Kurzgefasste Korpuslinguistik für GermanistInnen

Roland Schäfer

8. Juli 2014

Wir sind hier...

Korpusstudien

Überlegungen vor der quantitativen Korpusstudie:

- Ist eine Korpusstudie überhaupt das richtige Instrument?
- Welche Grundgesamtheit interessiert mich?
- Welches Korpus repräsentiert die Grundgesamtheit?
  Es gibt kein „repräsentatives Korpus“ an sich!
- Wie muss ich die Stichprobe passend zu meiner Hypothese durchführen, um die Hypothese valide zu testen?
- Ist das Korpus hinreichend linguistisch annotiert?
- Was müsste herauskommen, um meine Hypothese zu falsifizieren?

- Alles in allem müssen die Fragen aus Teil 2 der Statistik-VL vor der Studie lückenlos geklärt sein!

Probleme

Wann ein Korpus evtl. nicht passt:

- Wenn klassische „Grammatikalität“ ein zentraler Begriff der zu testenden Theorie ist.
- Bei seltenen Phänomenen (kann zumindest Probleme machen).
- Wenn das Phänomen nicht (halb-)automatisch zu suchen ist und keine Kapazitäten für die manuelle Suche vorhanden sind.
- Wenn keine passenden Korpora existieren.

Zitieren Sie immer die Artikel der Korpus-Ersteller!

Lesen Sie sie auch!
Es könnten relevante Informationen über das Korpus drinstehen!

Wir sind hier...

Grober Ablauf

1. Hypothesen formulieren
2. Hypothesen operationalisieren
3. Korpus wählen, Stichprobenverfahren definieren
4. Stichprobe ziehen, prüfen, evtl. neu ziehen
5. Stichprobe ggf. annotieren
6. Hypothesen inferenzstatistisch prüfen

Alle diese Schritte sollten in Ihrem Artikel/Ihrer Arbeit dokumentiert und alle Entscheidungen konzeptuell motiviert werden.

Zum Stichprobenverfahren

- Die Stichprobengröße kann i. d. R. nicht a priori festgelegt werden.
- Stichprobengrößen sind ein Kompromiss aus α-Niveau und Teststärke.
- Nur bei gerichteten Hypothesen (s. Statistik-VL) gibt es ideale Größen.
- Wahl zwischen Quotenstichprobe und Zufallsstichprobe:
  Ich bin vehementer Vertreter von Zufallsstichproben!
  - Quotenstichproben sind fehleranfällig,
  - sie bringen massenweise Auxiliarhypothesen mit,
  - und die Angst vor der großen Zufallsstichprobe ist meist unbegründet.
  - Die Suchanfrage(n) dürfen nur niemals relevante Daten ausschließen!
- Dringend empfohlen im Stichprobenprozess (und schon im Operationalisierungsprozess): Konkordanz-Surfen!
  Die besten Ideen kommen im Angesicht der Daten!

Zur Annotation

Was sollte man annotieren?

- NUR hypothesengebundene Variablen!
- Auf Verdacht „schnell noch Kasus (etc.) mit-annotieren“ ist Zeitverschwendung.
- Die Geschwindigkeit der AnnotatorInnen muss getestet werden.
- Lieber drei Variablen kodieren und genug Belege bekommen als acht Variablen und dann mit hundert Belegen dastehen!

Hinweis zur Korpusqualität

Grenzen der verfügbaren linguistischen Annotation in Korpora:

- Tokenizer/Tagger/Chunker usw. sind auf Standardsprache trainiert.
- Je weniger nah an dieser das gesuchte Phänomen ist, desto unzuverlässiger sind die Informationen im Korpus...
- ...zusätzlich zur allgemeinen Fehlerrate.
- Prüfen Sie immer an kleinen Stichproben, ob die Suchanfrage in linguistischer Annotation korrekt läuft.
- Eventuell sind bereits konzeptuelle Entscheidungen von den Taggern anders getroffen worden, als bei Ihnen.
- Beispiele:
  - Unterscheidung von Infinitiven und Präsensformen (PI) ist nicht 100% genau.
  - Suche nach Lemmata wie „bausparen“ führt nicht zu „baugespart“.

Wir sind hier...

Rüstzeug für KorpuslinguistInnen

Was muss man können?

- übliche Abfragesprachen von Korpus-Software
- Tabellenkalkulation
- Rohtextbearbeitung (wenigstens Notepad++ o. ä.)
- Statistik inkl. R, SPSS o. ä.
- Berechnung von Inter-Kodierer-Übereinstimmung

Was sollte man können?

- eine Skriptsprache
- reguläre Ausdrücke
- Stochastik

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Platzhalter

- Oft sucht man nach Zeichenketten, die bestimmte Platzhalter enthalten.
- Komposita mit *Haus* als Erstglied usw.
- Reguläre Ausdrücke (Regexe) erlauben es, Zeichenketten allgemeiner als durch literale Zeichenketten anzugeben.
- Vereinfacht kann man sich die Regex-Syntax als Platzhalter-Syntax vorstellen.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Zeichen

Man kann ein Zeichen in Regexen folgendermaßen definieren:

- literal
  z. B. `a` oder `4` ⇒ genau ein Vorkommen genau dieser Zeichen
- Zeichenklasse in `[]`, interpretiert als Aufzählung
  `[a4]` ⇒ genau ein Vorkommen von entweder a oder 4
- Zeichenklasse mit Bereich
  `[A-Z]` ⇒ genau ein Vorkommen eines Zeichens zwischen A und Z
- aufgezählte Zeichenklassen-Bereiche
  `[A-Za-z]` ⇒ genau ein Vorkommen eines Zeichens in A–Z oder a–z
- negierte Zeichnklasse mit `^` (Komplementbildung)
  `[^A-Z]` ⇒ genau ein Vorkommen irgendeines Zeichens **außer** A–Z

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Beliebige Zeichen

- irgendein Zeichen (ob Ziffer, Buchstabe, Sonderzeichen...)
  `.`
- Als Folge der besonderen Beduetung einiger Zeichen wie `.` in der Regex-Syntax, müssen die zugehörigen Literale geschützt werden:
  - `\.` ⇒ der Punkt
  - `\\` ⇒ der Backslash
  - `\[` ⇒ die eckige öffnende Klammer
  - ...das geht mit allen Zeichen so, die eine besondere Bedeutung haben.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Wir sind hier...

Korpora und Korpusstudien
Korpusstudien planen
Ablauf der Studie
Kenntnisse
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Logik und Klammerung
Morphologische Annotation
Grenz-Markierungen
DWDS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Wiederholung von Zeichen

- Alle Literale und Zeichnklassen sowie der Platzhalter stehen für **ein** Zeichen in der Zeichnkette.
- `Ha.s` findet also *Hans*, *Haus*, *Ha5s*, aber nicht *Hauses* usw.
- Nach jedem Zeichen oder jeder Zeichnklasse kann daher ein Quantor stehen.
- Der Quantor gibt an, wie oft das Zeichen (oder irgendein Zeichen der Klasse) wiederholt werden kann oder muss.
- `*` ⇒ null mal oder beliebig oft
- `+` ⇒ mindestens einmal
- `{3}` ⇒ genau drei Mal
- `{3,6}` ⇒ drei bis sechs Mal
- `{3,}` ⇒ drei Mal oder öfter

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Wir sind hier...

Korpora und Korpusstudien
Korpusstudien planen
Ablauf der Studie
Kenntnisse
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Logik und Klammerung
Morphologische Annotation
Grenz-Markierungen
DWDS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Grenzen von Zeichenketten

- Zwei besondere Zeichen markieren Positionen in der Zeichenkette.
- `^` ⇒ der Anfang der Zeichenkette
- `$` ⇒ das Ende der Zeichenkette

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Gruppen

- Man kann Teile von Zeichenketten zusammenfassen.
- `Haus(meister)` ⇒ nichts anderes als `Hausmeister`
- Die zusammengefassten Gruppen können allerings ein Öder`|` enthalten.
- `Haus(meister|tür)` ⇒ *Hausmeister* oder *Haustür*

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Grundlagen
Wiederholungen von Zeichen
Gruppen und Nicht-Zeichen
Liste der reservierten Zeichen
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP

Übungen

Schreiben Sie Regexe, die folgendes finden:

1. nur: das und welches
2. nur: Bus, Bussie und Busfahrer
3. Bus, Bussie und Busfahrer, aber nicht Bushaltestelle
4. gegenderte Nomina wie: StudentIn
5. dass. im Singular **und** Plural
6. alle Flexionsformen von rot
7. Wörter, die mit X oder x beginnen
8. Wörter, die mit Q beginnen und mit ant oder anten aufhören

Wir sind hier...

Reservierte Zeichen

Folgende Zeichen haben eine besondere Bedeutung in Regexen und müssen geschützt werden, wenn sie literal gemeint sind:

Zeichen	Bedeutung
.	beliebiges Zeichen
[ ]	Zeichenklasse
*	Quantor: 0–beliebig
+	Quantor: 1–beliebig
{ }	numerischer Quantor
\	Schutz für Literale
( )	Gruppe
|	Oder (Gruppe)
^	Anfang der Zeichenkette
$	Ende der Zeichenkette

Merkmale und operationalisierung

- Variablen, die normalerweise von Menschen handkodiert werden sind z. B.:
  - informationsstrukturelle Kategorien wie Topic
  - anaphorische Beziehungen
  - Kasus
  - Textsorte
- Alle Variablen, die von Menschen kodiert werden, müssen möglichst so genau operationalisiert werden, dass mehrere Menschen anhand der Operationalisierung die gleichen Entscheidungen treffen würden.
- Im Idealfall lässt man deshalb zwei oder mehr Kodierer dieselben Belege kodieren und prüft dann, wie gut die Ergebnisse übereinstimmen.

Zwei Kodierer

- Wenn mehrere Kodierer einfach irgendetwas anklicken, erwartet man trotzdem zufällige Übereinstimmung.
- Wie groß diese zufällige Übereinstimmung ist, hängt neben der Anzahl der Kodierer (hier jetzt erstmal immer 2) von der Anzahl der möglichen Ausprägungen der kodierten Variable ab.
- Cohens κ ist ein Maß dafür, um wieviel besser eine gegebene Übereintimmung gegenüber der per Zufall erwartbaren Übereinstimmung ist.

Kreuztabulierung der Ergebnisse

- Für das DECOW2012 haben zwei Kodierer eine Zufallsstichprobe von $n = 200$ Texten unter anderem für Zielpublikum nach den COWCat-Richtlinien kodiert.
- Die möglichen Werte waren: allgemein – informiert – professionell
- In der Kreuztabelle trägt man jetzt pro Fall ein, wie KodiererIn 1 (Zeilen) abhängig von KodiererIn 2 (Spalten) entschieden hat.
- Jede Abweichung von der Diagonalen markiert Nicht-Übereinstimmung.

Kreuztabelle

K1↓ K2→	allgemein	informiert	professionell	Summen
allgemein	151	14	4	169
informiert		20	6	26
professionell			5	5
Summen	151	34	15	200

Wie man sieht tendiert K1 (salopp gesagt) dazu, das Textniveau geringer einzuschätzen und es dadurch einem weniger informierten Publikum zuzutrauen.

Einfache Übereinstimmung

K1↓ K2→	allgemein	informiert	professionell	Summen
allgemein	151	14	4	169
informiert		20	6	26
professionell			5	5
Summen	151	34	15	200

Der einfache Anteil der Übereinstimmung ist:

$$p_0 = \frac{\text{Diagonalsumme}}{n}$$

$$\text{Bsp. } p_0 = \frac{151+20+5}{200} = \frac{176}{200} = 0.88$$

Zufällige Übereinstimmung

K1↓ K2→	allgemein	informiert	professionell	Summen
allgemein	151	14	4	169
informiert		20	6	26
professionell			5	5
Summen	151	34	15	200

Wenn wir die Spaltensumme in Spalte i mit SS_i und die Zeilensumme in Zeile j mit ZS_j bezeichnen, ist die zufällig erwartbare Übereinstimmung:

$$p_c = \frac{\sum_i SS_i \cdot ZS_i}{n^2}$$

Zufällige Übereinstimmung: Beispiel

K1↓ K2→	allgemein	informiert	professionell	Summen
allgemein	151	14	4	169
informiert		20	6	26
professionell			5	5
Summen	151	34	15	200

$$p_c = \frac{169 \cdot 151 + 26 \cdot 34 + 5 \cdot 15}{200^2} = \frac{25519 + 884 + 75}{200^2} = \frac{26478}{40000} = 0.66$$

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Cohens κ ausrechnen

Die Formel lautet:

$$\kappa_c = \frac{p_o - p_c}{1 - p_c}$$

Bsp.: $\kappa_c = \frac{0.88 - 0.66}{1 - 0.66} = \frac{0.22}{0.34} = 0.65$

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Interpretation von κ_c

- Wie immer ist die Interpretation eines Gütewerts keine vollständig normierte Sache.
- Ich würde den Wert prinzipiell einfach als Wert angeben.
- Die Faustregeln für die Praxis vielleicht:
  - $\kappa_c < 0.5$: Operationalisierung überprüfen und nochmal kodieren (lassen).
  - $\kappa_c > 0.7$: Ggf. im Text ausdrücklich auf die gute Übereinstimmung hinweisen.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Mehr als zwei Kodierer

- Wenn man mehr als zwei (also m) Kodierer hat, nimmt man Fleiss' κ_f.
- p_0 muss über m Dimensionen gerechnet werden, es gibt also mehr als nur Spalten und Zeilen.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

In R

Für die Berechnung muss meist eine Matrix oder ein Dataframe vorliegen, der die Ergebnisse folgendermaßen tabuliert:

Fall	K1	K2	...	Km
1	informiert	informiert	...	...
2	informiert	allgemein	...	...
3	professionell	informiert	...	...
n	...	...	...	...

Wenn so eine Tabelle im Objekt `kodierungen` vorliegt:

```
> library(irr)
> kappa2(kodierungen)
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Andere Maße, andere Daten

- Hier wurde implizit von Nominaldaten ausgegegangen.
- Es gibt andere Maße für Nominaldaten und andere Maße für andere Skalenniveaus.

Für einen Überblick die Liste aus `irr` anfordern:

```
> irr::<TAB>
```

Dann z. B.:

```
> ?kendall
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
Motivation
Cohens κ
Fleiss' κ
COSMAS
DWDS
CQP

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Logik und Klammerung
Morphologische Annotation
Grenz-Markierungen
DWDS
CQP

COSMAS

Vorteile:

- relativ groß
- nach Jahrgängen sortiert, immer aktuell

Nachteile:

- fast nur Zeitungssprache
- viele Regionalblätter, die die Stichproben dominieren (das berühmte *St. Galler Tageblatt*
- Momentan größter Nachteil: Obsoletheit... Das neue ganz große Ding kommt angeblich bald.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
Allgemeines
Zugang und Export
Einfache Suchanfragen
Abstände
Logik und Klammerung
Morphologische Annotation
Grenz-Markierungen
DWDS
CQP

Hinweis zum Namen

COSMAS ist der Name der Software.

Die IDS-Korpora pauschal als „COSMAS“ zu zitieren ist unprofessionell und unfair ggü. den Personen, die die Korpora machen, aber mit der Software COSMAS nichts zu tun haben!

Wir sind hier...

Zugang und Export

- Für das rein technische gibt es ein Videotutorial:
  `http://www.youtube.com/watch?v=EfxSegFzPZI`
- Wichtig: Die Konkordanzen dürfen nicht dauerhaft gespeichert werden, bitte sorgfältig die Nutzungsbedingungen lesen.
- Der Export kann mit Koka für eine Tabellenkalkulation vorbereitet werden:
  `http://hpsg.fu-berlin.de/~rsling/microsites/koka/`

Wir sind hier...

Wortformen(ketten) suchen

Einfache Wortform suchen:
> `Chuzpe`

Kette von aufeinanderfolgenden Wortformen:
(mit Weggelassener Verknüpfungsoperator bedeutet: Wortabstand /+w1)
> `hat Chuzpe`

Lemma-Suche und Wortformenexpansion

Alle Formen einer Grundform:
> `&Haus &kaufen`

Die Expansionsliste
Man kann die konkrete Liste der orthographischen Varianten bzw. Wortformen einsehen und bearbeiten, wenn man vor dem Absetzen der eigentlichen Anfrage auf `Liste zu: ...[Öffnen]` klickt.

Wir sind hier...

Wortabstände

Beispiel: Man sucht nach einem Artikel gefolgt von maximal einem Adjektiv, dann einem Nomen.

> `&ein /+w2 &Baum`

Optimaler, weil dann das seltenere Wort zuerst gesucht wird:
> `&Baum /-w2 &ein`

Satzabstand und im selben Satz suchen

Im selben Satz (ungerichteter Abstandsoperator):
> `&Haus /s0 &Hof`

In aufeinanderfolgenden Sätzen (...wirklich?):
> `&Haus /+s1 &Hof`

Abstandsbereiche

Die Abstandsoperatoren können alle ungerichtet (ohne + oder -) angegeben werden, und sie können Bereiche als Argument nehmen.

Simulation eines Nahe-Bei-Operators:
> `Chuzpe /w4 Verve`

Lösung des Problems mit aufeinanderfolgenden Sätzen – Bereiche:
> `&Haus /+s1:1 &Hof`

Also: Einzelne Zahlen n bezeichnen den Abstand bis maximal n.

Wir sind hier...

Logik-Operatoren UND ODER NICHT

Die Operatoren beziehen sich auf ein ganzes Dokument!
Man muss dabei die Argumente des Operators als Bedingungen an das Dokument lesen („muss enthalten“).

```
> Chuzpe UND Verve
```

Achtung: ODER ist inklusiv.

```
> &Kanzlerin ODER &Präsidentin
```

Achtung: NICHT ist eigentlich ein logisches UND NICHT.

```
> &Kanzlerin NICHT &Präsidentin
```

(Die letzte Anfrage liefert eine Teilmenge der vorletzten.)

Klammern

Mit Klammern muss man mehrere Abstandsoperatoren kombinieren.

```
> (&können /w1 &wollen) /w4 &haben
```

Vor allem ODER wird in Klammern aber eine Art echtes ODER:

```
> (&Bad ODER &Küche) /+w1 &schrubben
```

Objektsprachliches UND ODER NICHT

Die Operatoren müssen als normale Wörter in Anführungszeichen:

- "und"
- "oder"
- "nicht"

Wir sind hier...

MORPH-Operator

Mit dem MORPH()-Operator kann man auf POS-Tags zugreifen.

Achtung Wie immer funktioniert es nicht so, wie die Hilfe behauptet. Vergesst STTS, nehmt einfach den MORPH-Assistenten!

```
> MORPH(VRB inf v) /+w1:1 MORPH(VRB inf m) /+w1:1 MORPH(VRB fin a)
```

Beschränkungen auf Wortform und Tag

Will man gleichzeitig Lemma/Wortform und morphologische Information einschränken, muss man den 0-Wörter-Abstand-Trick benutzen:

```
> MORPH(VRB fin v) /w0 &laufen
```

Man sucht also nach einer finiten Verbform und im Abstand von 0 Wörtern einer Form von *laufen*.

Wie sucht man nach einer solchen Form im Abstand von maximal vier Wörtern nach links oder rechts zu einer nominalen Form von *Lauf*?

Wir sind hier...

Satz- und Absatz-Enden und -Anfänge

Es gibt die folgenden strukturellen Attribute:

- <sa> Satzanfang
- <se> Satzende
- <pa> Absatzanfang
- <pe> Absatzende

Auch sie sucht man mit dem 0-Wörter-Trick (z. B. VL-Sätze):

```
> MORPH(VRB fin) /w0 <se>
```

Wir sind hier...

DWDS-Kernkorpus

Vorteile:

- balanciertes Referenzkorpus
- deckt das 20. Jh. gleichmäßig ab
- Zusatzinformationen: Kollokationen, Wortverlauf,...

Nachteile:

- viel zu klein
- einige Strata sind verzerrt
  (z. B. 80er Jahre Wissenschaft = Slotterdijk)
- deckt per Definition nur das 20. Jh. ab
- zu viele Treffer nicht sichtbar
- schlechte Exportoptionen

- Hauptnachteil: das neue Interface

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Zugang und Export

- Für das rein technische gibt es ein Videotutorial:
  `http://www.youtube.com/watch?v=6pTW26PJeIc`
- Der Export kann mit Koka für eine Tabellenkalkulation vorbereitet werden:
  `http://hpsg.fu-berlin.de/~rsling/microsites/koka/`
- Die Anfragesprache ist einfacher,
  aber unterm Strich m. E. zielführender als mit COSMAS.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wortformen und Grundformen

Im Gegensatz zu COSMAS sucht DWDS immer auch nach Formvarianten!
> Haus
...sucht also auch *Hause*, *Häuser* usw.

Man muss umgekehrt den Wortform-Operator einsetzen:
> @Haus
...sucht definitiv nur *Haus*.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wortabstand

Alle Anfragen beziehen sich prinzipiell auf innerhalb eines Satzes.
Alle Sequenzanfragen müssen in " stehen:
> "Haus bauen"

Der einfache Abstand muss nicht markiert werden, alles darüber hinaus mit #:
> "der #1 Präsident"

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Bedingungen innerhalb eines Satzes

Der Satz ist sowieso die Suchdomäne. Wenn die Linearisierung egal ist, müssen logische Operatoren verwendet werden:

- `!Krieg` (Satz enthält nicht *Krieg*)
- `Frieden && Krieg` (Satz enthält *Frieden* und *Krieg*)
- `Frieden || Krieg` (Satz enthält *Frieden* und/oder *Krieg*)
- `Frieden && !Krieg` (Satz enthält *Frieden* und nicht *Krieg*)

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Zugriff auf STTS-Tags

Suche alle infiniten Modalverben ein oder zwei Wörter nach *haben*:
> "haben #2 $p=VMINF"

Kombinieren von Wortform und Tag-Information mit `with`...
Suche alle finiten Formen von *haben* als Vollverb:
> haben with $p=VVFIN

Kombiniert mit Abstand:
"haben with $p=VAFIN #4 Eis"

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
Allgmeines
Zugang und Export
Einfache Suchanfragen
Abstände
Morphologische Annotation
CQP

Wir sind hier...

CWB (bzw. CQP)

Vorteile:

- nicht an bestimmte Korpora gebunden
- auch für eigene Korpora
- sehr mächtig und flexibel
- eingebaute Statistikfunktionen

Nachteile:

- Rechenzeit- und Speichergierig
- Beschaffung der Korpora

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Von uns verfügbare Korpora

- COW2012 Token-Größen:
  - DE: 9 Mrd.
  - ES: 1,5 Mrd.
  - NL: 2,3 Mrd.
  - SE: 2,2 Mrd.
  - UK: 1 Mrd.
- COW2014: Alle über 10 Mrd, Deutsch wahrscheinlich über 20 Mrd.
- `http://corporafromtheweb.org/`

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Installation und Basics

- Video-Tutorial zur Installation (Win):
  `http://www.youtube.com/watch?v=5xR3Qikalcs`
- Video-Tutorial zur Korpusinstallation:
  `http://www.youtube.com/watch?v=Ku5Qvd6pax8&hd=1`
- Video-Tutorial zur Icon-Installation/Grafikanpassung (Win):
  `http://www.youtube.com/watch?v=eNL_KI3DAkI&hd=1`

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Das schwarze Fenster mit dem Text

- CQP funktioniert nach dem Textkonsolen/Shell-Prinzip.
- Man schreibt einen Befehl, drückt RETURN...
- ...und das System regiert mit einer Ausgabe.
- Pfeil-hoch und Pfeil-runter navigieren durch bereits getippte Kommandos.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Hinweis zu `rcqp`

Eine sehr praktische Art, auf CQP zuzugreifen,
ist die Schnittstelle `rcqp` für R.

Von den hier beschriebenen CQP-Methoden
ist nur die Syntax der Suchanfragen in `rcqp` gleich.
Korpusauswahl, KWIC-Anzeige usw. gehen grundlegend anders.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Wir sind hier...

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Korpuswahl und Basisbefehle

Verfügbare Korpora anzeigen:

```
> show corpora;
```

Informationen über ein Korpus abrufen:

```
> info DECOW2012-C00X1M;
```

Shell verlassen:

```
> exit;
```


Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Suchanfragen

- Neben einigen Steuerbefehlen ist die typische Eingabe in CQP eine Abfrage in dem Korpus.
- Man kann dabei nur abfragen:
  - Ketten von Tokens und
  - innerhalb bestimmter struktureller Attribute (Sätze, Absätze, Dokumente usw.).
- Die Reaktion der Shell ist die Ausgabe der Suchanfrage als KWIC.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines
Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Wortformen direkt oder mit regulären Ausdrücken

Einfache Wortform(enkette) abfragen, jedes Wort in ":
```
> "zu";
> "zu" "versuchen";
```

Mit regulären Ausdrücken:
```
> ".+lein";
> ".+lein.";
```


Volle reguläre Ausdruckssyntax

Character classes:
```
> "[MN].+lein"
```

Alternativen:
```
> "Mann" "und|oder" "Frau"
```

Quantoren:
```
> "[a-z]{2,3}"
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Übungen – bisher nicht unbedingt perfekt lösbar

1. Alle Flexionsformen von *rot*.
2. Alle Indikativ-Formen von *kaufen*.
3. Alle Indikativformen von *laufen*.
4. Dative mit und ohne Monoflexion der Adjektive.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Voller Zugriff auf Wortinformationen

- Wenn mehr als nur die Wortform eingeschränkt werden soll, muss jedes Such-Token in [] gesetzt werden.
- In den eckigen Klammern werden dann mit & die Suchkriterien für ein Token verbunden.
- Meist:
  - `word=` Wortform (wie bei der einfachen Suchanfrage).
  - `lemma=` Lemma.
  - `pos=` Wortart.
- Die einfache Suchanfrage mit ist eigtl. nur eine Abkürzung für `[word=]`.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Kombinationen, Beispiele

Finite Formen von *haben* als Auxiliar:
```
> [lemma="haben"& pos="VAFIN"];
```

ne als Artikel:
```
> [word="ne"& pos="ART"];
```

-lein-Diminutive:
```
> [pos="NN"& lemma=".+lein"];
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Abstände, endlich konsistent

`[]` ist das *matchall*-Token und kann daher
für die Festlegung von Abständen benutzt werden:

```
> [pos="ART"] []{0,4} [pos="NN"];
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Übungen

- Jetzt nochmal die Monoflexion bitte.
- Massangaben im Stil von *eine Tasse heißer/heißen Kaffe/s*.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Logische Operatoren

- In der [] können außer & auch andere logische Operatoren verwendet werden:
  - oder |
  - nicht !

Finite Formen von *gehen* außer Präteritum:

```
> [lemma="gehen"& pos="VVFIN"& !word="ging.*"];
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Strukturelle Attribute

- Auf strukturelle Attribute kann man entweder in <> direkt zugreifen...
- oder mit `within`.

Satzinititale Adverben:
```
> <s> [pos="ADV"];
```

Satzfinale Adverben:
```
> [pos="ADV"] </s>;
```

Innerhalb eines Satzes:
```
> [lemma="Fortuna"] [lemma="Düsseldorf"] []* [lemma="gewinnen"];
```

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Übungen

- Satzinitiale Adverben alleine im Vorfeld.
- Sätze, die auf einen Nebensatz enden.
- W-Exklamativa.

Kurzgefasste Korpuslinguistik für GermanistInnen
Roland Schäfer
Korpora und Korpusstudien
Reguläre Ausdrücke
Inter-Kodierer-Übereintimmung
COSMAS
DWDS
CQP
Allgemeines Installation und Basics
Einfache Suchanfragen
Anzeigeoptionen setzen
Suchanfragen speichern und ausgeben
Gruppieren und Sortieren

Wir sind hier...

Kontext

Kontext ist genau 1 Satz:

```
> set Context 1 s;
```

Kontext sind 20 Zeichen links und rechts:

```
> set Context 20;
```

PrintStructures

Dokument-URL anschalten:

```
set PrintStructures doc_url;
```

(Nicht definiert in COW-X-Korpora!)

Suchen in Dokumentnamen/Verfassernamen etc.:

```
> [lemma="Raubkopie"] :: match.doc_url = "http.+heise\.de.+";
```

Wir sind hier...

Suchanfragen benennen

Eigentlich sollte man Suchanfragen immer speichern und dann erst mit cat ausgeben:

```
> myQuery = [lemma="Chuzpe"]; cat myQuery;
```

...macht äußerlich dasselbe wie:

```
> [lemma="Chuzpe"];
```

Suchanfrage persisten machen:

```
> save myQuery;
```

Last

- Wenn die letzte Suchanfrage nicht explizit gespeichert wurde, ist sie als Last verfügbar.
- Man kann sie speichern in eine eigens benannte Suchanfrage: `myNewQuery = Last;`

Suchanfrage exportieren

```
> myQuery
> set Context 1 s;
> set PrintStructures doc_url;
> myQuery = "Chuzpe";
> cat myQuery > "dateiname.txt";
```

Übungen

- Zwei der gemachten Übungen nochmal machen, aber mit Speichern, catten und dann exportieren.
- Export durch Koka jagen, Konkordanz in LibreOffice formatieren.

Suchanfragen reduzieren

Wenn man sehr allgemeine Suchanfragen abfragt:

```
> [pos="NN"];
```

kann es sein, dass man nur eine Zufallsauswahl möchte.

Trefferanzahl abfragen:

```
> size myQuery;
```

Auf eine bestimmte Zufallstrefferzahl beschränken:

```
> reduce Last to 100;
```

oder

```
> reduce Last to 5%;
```

Übungen

Bitte bis zur Konkordanz bringen:

- Monoflexion als zwei 100er-Sample (Mono/nicht Mono).
- 10% aller *lein*-Diminutive im Korpus.

Wir sind hier...

match und matchend

- Jeder Treffer wird durch den Match charakterisiert:
  Anfang und Ende des Satzstücks, das der Suchanfrage entsprach.
- Das erste Token davon heißt `match[0]`, das letzte `matchend[0]`.
- Man kann auf Wörter davor oder danach zurückgreifen,
  indem man von 0 entsprechend Werte abzieht.
- Nach diesem Tokens kann man sortieren.

Sortieren

Nach einzelnem Token:

```
> Bla = "mit" [pos="ADJA"& word=".+em"] [pos="ADJA"] [pos="NN"];
> sort Bla by word on match[2];
```

Nach einem Bereich:

```
> sort Bla by word on match[1] .. match[2];
```

Gruppieren/Wortverteilungen

Mit group kann man Resultate gruppieren und die Zähldaten ausgeben lassen.

```
> mySchnell = [lemma="schnell"] [pos="VVFIN"];
> group mySchnell match[1] lemma;
```

Man kann nach allem groupen:

```
> group mySchnell match[-1] pos;
```

Übungen

- Vergleiche die Verteilung der finiten satzfinalen Verb-Arten
  (voll, aux, modal) mit der der nicht-satzfinalen.
- Später: Rechne einen χ^2-Test dazu!
- Finde heraus, mit welchen Wörtern (möglichst Satzzeichen und Zahlen
  und kreativen Sprachgebrauch ausschließen) *statt*
  in welcher Häufigkeit rechts-kombiniert (Lemmata und POS).

Mehrdimensionale Gruppierung

Man kann an die erste Grupierung eine zweite anhängen:

```
> myAdjKonj = [pos="ADJA"] "und" [pos="ADJA"];
> group myAdjKonj match lemma by matchend lemma;
```

Übungen

- Finde die häufigsten Adjektiv-Nomen-Bigramme.
- Finde die Häufigsten Bigramme aus infinitem Vollverb
  und infinitem Modalverb.