fremdsprachensyntax…

➢ Man kann eine Fremdsprache benutzen, obwohl man die Syntax nur rudimentär beherrscht…

➢ …aber ihre Ausdrucksmöglichkeiten kann man nur mit guten Syntaxkenntnissen ausnutzen.

Abfragesyntax

Welche Optionen benötigen wir überhaupt, wenn wir ein Korpus durchsuchen?

Beispiel 1:

Wir suchen alle Belege für das Verb *legen* in einem **nicht getaggten** und **nicht lemmatisierten** Korpus.

Abfragesyntax

Welche Optionen benötigen wir überhaupt, wenn wir ein Korpus durchsuchen?

Beispiel 2:

Wir suchen alle Belege für die Verben *setzen*, *stellen*, *legen* in einem **lemmatisierten** Korpus.

Abfragesyntax

Welche Optionen benötigen wir überhaupt, wenn wir ein Korpus durchsuchen?

Beispiel 3:

Wir suchen Belege für die Wendung *je* X-*er desto* Y-*er* in einem getaggten Korpus.

Abfragesyntax

Welche Optionen benötigen wir überhaupt, wenn wir ein Korpus durchsuchen?

Beispiel 4:

Wir suchen Belege für *weil* + Substantiv und *weil* + Adjektiv in einem **getaggten** Korpus:

- Ich kann heute nicht ins Kino, weil Seminar.
- Ich will heute nicht ins Kino, weil müde.
- aber **nicht:** …weil Seminar ist.

Abfragesyntax

Welche Optionen benötigen wir überhaupt, wenn wir ein Korpus durchsuchen?

Beispiel 5:

Wir suchen Belege für V-*en gehen* in einem getaggten Korpus.

- Ich gehe heute schwimmen.
- Ich will heute schwimmen gehen.
- Ich gehe heute mit meinem Freund schwimmen.

Was brauchen wir also?

Logische Operatoren

- UND
- ODER
- NICHT

Wildcards

- leg*, setz*, stell*

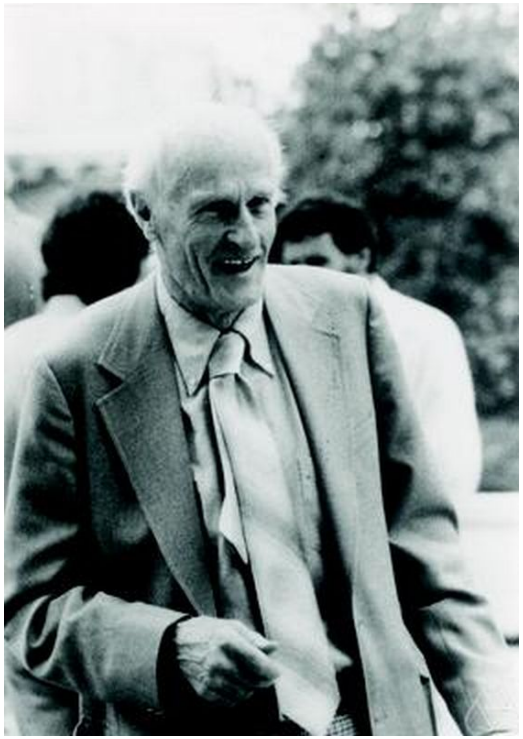
Wortabstandsoperatoren

- Ich gehe {0-5} schwimmen

Wie finde ich, was ich suche?

Reguläre Ausdrücke

- Zeichensequenz, die ein Suchmuster definiert
- Ursprung: Mathematik und Informatik



Stephen Kleene

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!)

Die wichtigsten regulären Ausdrücke

Wildcards und Wiederholungsoperatoren

- . irgendein Zeichen
- ? das Zeichen unmittelbar davor tritt 0- oder 1-mal auf.
- * das Zeichen unmittelbar davor tritt 0- oder x-mal (in unmittelbarer Folge) auf.
- + das Zeichen unmittelbar davor tritt 1- oder x-mal (in unmittelbarer Folge) auf.
- {n} das Zeichen unmittelbar davor tritt genau n-mal (in unmittelbarer Folge) auf.
- {x,} das Zeichen unmittelbar davor tritt mindestens x-mal (in unmittelbarer Folge) auf.
- {x,y} das Zeichen unmittelbar davor tritt mindestens x-, maximal y-mal (in unmittelbarer Folge) auf.

Die wichtigsten regulären Ausdrücke

Weitere Operatoren

- | oder-Operator
- \ Escape-String, z.B. um "echte" Fragezeichen zu finden
- ^ Anfangsposition
- $ Endposition

Zum Gebrauch regulärer Ausdrücke

- Kleinere und größere Unterschiede je nach Korpusabfragesprache
- z.B. in DWDS teilweise doppelter ODER-Operator erforderlich
- in einigen Korpora benutzt man statt oder-Operator das Wort oder (oder OR, oder ODER, ….)
- Ergo: Jede Korpusabfragesprache will gelernt sein (wie echte Sprachen auch…)

Das nächste Level:

Bracket expressions

- [:alnum:] alphanumerisch (a, b, 1, 2)
- [:alpha:] alphabetisch (a, b, c, nicht 1, 2)
- [:digit:] Ziffern (1, 2, 3, … nicht a, b, c)
- [:blank:] Leerzeichen, Tabstopps
- [:punct:] Interpunktion

Lookaround / Lookahead

(?=foo)	Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist *foo*
(?<=foo)	Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist *foo*
(?!foo)	Negative Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist nicht *foo*
(?<!foo)	Negative Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist nicht *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist *foo*
(?<=foo)	Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist *foo*
(?!foo)	Negative Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist nicht *foo*
(?<!foo)	Negative Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist nicht *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist *foo*
(?<=foo)	Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist *foo*
(?!foo)	Negative Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist nicht *foo*
(?<!foo)	Negative Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist nicht *foo*

Lookaround / Lookahead

(?=foo)	Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist *foo*
(?<=foo)	Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist *foo*
(?!foo)	Negative Lookahead	Der String, der der gesuchten Position unmittelbar folgt, ist nicht *foo*
(?<!foo)	Negative Lookbehind	Der String, der der gesuchten Position unmittelbar vorausgeht, ist nicht *foo*

Übungen zu regulären Ausdrücken

- s. Github-Verzeichnis *Regex!*

Übung zu regulären Ausdrücken

- Wie können wir folgende Suchanfragen im DWDS formulieren? (mit Hilfe regulärer Audrücke + der DWDS-Hilfe zur Suche)

  Wir suchen…

  – alle Komposita mit dem Bestimmungsglied *-papst*
  – …alle Verben mit dem Präfix *be-*
  – alle Eigennamen (NE) und Appellative (NN)
  – die Konstruktion *je* X-*er desto* Y-*er*
  – Varianten von *je* X-*er desto* Y-*er* wie z.B. *je* X-*er umso* Y-*er* oder auch *je* X-*er je* Y-*er, umso* Y-*er umso* Y-*er*

Übung zu regulären Ausdrücken

- Wie können wir folgende Suchanfragen im DWDS formulieren?

  Wir suchen…

  – Partikelverben mit "Fronting", z.B. *an hat sie das Licht gemacht, nicht aus*

  – voran- und nachgestellte Genitive

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 benutzt standardmäßig Windows-1252 (ähnlich ISO/IEC 8859-1 / Latin-1 / ASCII)
- Unix und Linux benutzen standardmäßig UTF-8 (Unicode-basiert)
- Die meisten deutschsprachigen historischen Korpora sind (wegen der Sonderzeichen) UTF-8-kodiert.
- Windows kann UTF-8, aber manchmal nur widerwillig…
- Daher bei Arbeit mit Windows oder Microsoft-Programmen (Excel!!) immer darauf achten, dass keine Sonderzeichen verlorengehen.

Encoding Hell

- Im Blick auf Encoding hat das kostenlose Calc einige Vorteil ggü. Excel
- (dafür jedoch teils schlechtere Performance und weniger Optionen)

Hinter den Kulissen eines Korpus

- tinyurl.com/korpling-siegen1

Was bringt der Blick hinter die Kulissen?

- Korpusdateien sehen oft furchteinflößend komplex aus…
- …aber sie sind hochstrukturiert!
- Das hat viele Vorteile, wenn man Suchabfragen machen will, die die jeweilige Suchabfragesyntax nicht kann.

Was bringt der Blick hinter die Kulissen?

- Beispiel: Das Referenzkorpus Mittelhochdeutsch verfügt über eine Satzgrenzenannotation…
- … aber es ist derzeit nicht möglich, ANNIS zu sagen: "Suche nur innerhalb eines bestimmten Satzes!"
- Wenn man das Korpus hingegen mit eigenen Skripten durchsucht, ist das kein Problem.

Beispiel: 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>
```

Beispiel: 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>
```

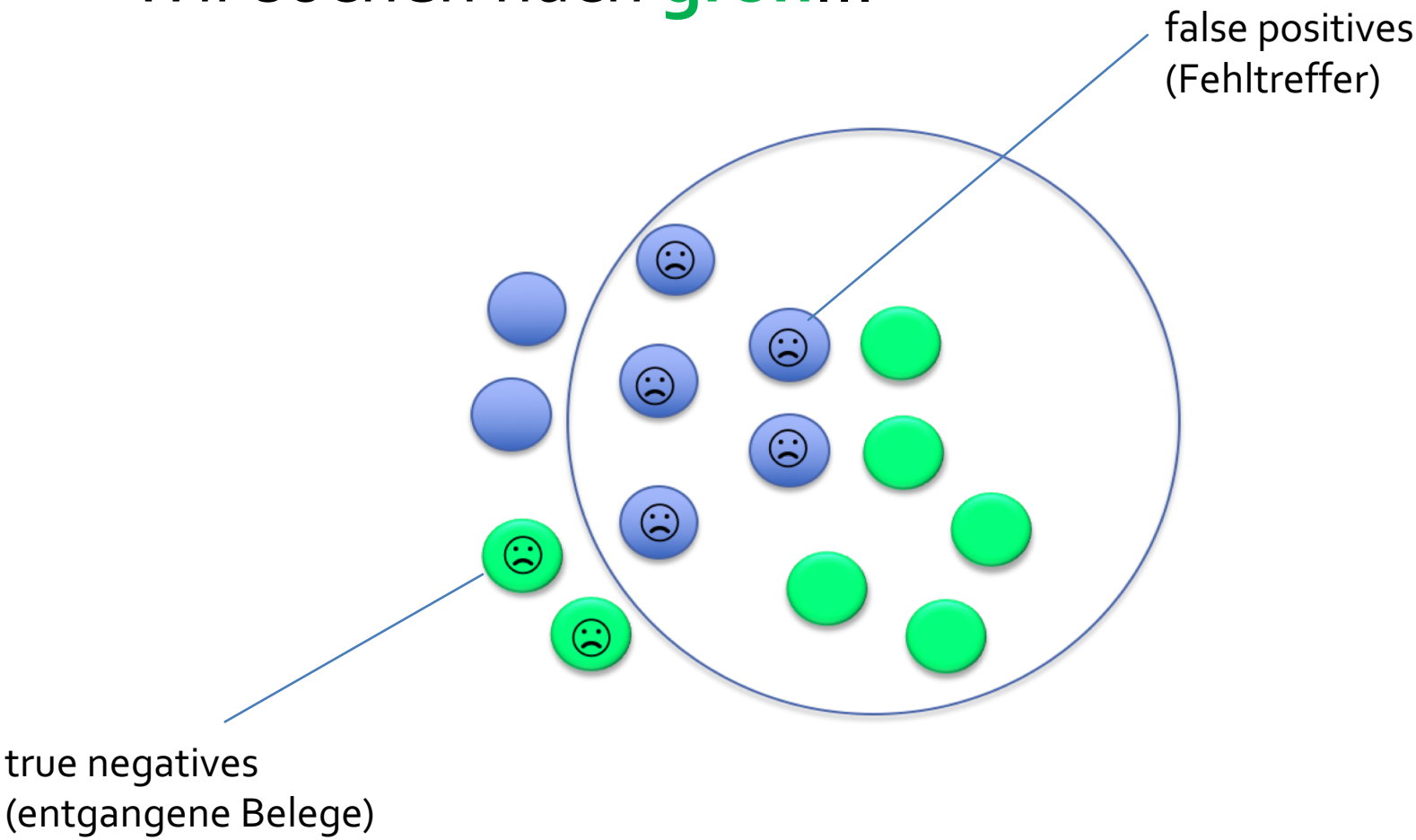
Precision und Recall

Finde ich, was ich suche?

- Bei einer Korpusrecherche wollen wir möglichst **alle** für uns relevanten Belege finden.
- Gleichzeitig möchten wir die Zahl der Fehltreffer möglichst **gering** halten.
- Man spricht hier auch von *Precision* und *Recall*

Precision und Recall

- Wir suchen nach **grün**…



Precision und Recall

- $Precision = \frac{Richtige\ Treffer}{Richtige\ Treffer + Fehltreffer}$

- $Recall = \frac{Richtige\ Treffer}{Richtige\ Treffer + entgangene\ Belege}$

- Ideal: 100% Precision und 100 % Recall

- Was ist wichtiger: Precision oder Recall?

Precision und Recall

Bitte überlegen Sie:

- Welche Faktoren können dazu führen, dass uns Treffer **entgehen**?

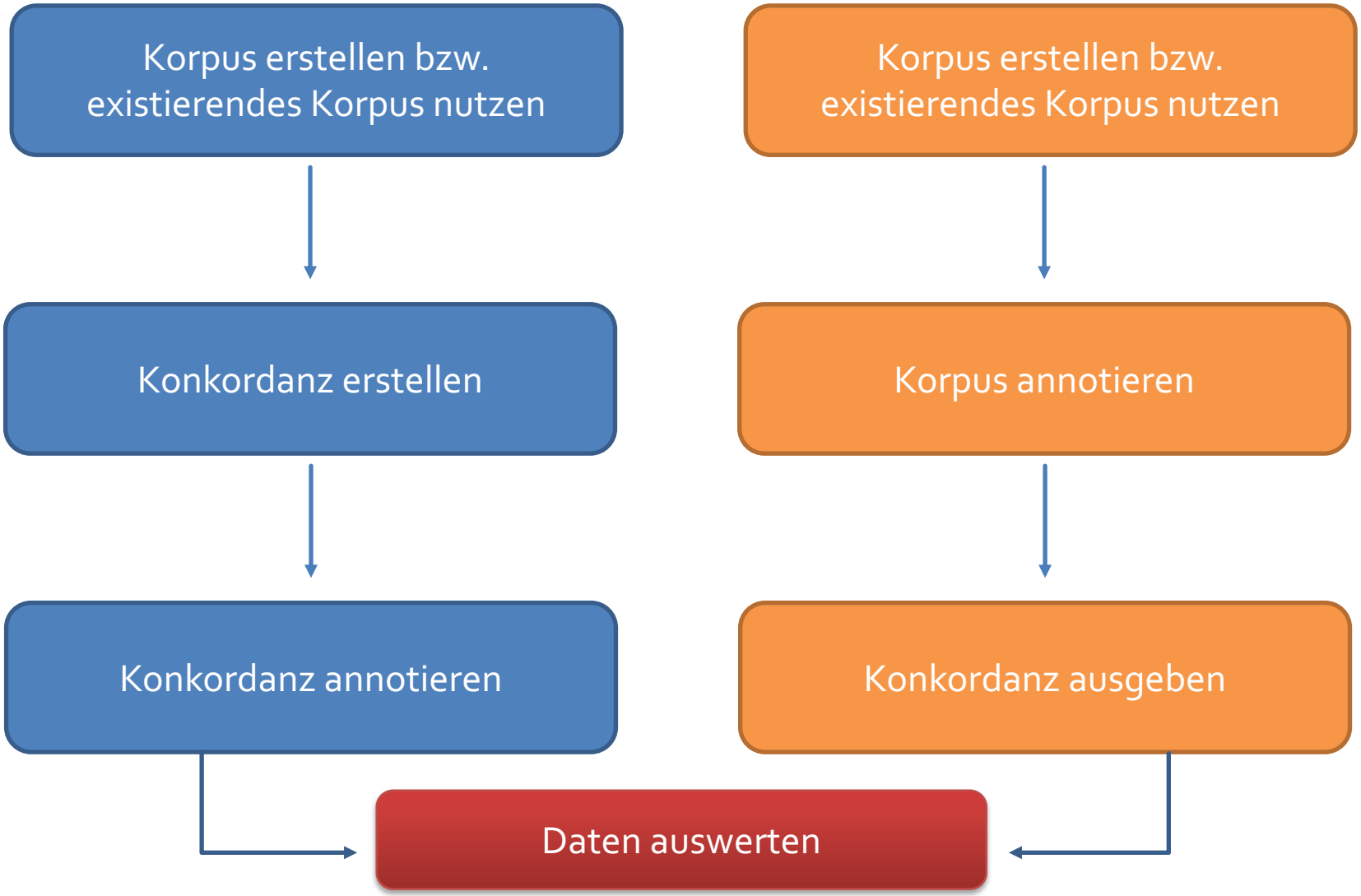
- Wie können wir die Zahl der Fehltreffer und der entgangenen Belege gering halten?

Annotation

Annotation

- Anreicherung von Sprachdaten mit zusätzlichen Informationen
- kann auf die gesamten Korpusdaten angewandt werden oder auf Konkordanzen (s.o.)

Mögliche Workflows



Annotation

- Zentrales Element jedes Annotationsvorhabens ist das **Annotationsschema**
- Es definiert für eine **Annotationseinheit** (z.B. Wort, Phrase, Satz) ein Inventar an klar definierten "Labels" (vgl. Ide 2017)
- Beispiel: STTS-Tagset als Annotationsschema

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

Annotation

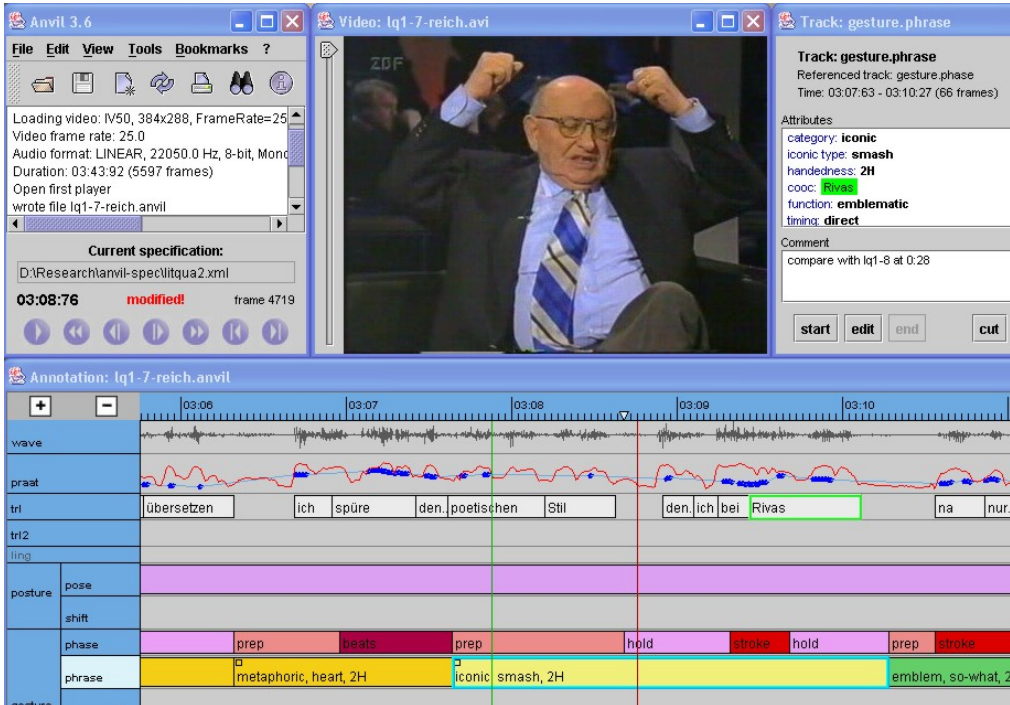
- Annotation kann automatisch oder manuell erfolgen
- Bei automatischer Annotation wird ein Computerprogramm darauf "traniert", das Annotationsschema anzuwenden
- Bei manueller Annotation annotiert man selbst (oder lässt ~~Sklaven~~ Hiwis annotieren)

Annotationstools

- Multi-Layer: z.B. GATE, MMAX2

- für Multimedia geeignet und verbreitet:
  - ELAN
  - EXmaralda

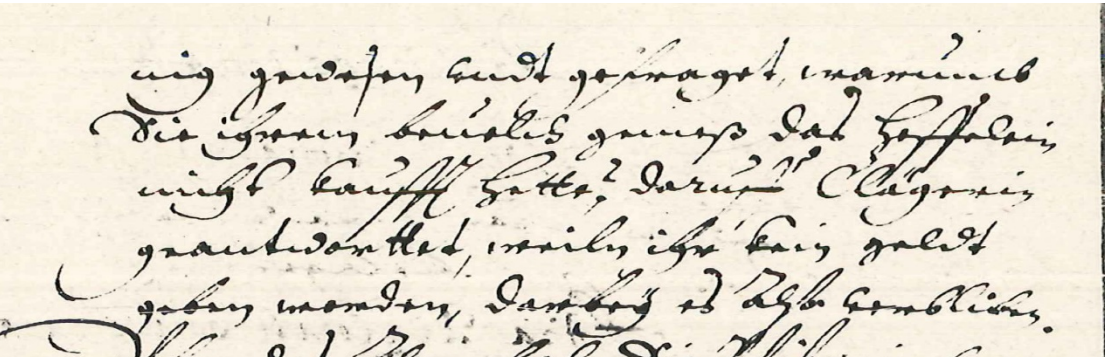
EXMARaLDA



Beispiel SiGS-Korpus: Satzannotation

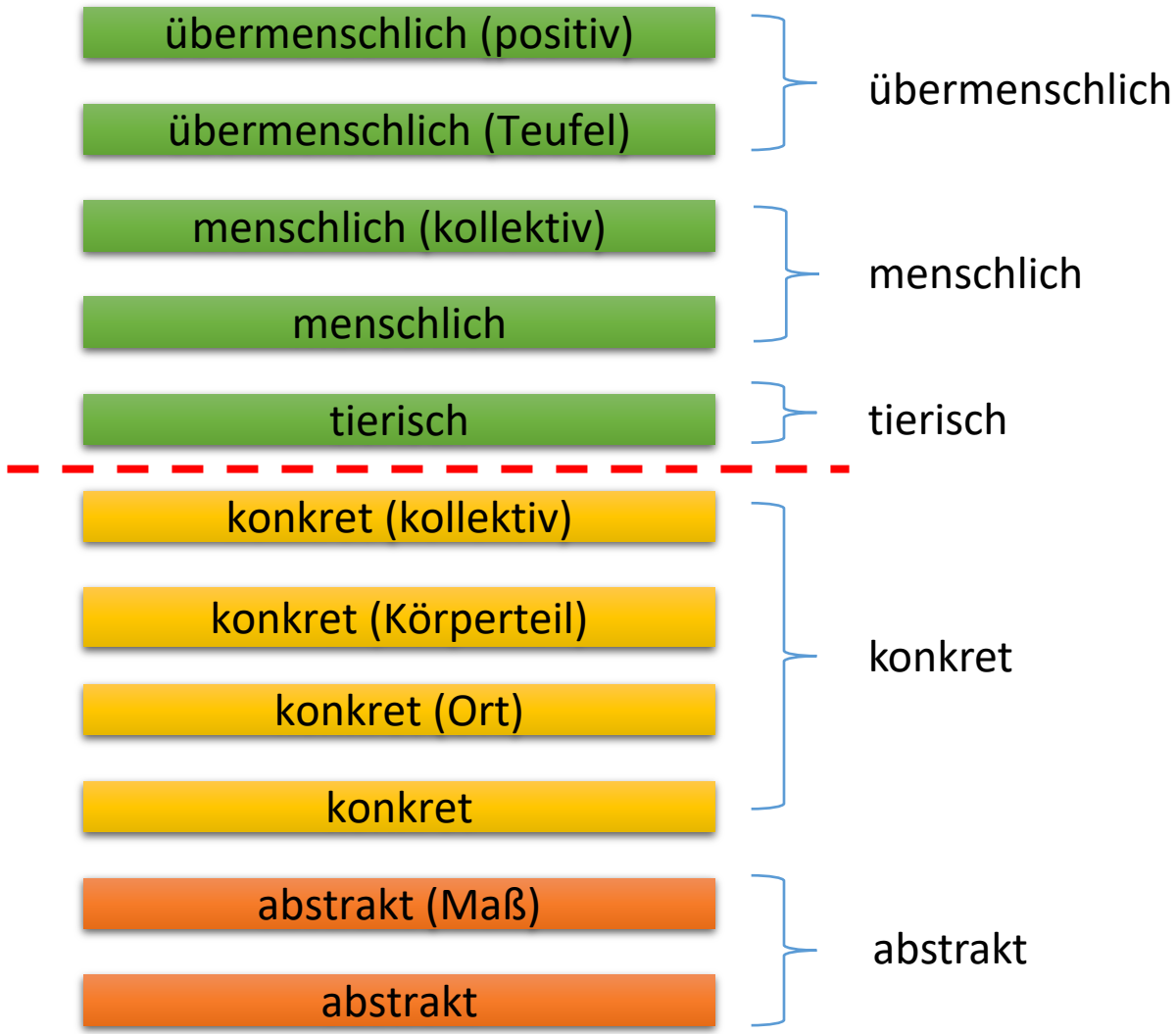
- "konservative" Annotation der Satzgrenzen, um zu verhindern, dass Instanzen **satzinitialer** Großschreibung als Datenpunkte in die Untersuchung eingehen
- Interpunktion bei fnhd. Texten kein verlässlicher Indikator für Satzgrenzenannotation
- daher sog. "Minimalsätze", bestehend aus Vollverb und seiner unmittelbaren Umgebung
- [daruff Clägerin geantworttet, weiln ihr kein geldt geben worden], [darbey es Also verblieben].

Beispiel SiGS-Korpus: Satzannotation



daruff Clägerin geantworttet, weiln ihr kein geldt geben worden, darbey es Also verblieben.

Beispiel: Belebtheit im SiGS-Korpus



Beispiel: Belebtheitsannotation

- https://www.soscisurvey.de/tutorial151376/

Beispiel SiGS-Korpus: Belebtheit

- Regeln zur metaphorischen Verwendung erzeugen Abweichungen:
  - gerihte (bei 5. Item)
  - höret her gi grawen, gy grunen, schwarten, witten, bunten
  - wege und weise (s.u.)
  - Ausdehnung von w_syn: ihne <w_syn>hans ganß</w_syn> geheißen
    der herr Jesus
    heiligen geists

  - direkt übermenschliche Begriffe, wie Teuffel und Gott mit Metapher annotiert
    was ist mit Jesus? und in vater, son und heiliger geist?

- auf w: Teile von größeren syntaktsichen Token
  toppfer knecht -> toppfer als konkret oder menschlich?

Beispiele für Zweifelsfälle

- Landmark, Feldmark konkret oder Ort?
- Elben: menschlich, übermenschlich (positiv) oder übermenschlich (Teufel)?
- Buhle: menschlich oder übermenschlich (Teufel)?
- Heilige als übermenschlich (positiv) oder menschlich?
- Dorf: konkret (Kollektivum) oder Ort?

Bitte nicht wundern, dass *kopff* zweimal als „konkret" anstatt als „konkret (Körperteil)" annotiert ist. In der Edition steht, dass es sich hier um einen Becher handelt.

Annotationsrichtlinien

- möglichst genaue, transparente Darlegung der Annotationskriterien
- Prinzip der **Reproduzierbarkeit** bzw. **Replizierbarkeit**
- möglichst eindeutige Kriterien: bei Anwendung der Kriterien auf dieselben Daten soll ein unabhängiger Dritter zu den gleichen Urteilen/Ergebnissen kommen können

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

- Den Diagonalwerten ist zu entnehmen, wie viele Beobachtungen von den AnnotatorInnen der gleichen Kategorie zugeordnet wurden.
- Durch Aufsummieren der Diagonalwerte und Teilen durch die Gesamtzahl der Beobachtungen, erhält man eine einfache Kennzahl für die Beobachterübereinstimmung:

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

Inter-Annotator Agreement

- Welchen Nachteil hat diese Kennzahl?

Inter-Annotator Agreement

- Welchen Nachteil hat diese Kennzahl?

Auch bei zufälliger Klassifizierung stimmen einige Beobachtungen überein. Dieser Prozentsatz ist umso höher, je weniger Kategorien verwendet werden.

→ Daher: Zufallskorrektur des Übereinstimmungsmaßes!

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

Inter-Annotator Agreement

- Beispiel: Cohen's Kappa

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

$$p_e = \frac{1}{n^2} \sum_{j=1}^{k} Zeilensumme_j \, Spaltensumme_j$$

Inter-Annotator Agreement

	belebt	unbelebt	abstrakt	Summe
belebt	5	7	10	22
unbelebt	7	8	5	20
abstrakt	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: Fazit

- Annotation als "art" und "science"
- Annotation bringt häufig eine qualitative, interpretative Komponente mit sich
- im Sinne der Reproduzierbarkeit ist es jedoch wichtig, Annotationskriterien klar, transparent und intersubjektiv nachvollziehbar zu gestalten.

Von der Konkordanz zur Analyse:

Praktische Beispiele für die Arbeit mit Tabellenkalkulations-programmen

Tabellenkalkulationsprogramme

- im Wesentlichen zwei Alternativen: Excel (z.B. über Uni-Lizenz), Calc (kostenlos)
- beide haben Vor- und Nachteile:
  - Excel kann insgesamt etwas mehr und ist z.T. intuitiver zu bedienen…
  - Calc kann dafür etwas unkomplizierter mit Unicode-Daten umgehen.
- Für kollaboratives Arbeiten eignet sich außerdem auch GoogleSheets.

Beispiel: *programmiert* vs. *vorprogrammiert*

- Bastian Sick: *vorprogrammiert* macht keinen Sinn, weil man ja einen automatischen Ablauf immer im Voraus programmiert
- → Wirkt sich dieses sprachkritische Urteil auf den Sprachgebrauch aus?
- Methode: Suche nach *programmiert* und *vorprogrammiert* im DeReKo
- Konkordanz als csv-Datei im VC!

Aufgabe – Stufe 1: einfach

- Bitte öffnen Sie die Datei *programmiert.csv* mit Calc oder Excel.
- Sie enthält die Belege für *programmiert* und *vorprogrammiert* aus dem DeReKo aus den 90er- und 00er-Jahren.

Aufgabe – Stufe 2: mittel

- Erstellen Sie eine Pivot-Tabelle, die die (absolute und/oder relative) Frequenz von *programmiert* und *vorprogrammiert* nach Jahrzehnt darstellt.
- Überführen Sie die Tabelle nach Möglichkeit in eine passende grafische Darstellung.

Aufgabe – Stufe 3: nicht völlig trivial

- Erstellen Sie eine Spalte, die das **letzte Wort** aus der Spalte "Left" enthält, also das letzte Wort aus dem linken Kontext.
- Sortieren Sie die Tabelle nach dem letzten Wort im linken Kontext.
- Erstellen Sie eine Pivot-Tabelle, die Ihnen zeigt, was besonders gerne "programmiert" und was eher "vorprogrammiert" wird.

Beispiel: *Sinn machen* vs. *Sinn ergeben*

- Bastian Sick: *Sinn machen* macht keinen Sinn, weil man Sinn nicht machen kann.
- Auch hier fragen wir: Wirkt sich dieses sprachkritische Urteil auf den Sprachgebrauch aus?
- Methode: Suche im DWDS nach *Sinn machen* vs. *Sinn ergeben*

Aufgabe – Stufe 1: fast noch trivial

- Bitte überlegen Sie sich eine geeignete Suchanfrage für *Sinn machen* vs. *Sinn ergeben* im DWDS.