

# Humboldt-Universität zu Berlin

## Korpuslinguistik

Einführung:  
CQP  
(Corpus Query Processor)

Anke Lüdeling

Amir Zeldes

Hagen Hirschmann

[hirschhx@hu-berlin.de](mailto:hirschhx@hu-berlin.de)

(und andere Mitarbeiter der Professur für Korpuslinguistik und Morphologie)

# Was ist CQP?

- Eine Abfragesprache für Textkorpora
- Findet linguistische Phänomene
- Erlaubt komplizierte Mustersuchen nach:
  - Zeichenketten
  - Sequenzen von Zeichenketten (Wortfolgen)
  - Eigenschaften wie z.B. Wortarten
- CQP wurde im IMS Stuttgart entwickelt:  
[www.ims.uni-stuttgart.de/forschung/projekte/corpus-workbench/](http://www.ims.uni-stuttgart.de/forschung/projekte/corpus-workbench/)

# Das Web-Interface



Login CQP-Webinterface

username

password

clear login

Anmeldung mit eigenem Account oder:

username: **CQP\_Demo**

password: **TestSuchen**

Login-Seite: [hu-berlin.de/cqp](https://hu-berlin.de/cqp)

Link, um eigenen Account zu beantragen:

<https://korpling.german.hu-berlin.de/korpusregistrierung/>

# Das Web-Interface

The image shows a screenshot of the CQP-Webinterface. The interface is dark blue with white text and form elements. At the top, the title "CQP-Webinterface" is centered. Below the title, there are two main input fields: "query" and "corpus". The "query" field is empty, and the "corpus" field contains "AD 2006". To the right of the "corpus" field is an information icon (i). Below these fields, there are several sections for configuring the search results. The "output" section has three radio buttons: "matches" (selected), "frequencies", and "matches + frequencies". The "basic options" section contains two dropdown menus: "result set" (set to "all") and "output format" (set to "HTML"). The "options for matches output" section has two dropdown menus: "left context" (set to "5 tokens") and "right context" (set to "5 tokens"). Below these are three sections for attributes: "positional attributes" (with a list containing "word", "pos", and "lemma"), "structural attributes" (with a list containing "abstract", "abstract\_thema", "abstract\_autor", "abstract\_jahr", and "abstract\_urldiss"), and "alignment attributes" (with an empty box). At the bottom of the interface, there are two buttons: "clear" and "search". The "search" button is circled in red. Annotations include a red box labeled "Anfrage" pointing to the query field, a red box labeled "Korpus aussuchen" pointing to the corpus dropdown, a black arrow labeled "Informationen" pointing to the information icon, and a red box labeled "Los!" pointing to the search button.

**Anfrage**

**Korpus aussuchen**

**Informationen**

**Los!**

**CQP-Webinterface**

query  ?

corpus  ?

output

- matches
- frequencies
- matches + frequencies

basic options

result set  ?

output format  ?

options for matches output

left context  ?

right context  ?

positional attributes

- word
- pos
- lemma

structural attributes

- abstract
- abstract\_thema
- abstract\_autor
- abstract\_jahr
- abstract\_urldiss

alignment attributes

# Beispiel: Suche nach fester Zeichenkette

- Suchen Sie nach allen Vorkommen der Wortform *trocken*:
  - Anfrage: **[word="trocken"]** )
    - Im Bonner Zeitungskorpus
    - Im c't-Magazin
  - Wie viele Bedeutungen können Sie finden?
  - Gibt es Unterschiede zwischen den Korpora? (Metadaten!)
  - Was wird nicht gefunden, wofür Sie sich aber interessieren könnten?

# Lemmata

- Lemma  $\triangleq$  „Grundform“ eines Wortes
- Suchen Sie nach allen Vorkommen des Lemmas *trocken*:
  - Anfrage: **[lemma="trocken"]**
    - Im Bonner Zeitungskorpus
    - Im c't-Magazin
  - Was wird jetzt gefunden?
  - Was wird trotzdem nicht gefunden, was Sie aber interessieren könnte?

# Mustersuche: \* + und .

- Der Punkt . steht für ein beliebiges Zeichen
  - z.B. **[word="trocken."]** oder **[word="trocken.."]** oder **[word="trocken..."]**
- Das Sternchen \* steht für „beliebig viel“
  - z.B. **[word/lemma="das\*"]** findet „da“, „das“, „dass“ (im Bonner Korpus gibt es noch kein „dass“)
- Das + Zeichen bedeutet „mindestens einmal“
  - z.B. **[word/lemma="das+"]** findet „das“ und „dass“, nicht „da“.
- Was findet man mit **[word="trock.\*"]**?

# Der Operator ?

- Der Operator ? bedeutet „optional“
  - Welche Formen finden Sie mit **[word="dunke?le?n"]** ?
  - Finden Sie alle Genitivformen des Wortes ***Kind***

# a oder b: **(a|b)**

- Mit Klammern kann man gleichzeitig nach verschiedenen Wörtern suchen:
  - **[lemma="(Kind|Mädchen|Junge)"]**
- Nach verschiedenen Formen:
  - **[word="(Kinds|Kindes)"]**
- Oder Zeichenketten:
  - **[word="bes(ser|t).\*"]**
- Finden Sie alle Formen von *sehen* im Präsens

# Zeichenmengen [ ] und ^

- Mit [ ] und ^ kann man Mengen definieren:
  - [aeiou] – einfache Vokale
  - [^äöüÄÖÜ] – alles außer Umlauten
  - [A-ZÄÖÜ] – Großgeschriebene Buchstaben
  - [0-9]+ - Zahlen
- Beispiele:
  - Flektierte Formen von *Kind*: **[word="Kinde?[sr]n?"]**
  - Bindestrichkomposita:  
**[word/lemma="[A-ZÄÖÜ][a-zäöüß]+-[A-ZÄÖÜ][a-zäöüß]+"]**  
Was für Wörter finden Sie?
  - Versuchen Sie alle Formen von *geben* zu finden

# Suche auf mehreren Ebenen

- Suche nach Lemma und Wortform:
  - Finden Sie alle Formen des Verbs *geben*  
*außer der Form geben*:  
**[lemma="geben" & word!="geben"]**
- Finden Sie Wörter, die von „trocken“ abgeleitet werden

# Suche nach Wortart

- Die meisten deutschen Korpora benutzen das Tagset STTS (Übersicht: <https://hu.berlin/stts>)
  - ADJA            attributives Adjektiv
  - ADV            Adverb
  - ART            Artikel
  - NN            normales Nomen
  - VVFIN        finites Verb
  - ...

# Wortart

- Die CQP-Korpora an der HU Berlin sind (fast) alle STTS-getaggt.
- → Liste von Wortarten, entsprechenden Kurzbezeichnungen (Tags) und Beispielen
- Suchen Sie nach Pronominaladverbien.
  - **[pos="PAV"]**  
(Anm.: in manchen Korpora: PROAV)
- Suchen Sie nach *laut* als Adjektiv (Liste)
  - Anfrage: **[word="laut" & pos="ADJ."]**

# Suche nach Wortart

- Finden Sie Verkleinerungsformen/Diminutiva:  
`[lemma=".+chen" & pos="NN"]`  
bzw.  
`[lemma=".+[^aeiouäöü]chen" & pos="NN"]`
- Oder umgelautete Pluralformen bei Nomina:  
`[lemma="[^äöü]+" & word=".+[äöü].+" & pos="NN"]`

# Häufigkeit

- Man kann auch Häufigkeitsverteilungen untersuchen:

output  
 matches  
 frequencies  
 matches + frequencies

- Vergleichen Sie die häufigsten Verkleinerungsformen in beiden Korpora

# Wortfolgen

- Jedes Wort wird durch einen Ausdruck zwischen [ ] dargestellt:

**[word="[li]ch" [word="(gehe|ging)"]**

- Welche Nomina kommen vor Jahreszahlen vor?

**[pos="NN" [word="[12][09][0-9][0-9]"]**

- Suchen Sie im c't-Magazin nach Adjektiven ("ADJA"), die vor *Computer* vorkommen

# Wortfolgenmuster

- den N dem N:

**[word="den"][pos="NN"][word="dem"][pos="NN"]**

- Auch mit beliebig vielen Adjektiven:

**[word="den"][pos="ADJA"]\*[pos="NN"]**

**[word="dem"][pos="ADJA"]\*[pos="NN"]**

- Sätze mit 20-50 Wörtern zwischen Hilfsverb und Partizip:

– **[pos="VAFIN"]**

**[pos!="V.\*" & word!="."]{20,50}**

**[pos="VVPP"]**

# Metadaten

- Informationen über die Daten innerhalb der Korpora
- z.B.:
  - Ich will nur in Artikeln zu bestimmten Themen suchen.
  - Ich will nur in bestimmten Jahrgängen suchen.
  - ...
- Ermitteln der verfügbaren Metadaten: Info-Knopf oder Liste 'structural attributes'
  - Metadatum anklicken, einfache Suche abschicken ...
- Bsp: `[lemma="Golfkrieg"]::match.quelle_year="1996"` findet im Korpus 'Parlamentsreden' Vorkommen des Lemmas *Golfkrieg*. Wenn man mit Treffern der Folgejahre vergleicht, kann man den politischen Diskurs über das Thema verfolgen

# Metadaten – komplexe Aufgabe

- In "Akademisches Deutsch" vorhandene Fachbereichs-Metadaten: *architektur, bauing, biologie, chemie, etechnik, geisteswiss, geschichte, informatik, jura, kunstgeschichte, linguistik, maschbau, mathe, mathematik, medizin, musik, paedagogik, philosophie, physik, politik, psychologie, sinologie, soziologie, sport, wirtschaft* (u.a.)
- Formulieren Sie zwei Suchanfragen für den Vergleich des Personalpromens der ersten Pers. Singular im Fachbereich Medizin auf der einen Seite und **allen Geisteswissenschaften** auf der anderen
- `[lemma="ich"]::match.abstract_sachgebiet="medizin"`
- `[lemma="ich"]::match.abstract_sachgebiet="(kunstgeschichte|soziologie|geisteswiss|geschichte|politik|linguistik|musik|paedagogik|sinologie|linguistik|psychologie)"`

# Zusammenfassung

- Mit CQP kann man:
  - nach Phänomenen suchen, die sich als Muster definieren lassen
  - Phänomene quantifizieren
  - Korpora vergleichen
- Man kann verschiedene Gebiete erforschen:
  - Phonologie (bzw. Orthographie)
  - Morphologie
  - Lexik
  - Semantik
  - Syntax

# Zusammenfassung

- Suche nach Tokens, bezüglich:
  - Wortform
  - Lemma
  - Wortart
  - Auch Negation ist möglich (!=  $\triangleq$  "statt =")
- Suche nach Wortfolgen:
  - Feste Wortfolgen
  - Ketten von Wortarten usw.
  - Muster mit variabler Tokenanzahl

# Zusammenfassung

- Operatoren:

- Ein beliebiges Zeichen
- Beliebig viel (0 bis unendlich)
- Mindestens einmal
- Optional
- Menge (oder  $[^abc]$  = nicht die Menge)
- a oder b
- a 2 bis 3 mal

# zum Schluss: Vorsicht...

- Ein Korpus entspricht nicht der ganzen Sprache
- Unterschiedliche Korpora zeigen unterschiedliche Ergebnisse (das ist auch interessant!)
- Manchmal sind Korpora fehlerhaft
- Trotzdem können Korpora Hypothesen gut unterstützen oder widerlegen
- Methodische Einführungen in die Korpuslinguistik sind wichtig, um Fehler beim Arbeiten mit Korpora zu vermeiden

**Vielen Dank!**

[hirschhx@hu-berlin.de](mailto:hirschhx@hu-berlin.de)

# Ergebnisse: Zahlen zu *ich* in "Akademisches Deutsch"

- Absolute Treffer "Medizin": 21
- Anzahl Personalpronomina: 7833
- Absolute Treffer: 5  
kunstgeschichte|soziologie|geisteswiss|geschichte|politik|linguis  
tik|musik|paedagogik|sinologie|linguistik|psychologie
- Anzahl Personalpronomina: 174  
kunstgeschichte|soziologie|geisteswiss|geschichte|politik|linguis  
tik|musik|paedagogik|sinologie|linguistik|psychologie