# Computergestützte Korpuslinguistik und die Kollokationstheorie

PS: Computerlinguistik

Kristin Dill

# Korpuslinguistik

- Die **Korpuslinguistik** ist ein Bereich der Linguistik, in dem Theorien über Sprache anhand von Belegen oder statistischen Daten aus Textkorpora aufgestellt oder überprüft werden.
- induktive/empirische Methode
- steht im Gegensatz zu deduktiven Methoden:
- Ziel: bestimmte sprachliche Phänomene aufzuzeigen oder bestehende Theorien zu falsifizieren.

# Das Korpus/Corpus

- Das Korpus/Corpus
- „[e]ndliche Menge von konkreten sprachlichen Äußerungen, die als empirische Grundlage für sprachwiss. Untersuchungen dienen.“ (Bußmann)
- eine Sammlung bestehend aus Texten mündlicher und/oder schriftlicher Äußerungen
- wurde nach einem bestimmten Verfahren mit einem bestimmten Zweck erstellt
- Beispiele: Zeitungsarchive, Lehrkorpora zu Übungszwecken, Sammlung von Hörtexten oder Hörbeispielen, das Internet, **sehr große digitale Datenbanken**

# Beispiele von sehr großen digitalen Datenbanken in der deutschen Sprache

- Leipzig
wortschatz.uni-leipzig.de

- Cosmas II
www.ids-mannheim.de

- Tiger
www.tigersearch.de

- DWDS
www.dwds.de


DWDS

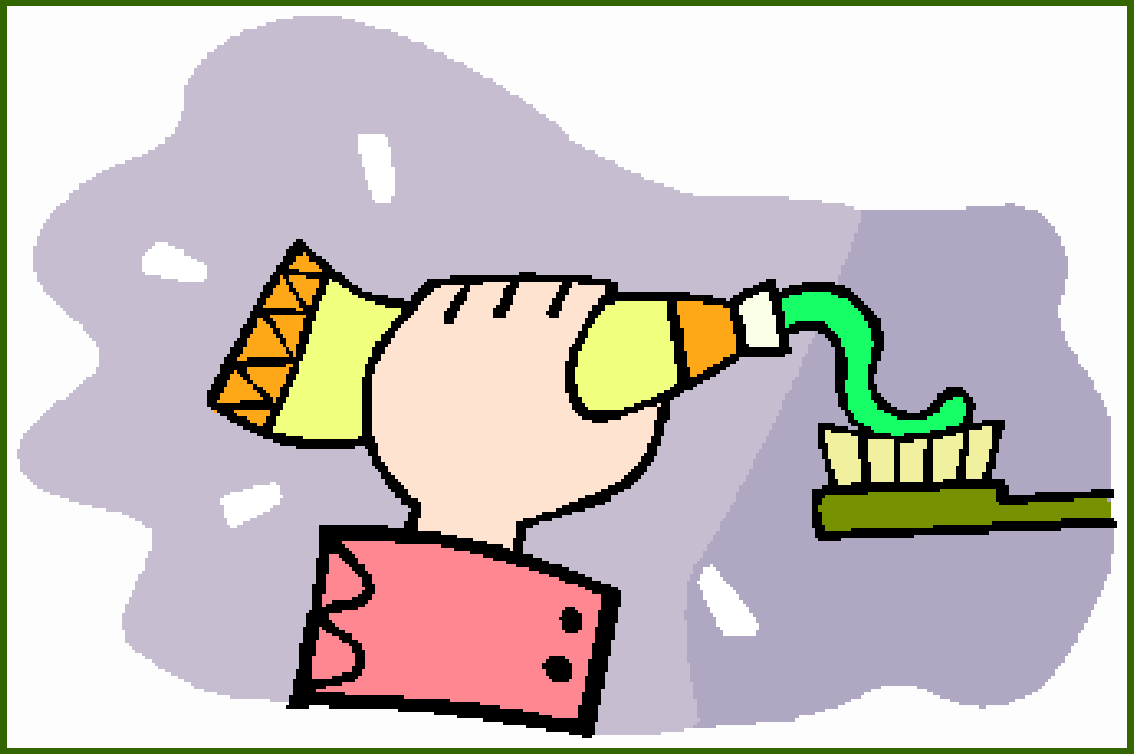
# Benützung:

- Lexik: Herstellung und Verbesserung von Wörterbüchern
- Semantik: Hilfe mit Fremdsprachenlernen
- Linguistik allgemein: sprachwissenschaftliche Untersuchungen,
- statistische Kollokationsbestimmung

# Der Kollokationsbegriff und Korpuslinguistik

- eingeführt von J.R. Firth von der Londoner Schule (Kontextualismus)
- seit den 80er und 90er Jahren sind Kollokationen in der Linguistik relevant geworden.

# die Zähne reinigen?



# Sind die rechten Ausdrücke falsch?

- die Zähne putzen
- einen Vortrag halten
- den Faden verlieren
- eine Entscheidung fällen
- to commit murder

?die Zähne bürsten

?einen Vortrag geben

?sie verliert schnell einen Faden

?eine Auswahl/Wahl fällen

?to perform murder

# Was sind Kollokationen?

- Terminus für charakteristische, häufig auftretende Wortverbindungen, deren Miteinandervorkommen auf einer Regelhaftigkeit gegenseitiger Erwartbarkeit beruht, also primär semantisch  (nicht grammatisch) begründet wird. (Bußmann)
- besteht aus zwei oder mehreren Wörtern und drückt einen Inhalt aus
- Bestandteile nicht substituierbar und nicht modifizierbar
- auch Begriff für das Miteinandervorkommen von Wörtern

# Unterschiedlicher Gebrauch des Begriffes

- Kollokationsbegriff beschreibt eine heterogene Sammlung von Wortverbindungen
- semantische, syntaktische und **statistische** Kriterien zur Ausgrenzung K. gegenüber anderen Wortarten
- freie Wortverbindungen bis Idiome
- benachbarte Konzepte: Sprichwörter (Wer nicht wagt, der nicht gewinnt), Floskeln,  (soviel ich weiß, ...) Idiome (an den Nagel hängen), Funktionsverbgefüge (zum Ausdruck kommen)

# Wichtiger Bestandteil der Sprache

- **wichtiger Bestandteil der Sprache**
- **vielleicht 70% von Sprachäußerungen in einem Korpus bestehen aus kombinatorischen wiederkehrenden Wortverbindungen (Altenberg)**
- **obwohl Möglichkeit einer unbegrenzten Kombination von einfachen linguistischen Einheiten (Chomsky) existiert, verwenden Sprecher etablierte Mehrwortverbindungen**

# Kollokationen in der computergestützten Korpuslinguistik

- empirisch bestimmen, wie oft gewisse Wörter in Kombination mit anderen auftreten (vs. Intuition)
- je mehr Sprachäußerungen und je variabler die Kontexte in dem Korpus, desto besser die Ergebnisse für empirische Untersuchungen
- maschinelle Verarbeitung von sehr großen Textsammlungen vor den 90er Jahren kaum durchführbar

# Statistische Kollokationsbestimmung

- statistisch gesehen sind Kollokationen Wörterkombinationen, die signifikant häufiger in einem Text miteinander vorkommen (Kookkurenz), als vom Zufall erwartet werden kann
- statistische Methoden werden angewendet, um eine Liste von Kollokationskanditaten aus einem Korpus zu erzeugen
- dazu braucht man durchsuchbare Textsammlungen: Text Mining, POS-Tagging, Konkordanzprogramme

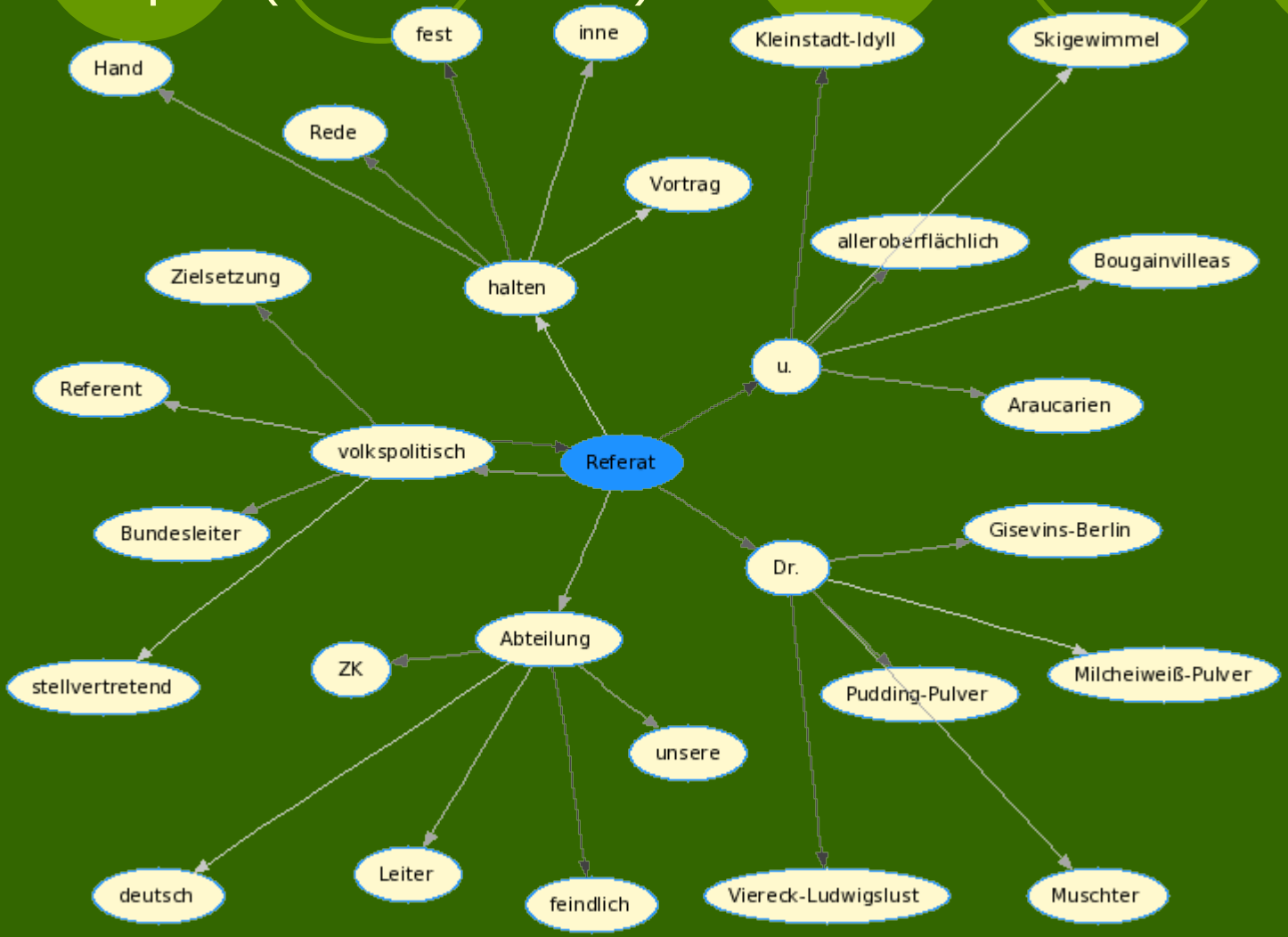
# DWDS: Das digitale Wörterbuch der deutschen Sprache des 20. Jh.

- DWDS-Kerncorpus: zeitlich und nach Textsorten ausgewogenes Corpus des gesamten 20. Jahrhunderts.
- Umfang: 100 Millionen Textwörter (tokens) in 79.830 Dokumenten.
- Textgrundlage: zur Bibliographischen Datenbank der Corpus-Texte
- lemmatisiert, mit Wortartinformationen versehen und mit einer linguistischen Suchmaschine abfragbar
- annotiert gemäß XML/TEI.
- eingebaute statistische Verfahren, um Kollokationen automatisch zu generieren
- http://www.dwds.de/cgi-bin/rest/loginstart?&kompakt=1&qu=Gefühl

# Automatisch berechnete Kollokationen

| | |
|---|---|
| #w1 | Suchbegriff (Kollokant) |
| F(w1) | Frequenz des Kollokants |
| W2 | Kollokat (node) |
| F(w2) | Frequenz des Kollokats |
| F(w1,w2) | Frequenz des Bigrams |
| MI | Mutual Information |
| T-Score | T-Score |
| Log-L. | Log-Likelihood |

# Automatisch berechnete Kollokationen aus dem DWDS Kerncorpus (lemmabasiert)



# Statistik als erster Schritt der Korpuslinguistik

- nur Kookurrenz und Kollokations-Kandidaten
- empirische Basis der Recherche
- qualitative linguistische Analyseverfahren notwendig: syntaktische und semantische Untersuchungen
- bestimmte sprachliche Phänomene aufzuzeigen oder bestehende Theorien zu falsifizieren

# Literatur:


- Anderman, Gunilla M. [Hrsg.] : Incorporating corpora : the linguist and the translator / ed. by Gunilla Anderman ... . - Clevedon [u.a.] : Multilingual Matters , 2008 .
- Bahns, Jens : Kollokationen als lexikographisches Problem : eine Analyse allgemeiner und spezieller Lernerwörterbücher des Englischen / Jens Bahns . - Tübingen : Niemeyer , 1996 .
- Dietrich, Wolf (Philologe) [Hrsg.] : Lexikalische Semantik und Korpuslinguistik / Wolf Dietrich ... (Hrsg.) . - Tübingen : Narr , 2006 .
- Lehr, Andrea : Kollokationen und maschinenlesbare Korpora : ein operationales Analysemodell zum Aufbau lexikalischer Netze / Andrea Lehr . - Tübingen : Niemeyer , 1996 . Porzig, Walter : Das Wunder der
- Ludewig, Petra : Korpusbasiertes Kollokationslernen : Computer-Assisted Language Learning als prototypisches Anwendungsszenario der Computerlinguistik / Petra Ludewig . - Frankfurt a.M. ; Wien [u.a.] : Lang , 2005 .
- Reder, Anna. Kollokationen in der Wortschatzarbeit . Wien: Praesens-Verl. 2006
- Sprache : Probleme, Methoden und Ergebnisse der Sprachwissenschaft / Walter Porzig. Hrsg. von Andreas Jecklin und Heinz Rupp . - 9. Aufl., unveränd. Nachdr. der 8. Aufl. . - Tübingen [u.a.] : Francke , 1993 .

# Konkordanz

- Stefanie Preiner

- PS Computerlinguistik
  - WiSe 08/09

# Gliederung des Referats

**Was ist Konkordanz?**

- Begriffsabgrenzung
- Definitionsversuch

**Beispiele von Konkordanzen**

- Konkordanzen in vordigitalen Zeiten
- Erste computergestützte Auswertungen
- Internet: Shakespeares gesammelte Werke

**Erstellung von Konkordanzen mit dem DWDS**

- Konkordanzprogramm DDS
- Das Stuttgart-Tübingen Tagset (STTS)

# Was ist Konkordanz?

„Konkordanz (lat.mlat.) *die*,., -en: 1. a) alphabetisches Verzeichnis von Wörtern od. Sachen zum Vergleich ihres Vorkommens u. Sinngehaltes an verschiedenen Stellen eines Buches (bes. als Bibelkonkordanz); b) Vergleichstabelle von Seitenzahlen verschiedener Ausgaben eines Werkes. 2. gleichlaufende Lagerung mehrerer Gesteinsschichten übereinander (Geol.) 3. die Übereinstimmung in Bezug auf ein bestimmtes Merkmal (z.B. von Zwillingen; Biol.). 4. ein Schriftgrad (Maßeinheit von 4 Cicero; Druchw.). 5. (in bestimmten Sprachen) Ausdruck grammatischer Zusammenhänge durch formal gleiche Elemente bes. durch Präfixe (Sprachw.). “

DUDEN – Das Fremdwörterbuch. Mannheim: Brockhaus 1997, S. 437.

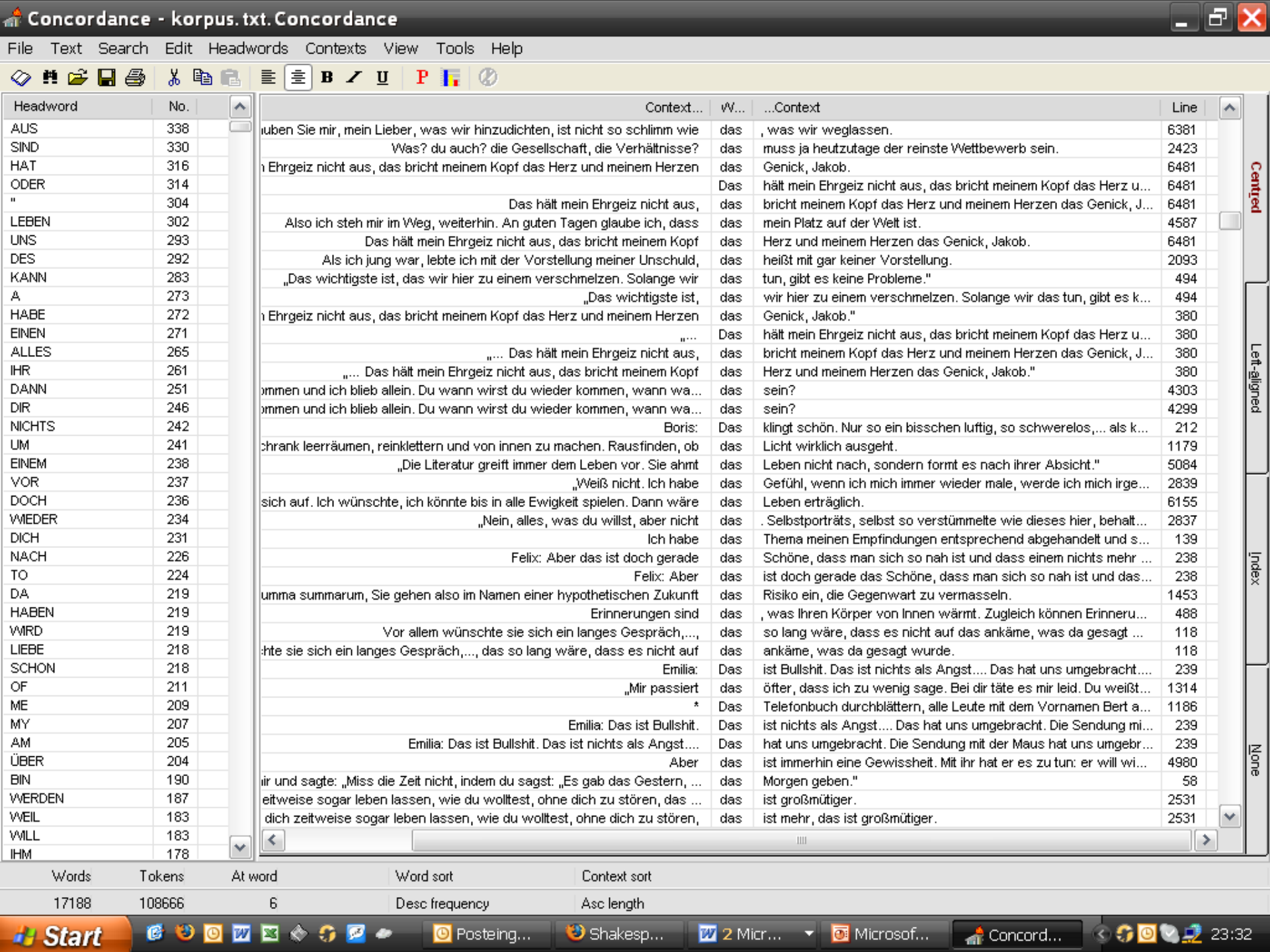
# Was ist Konkordanz?

„Concordance: A list of words, normally in alphabetical order, where each occurrence of each word is shown with surrounding context and identified by a reference indicating where it occurs in the text.“

(Lawler / Dry 1998, S. 259)

# Darstellungsmöglichkeiten

- Index
- Wortliste
- Konkordanz


Concordance - korpus.txt.Concordance
File Text Search Edit Headwords Contexts View Tools Help
Headword No.
AUS 338
SIND 330
HAT 316
ODER 314
" 304
LEBEN 302
UNS 293
DES 292
KANN 283
A 273
HABE 272
EINEN 271
ALLES 265
IHR 261
DANN 251
DIR 246
NICHTS 242
UM 241
EINEM 238
VOR 237
DOCH 236
WIEDER 234
DICH 231
NACH 226
TO 224
DA 219
HABEN 219
WIRD 219
LIEBE 218
SCHON 218
OF 211
ME 209
MY 207
AM 205
ÜBER 204
BIN 190
WERDEN 187
WEIL 183
WILL 183
IHM 178
Context... W... ...Context Line
uben Sie mir, mein Lieber, was wir hinzudichten, ist nicht so schlimm wie das , was wir weglassen. 6381
Was? du auch? die Gesellschaft, die Verhältnisse? das muss ja heutzutage der reinste Wettbewerb sein. 2423
Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, Jakob. 6481
Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz u... 6481
Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, J... 6481
Also ich steh mir im Weg, weiterhin. An guten Tagen glaube ich, dass das mein Platz auf der Welt ist. 4587
Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, Jakob. 6481
Als ich jung war, lebte ich mit der Vorstellung meiner Unschuld, das heißt mit gar keiner Vorstellung. 2093
„Das wichtigste ist, das wir hier zu einem verschmelzen. Solange wir das tun, gibt es keine Probleme." 494
„Das wichtigste ist, das wir hier zu einem verschmelzen. Solange wir das tun, gibt es k... 494
Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, Jakob." 380
„... Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz u... 380
„... Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, J... 380
„... Das hält mein Ehrgeiz nicht aus, das bricht meinem Kopf das Herz und meinem Herzen das Genick, Jakob." 380
mmen und ich blieb allein. Du wann wirst du wieder kommen, wann wa... das sein? 4303
mmen und ich blieb allein. Du wann wirst du wieder kommen, wann wa... das sein? 4299
Boris: Das klingt schön. Nur so ein bisschen luftig, so schwerelos,... als k... 212
chrank leerräumen, reinklettern und von innen zu machen. Rausfinden, ob das Licht wirklich ausgeht. 1179
„Die Literatur greift immer dem Leben vor. Sie ahmt das Leben nicht nach, sondern formt es nach ihrer Absicht." 5084
„Weiß nicht. Ich habe das Gefühl, wenn ich mich immer wieder male, werde ich mich irge... 2839
sich auf. Ich wünschte, ich könnte bis in alle Ewigkeit spielen. Dann wäre das Leben erträglich. 6155
„Nein, alles, was du willst, aber nicht das . Selbstporträts, selbst so verstümmelte wie dieses hier, behalt... 2837
Ich habe das Thema meinen Empfindungen entsprechend abgehandelt und s... 139
Felix: Aber das ist doch gerade das Schöne, dass man sich so nah ist und dass einem nichts mehr ... 238
Felix: Aber das ist doch gerade das Schöne, dass man sich so nah ist und das... 238
umma summarum, Sie gehen also im Namen einer hypothetischen Zukunft das Risiko ein, die Gegenwart zu vermasseln. 1453
Erinnerungen sind das , was Ihren Körper von Innen wärmt. Zugleich können Erinneru... 488
Vor allem wünschte sie sich ein langes Gespräch,..., das so lang wäre, dass es nicht auf das ankäme, was da gesagt ... 118
hte sie sich ein langes Gespräch,..., das so lang wäre, dass es nicht auf das ankäme, was da gesagt wurde. 118
Emilia: Das ist Bullshit. Das ist nichts als Angst.... Das hat uns umgebracht.... 239
„Mir passiert das öfter, dass ich zu wenig sage. Bei dir täte es mir leid. Du weißt... 1314
* Das Telefonbuch durchblättern, alle Leute mit dem Vornamen Bert a... 1186
Emilia: Das ist Bullshit. Das ist nichts als Angst.... Das hat uns umgebracht. Die Sendung mi... 239
Emilia: Das ist Bullshit. Das ist nichts als Angst.... Das hat uns umgebracht. Die Sendung mit der Maus hat uns umgebr... 239
Aber das ist immerhin eine Gewissheit. Mit ihr hat er es zu tun: er will wi... 4980
ir und sagte: „Miss die Zeit nicht, indem du sagst: „Es gab das Gestern, ... das Morgen geben." 58
eitweise sogar leben lassen, wie du wolltest, ohne dich zu stören, das ... das ist großmütiger. 2531
dich zeitweise sogar leben lassen, wie du wolltest, ohne dich zu stören, das ist mehr, das ist großmütiger. 2531
Centred
Left-aligned
Index
None
Words Tokens At word Word sort Context sort
17188 108666 6 Desc frequency Asc length
Start
Posteing...
Shakesp...
2 Micr...
Microsof...
Concord...
23:32

# Beispiele von Konkordanzen

- Konkordanzen in vordigitalen Zeiten
- Erste computergestützte Auswertungen
- Internet: Shakespeares gesammelte Werke

# Erstellung von Konkordanzen mit dem DWDS

- Beispiel: Konkordanz von „Lehrer“
- Das Stuttgart-Tübingen Tagset

# Text Mining und Informationsextraktion

Christian Wimplinger

# Gliederung

- Text Mining – Begriffserklärung
- Volltextanalyse
- Analyse von Einzeltexten
- Merkmalsextraktion
- Beispiel für Merkmalsextraktion
- Zusammenfassung

# Text Mining – Anwendung

- Aus Textquantität informative Qualität herausfiltern
- Textmaterial ist unstrukturiert
  - Im Gegensatz zu Data Mining: untersucht strukturierte Datenbanken
- Textimmanent und intertextuelle Anwendung
  - Textimmanent: Schlüsselwortextraktion, automatische Textzusammenfassung, etc.
  - Intertextuell: Clusterbildung von semantisch zusammengehörender Texte

# Volltextanalyse

- Funktioniert durch Indexbildung
  - (Fast) jedes Wort ein Indexterm
  - Normalisierung der Wörter
  - Stoppworte ausfiltern
- Bei Anfrage wird die Textposition angegeben
  - Komplexe Suche durch Retrievalmodelle

# Analyse von Einzeltexten

- Vorbereitung zur besseren Informationsauswertung
  - Textrelevante Daten werden in eine einheitliche Form gebracht.
    - Erkennung des Datenformats
    - Erkennung der Zeichenkodierung
    - Hinweis auf Textstruktur
    - Erkennung der Sprache
- Erstellung einer Merkmalsmatrix

# Merkmalsextraktion

- Tokenisierung – Zerlegung in Worte
  - Satzgrenzen erkennen
  - Mittels Wörterbücher rückführen auf Stamm
  - Datumsangaben durch Algorithmen ersetzen
  - Akronyme entschlüsseln
- Semantische Probleme

# Beispiel für Merkmalsextraktion

PersonOut

PersonIn

Position

Organisation

TimeOut

TimeIn

# Zu verarbeitender Text

„Dr. Hermann Wirth, bisheriger Leiter der Musikhochschule München, verabschiedet sich heute aus dem Amt. Der 65-jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde Sabine Klinger benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.“

# Ausgabe der Merkmalsextraktion

| | |
|---|---|
| PersonOut | Dr. Hermann Wirth |
| PersonIn | Sabine Klinger |
| Position | Leiter |
| Organisation | Musikhochschule München |
| TimeOut | Heute |
| TimeIn | |

# AUS

## Literatur:

Dörre, Jochen/Gerstl, Peter/Seiffert, Roland: Volltextsuche und Text Mining. In: *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Hrsg. V. Carstensen/Ebert, u.a., 2., überarbeitete und erweiterte Auflage. München: Elsevier GmbH (2004).