



DH 1: Sprache und Text

Korpora und Annotation

Andreas Blombach, Stephanie Evert

Lehrstuhl für Korpus- und
Computerlinguistik

<https://www.linguistik.phil.fau.de>



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie



Korpuslinguistik ...

Universität Erlangen-Nürnberg • Postfach 3520 • 91023 Erlangen

Prof. Dr. Stefan Evert
Professur für Körperlinguistik
Bismarckstraße 6
91054 Erlangen



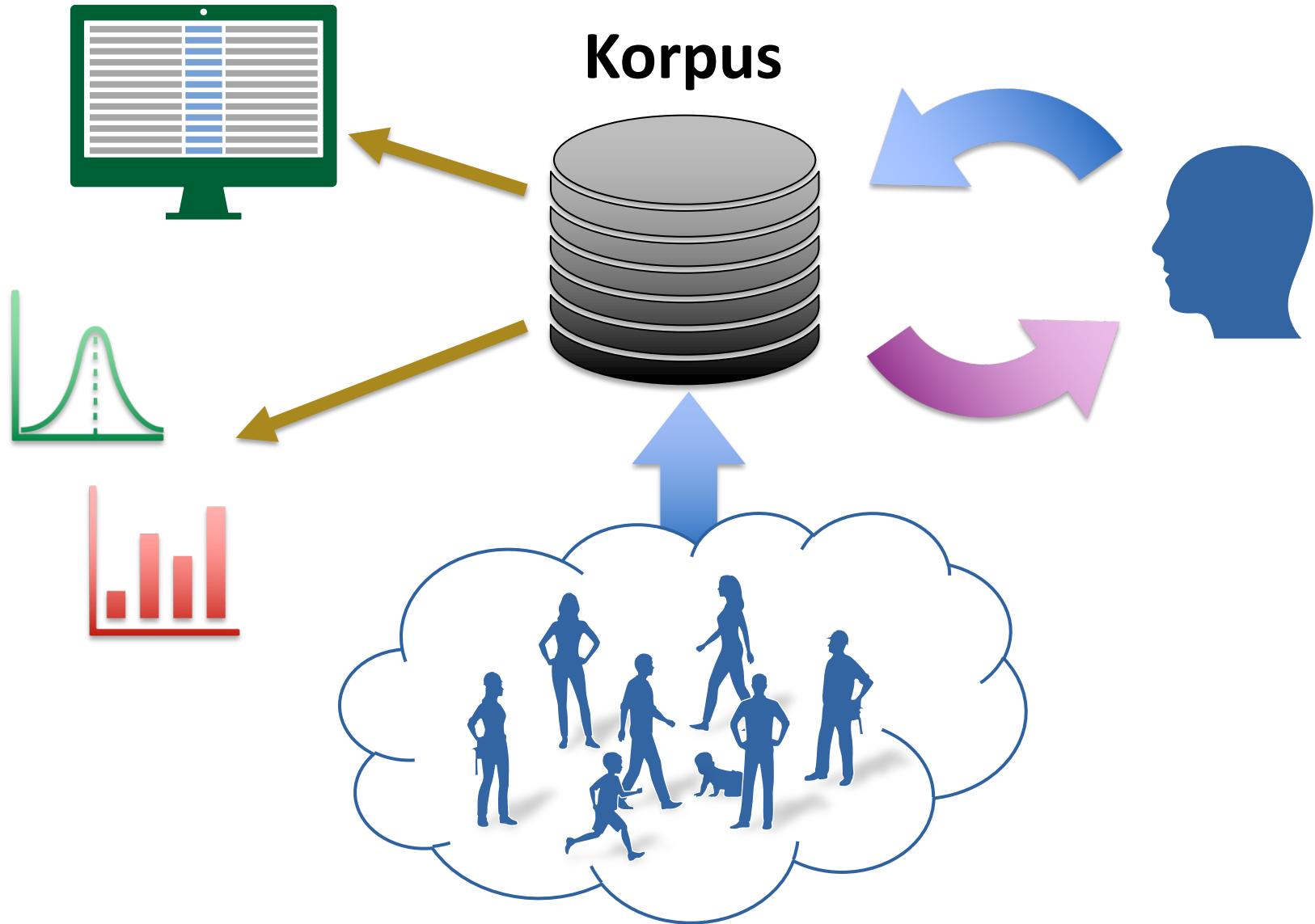
Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Was ist ein Korpus?

- Korpus im weiten Sinn = **Textarchiv** ([electronic text collection](#))
 - Sammlung beliebiger maschinenlesbarer Texte
 - oft opportunistisch → möglichst viel verfügbares Material
 - kann heterogen sein → z.B. Oxford Text Archive, DeReKo
- **vs.** Korpus als **repräsentative Stichprobe** der Sprachwirklichkeit
 - Designkriterien ([sampling frame](#))
 - zufällige Stichprobe auf Basis der Designkriterien ([random sample](#))
 - strebt **Repräsentativität** (für Sprachvarietät oder Teilsprache) an
 - einheitliche Metadaten, Annotation, Repräsentationsformat
- Achtung: *das Korpus, die Korpora*

Empirische Basis der Sprachwissenschaft

- Introspektion
 - bestenfalls als Anstoß zur Theoriebildung geeignet
- Sprecherurteile
 - z.B. Grammatikalität, Plausibilität, ... einer Äußerung
- Psycholinguistische Experimente
 - Überprüfung von spezifischen Hypothesen
 - unter Laborbedingungen
- Korpora
 - authentischer Sprachgebrauch
 - skalierbar: breite Abdeckung von Phänomenen in großen Korpora



Ein Beispiel: Kollokationen

- Kollokationen sind übliche Wortverbindungen
 - nicht nur Redewendungen (ins Wasser fallen)
 - oft kompositionell verständlich, aber nicht vorhersagbar

■ *Zähne* ...

bürsten

putzen

waschen

reinigen

■ *Kritik* ...

erheben

anführen

aussprechen

üben

■ *Kritik ist* ...

schwer

scharf

streng

stark

■ *als Streithelfer* ...

mitmachen

eintreten

beitreten

hinzukommen

| ↔ :: ◻ ✖ | | | | |
|--|--------|-------|-----|---|
| modifiers of "Kritik" | | | | |
| konstruktiv | 21,880 | 10.24 | ... | |
| scharf | 20,061 | 9.72 | ... | |
| heftig | 19,471 | 9.7 | ... | |
| harsch | 8,941 | 9.45 | ... | |
| berechtigt | 10,278 | 9.18 | ... | |
| massiv | 11,061 | 8.6 | ... | |
| geäußert | 5,088 | 8.56 | ... | |
| sachlich | 4,366 | 7.98 | ... | |
| vernichtend | 2,675 | 7.7 | ... | |
| negativ | 7,454 | 7.69 | ... | |
| herb | 2,401 | 7.42 | ... | |
| radikal | 2,083 | 6.85 | ... | |
| ▼ | | | | ▼ |
| ↔ :: ◻ ✖ | | | | |
| verbs with "Kritik" as accusative object | | | | |
| üben | 34,203 | 11.02 | ... | |
| äußern | 13,175 | 9.77 | ... | |
| geraten/raten | 6,414 | 9.06 | ... | |
| hageln | 4,009 | 8.79 | ... | |
| einstecken | 4,350 | 8.74 | ... | |
| ernten | 4,111 | 8.48 | ... | |
| weisen | 5,328 | 7.76 | ... | |
| anbringen | 2,351 | 7.64 | ... | |
| vertragen | 1,792 | 7.37 | ... | |
| formulieren | 2,224 | 7.34 | ... | |
| stoßen | 3,094 | 7.32 | ... | |
| nachvollziehen | 1,808 | 7.32 | ... | |
| ▼ | | | | ▼ |
| ↔ :: ◻ ✖ | | | | |
| genitive objects of "Kritik" | | | | |
| Vernunft | 5,230 | 10.27 | ... | |
| Ökonomie | 3,615 | 9.97 | ... | |
| Urteilskraft | 1,147 | 8.71 | ... | |
| Opposition | 1,362 | 8.33 | ... | |
| Kant | 703 | 7.81 | ... | |
| Relativitätstheorie | 419 | 7.2 | ... | |
| Rechnungshof | 430 | 7.18 | ... | |
| Verhältnis | 632 | 6.92 | ... | |
| Kapitalismus | 567 | 6.92 | ... | |
| Kritik | 490 | 6.85 | ... | |
| Rechtsphilosophie | 282 | 6.71 | ... | |
| Bundesrechnungshof | 260 | 6.55 | ... | |
| ▼ | | | | ▼ |
| ↔ :: ◻ ✖ | | | | |
| nouns with "Kritik" as genitive object | | | | |
| Kreuzfeuer | 2,614 | 10.62 | ... | |
| Ökonomie | 923 | 8.65 | ... | |
| Vernunft | 654 | 8.06 | ... | |
| Zielscheibe | 382 | 7.83 | ... | |
| Fokus | 887 | 7.64 | ... | |
| Kernpunkt | 321 | 7.42 | ... | |
| Grundriss | 327 | 7.35 | ... | |
| Gegenstand | 1,383 | 7.34 | ... | |
| Welle | 379 | 6.79 | ... | |
| Kern | 610 | 6.77 | ... | |
| Zentrum | 1,465 | 6.71 | ... | |
| Mittelpunkt | 1,088 | 6.63 | ... | |
| ▼ | | | | ▼ |

| ↔ :: ◻ ✖ | | | | |
|--------------------------------|---------|--------|-----|---|
| verbs with "Kritik" as subject | | | | |
| üben | 2,376 | 8.75 | ... | |
| richten | 4,615 | 7.83 | ... | |
| entzünden | 551 | 7.62 | ... | |
| zielen | 967 | 7.56 | ... | |
| beziehen | 2,170 | 7.33 | ... | |
| ernten | 396 | 7.01 | ... | |
| ▼ | | | | ▼ |
| ↔ :: ◻ ✖ | | | | |
| "Kritik" and/or ... | | | | |
| Lob | 10,899 | 10.65 | ... | |
| Anregung | 19,032 | 10.52 | ... | |
| Verbesserungsvorschlag | 3,154 | 9.25 | ... | |
| Vorschlag | 3,044 | 8.65 | ... | |
| Publikum | 2,703 | 8.48 | ... | |
| Meinung | 2,816 | 8.14 | ... | |
| ▼ | | | | ▼ |
| ↔ :: ◻ ✖ | | | | |
| prepositional phrases | | | | |
| "Kritik" an + noun | 183,797 | 14.92% | ... | |
| "Kritik" von + noun | 23,131 | 1.88% | ... | |
| "Kritik" in + noun | 21,967 | 1.78% | ... | |
| noun + in "Kritik" | 18,080 | 1.47% | ... | |
| noun + auf "Kritik" | 12,929 | 1.05% | ... | |
| noun + zu "Kritik" | 12,372 | 1% | ... | |
| ▼ | | | | ▼ |

| modifiers of "Streithelfer" | | | | |
|-----------------------------|----|------|-----|--|
| beigetreten | 10 | 7.37 | ... | |
| vernommen | 9 | 7.09 | ... | |
| beklagt | 9 | 1.92 | ... | |
| hiesig | 8 | 0.94 | ... | |

| verbs with "Streithelfer" as accusative object | | | | |
|--|----|------|-----|--|
| beauftragen | 10 | 1.46 | ... | |

| genitive objects of "Streithelfer" | | | | |
|------------------------------------|-----|------|-----|--|
| Bekl | 5 | 5.21 | ... | |
| Beklagte | 116 | 4.14 | ... | |
| Klägerin | 76 | 4.14 | ... | |
| Kl. | 5 | 3.74 | ... | |
| Kläger | 50 | 3.26 | ... | |
| Berufung | 5 | 3.14 | ... | |
| Kosten | 15 | 1.19 | ... | |

| nouns with "Streithelfer" as genitive object | | | | |
|--|----|------|-----|--|
| Bösgläubigkeit | 11 | 9.23 | ... | |
| Nebenintervention | 5 | 8.33 | ... | |
| Zeugenaussage | 6 | 5.68 | ... | |
| Schriftsatz | 10 | 5.49 | ... | |
| Rechtsmittel | 6 | 4.69 | ... | |
| Prozessbevollmächtigte | 5 | 4.34 | ... | |
| Berufung | 24 | 4.31 | ... | |
| Vernehmung | 6 | 4.11 | ... | |
| Beitritt | 12 | 3.91 | ... | |
| Vorbringen | 7 | 3.86 | ... | |
| Befugnis | 7 | 3.82 | ... | |
| Verschulden | 7 | 3.73 | ... | |

| verbs with "Streithelfer" as subject | | | | |
|--------------------------------------|----|------|-----|--|
| beitreten | 6 | 5.86 | ... | |
| beantragen | 13 | 2.68 | ... | |
| behaupten | 8 | 0.63 | ... | |

| "Streithelfer" and/or ... | | | | |
|---------------------------|-----|------|-----|--|
| Hauptpartei | 7 | 9.7 | ... | |
| HABM | 8 | 9.19 | ... | |
| Kl. | 8 | 7.81 | ... | |
| Beklagte | 102 | 7.6 | ... | |
| Klägerin | 39 | 6.3 | ... | |
| Kläger | 50 | 6.15 | ... | |

| prepositional phrases | | | | |
|---------------------------|----|-------|-----|--|
| "Streithelfer" in + noun | 55 | 1.96% | ... | |
| noun + als "Streithelfer" | 49 | 1.74% | ... | |
| "Streithelfer" auf + noun | 29 | 1.03% | ... | |
| "Streithelfer" zu + noun | 28 | 1% | ... | |
| noun + von "Streithelfer" | 8 | 0.28% | ... | |
| noun + zu "Streithelfer" | 6 | 0.21% | ... | |

Spezialisiertes Korpus BGH

www.bundesgerichtshof.de

Kontakt und Anfahrt Inhaltsverzeichnis Impressum Datenschutz Deutsch Français English Portugiesisch Griechisch Leichte Sprache RSS-Feed



Bundesgerichtshof



Das Gericht

Wir heißen Sie beim Bundesgerichtshof zu Aufgaben und Organisation des Gerichts

Das Gericht



<text id="4_str_540-01" type="Beschluss" az="4 StR 540/01" date="2001-12-20" topic="wegen gefährlicher Körperverletzung u.a.">

<p>Der 4. Strafsenat des Bundesgerichtshofs hat nach Anhörung des Generalbundesanwalts und der Beschwerdeführerin am 20. Dezember 2001 gemäß § 349 Abs. 2 und 4 StPO beschlossen:

</p>

enum="1.">Auf die Revision der Angeklagten wird das Urteil des Landgerichts Saarbrücken vom 11. September 2001 im Maßregelausspruch mit den Feststellungen aufgehoben.

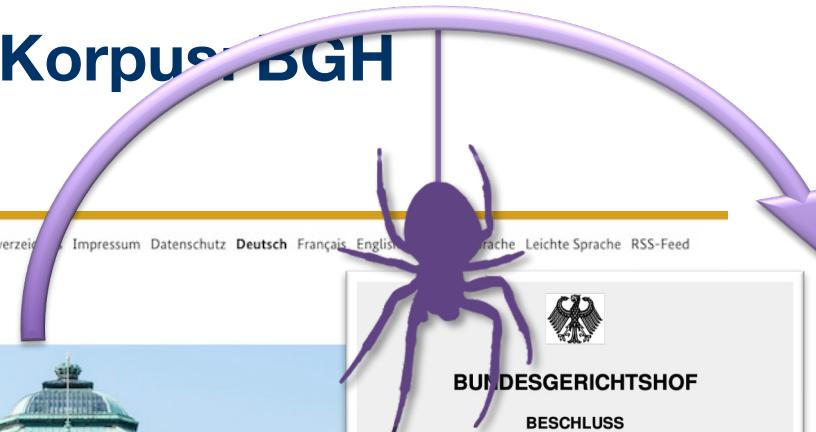
</p>

enum="2.">Im Umfang der Aufhebung wird die Sache zu neuer Verhandlung und Entscheidung, auch über die Kosten des Rechtsmittels, an eine Strafkammer des Landgerichts Landau zurückverwiesen.

</p>

...

</text>



- 2 -
Der 4. Strafsenat des Bundesgerichtshofs hat nach Anhörung des Generalbundesanwalts und der Beschwerdeführerin am 20. Dezember 2001 gemäß § 349 Abs. 2 und 4 StPO beschlossen:

1. Auf die Revision der Angeklagten wird das Urteil des Landgerichts Zweibrücken vom 11. September 2001 im Maßregelausspruch den Feststellungen aufgehoben.

2. Auf die Aufhebung wird die Sache zu neuer Verhandlung und Entscheidung, auch über die Kosten des Rechtsmittels, an eine Strafkammer des Landgerichts Landau zurückverwiesen.

gehende Revision wird verworfen.

Gründe:

Die Angeklagte hat die Angeklagte des "Widerstandes gegen Vollzug in Tateinheit mit gefährlicher Körperverletzung und verdeckter Körperverletzung, der Bedeutigung in Tateinheit mit Bedrohung: in 4 tateinheitlich zusammen treffenden Fällen] und schuldig befunden und sie unter Einziehung "des Urteils Zweibrücken [richtig: der Strafe aus dem Strafbefehl des Saarbrücken] vom 23.01.2001" zu einer Gesamtfreiheitsstrafe verurteilt und ihre Unterbringung in einem psychiatrischen Klinikum ordnet. Hiergegen wendet sich die Angeklagte mit ihrer Revision sachlichen Rechts rügt. Das Urteil hat um

Was ist Korpuslinguistik?

Korpuslinguistik als **Methodenwissenschaft**
(*corpus-based research*)

- Erstellung von Korpora als repräsentative Stichproben
(Grundgesamtheit, *sampling frame*, Stichprobenplanung, ...)
- Automatische linguistische Annotation
(schwierig z.B. für literarische oder historische Texte)
- Werkzeuge für Suche (*corpus query*) & Visualisierung
- Überprüfung von Hypothesen mit statistischen Tests
(→ methodologische Herausforderungen)
- Statistische Modellierung & explorative Verfahren
- Weitgehend unabhängig von spezifischen Fragestellungen

Was ist Korpuslinguistik?

Korpuslinguistik als eigene **Teildisziplin der Sprachwissenschaft**
(*corpus-driven research*)

- Britische Forschungstradition seit 1950er Jahren
(Firth, Sinclair, Halliday, Leech, McEnery, ...)
- Qualitative Analyse von Konkordanzen (John Sinclair: *trust the text!*)
- Quantitative Methoden: Keywords, Kollokationen, Häufigkeitsvergleich, ...
- Traditionell vorwiegend deskriptiv, Schwerpunkt auf Sprachgebrauch
(→ angewandte Korpuslinguistik, *Applied CL*)
- Themen: Sprachvarietäten, funktionale Aspekte, soziologische und politische Diskurse, Spracherwerb und -unterricht, ...
- Neu: enge Verbindung zu kognitiver Linguistik (*usage-based approaches*)

Korpuslinguistische Fragestellungen

- Welche Struktur hat (eine) Sprache?
 - Grammatiken (CGEL¹), Syntaxtheorie, Morphologie, ...
 - Wörterbücher (→ Computerlexikographie)
 - Sprachdokumentation (insb. bedrohte Sprachen), Typologie
- Wie funktioniert Sprachverarbeitung im Gehirn?
 - Psycholinguistik, Neurolinguistik
 - kognitive Linguistik (→ Linguistics Lab)
- Wie wird Sprache gelernt (und gelehrt)?
 - Spracherwerb bei Kindern (L1)
 - Fremdsprachenerwerb (L2, L3), Lernersprache
 - Sprachunterricht (→ CALL²)

¹ Quirk/Greenbaum/Leech/Svartvik (1985): A *Comprehensive Grammar of the English Language*

² Computer-Assisted Language Learning

Korpuslinguistische Fragestellungen

- Wie sehr variiert Sprache?
 - Varietäten & Dialekte (Aussprache, Vokabular, Häufigkeiten)
 - Soziolinguistik, individuelle Unterschiede zw. Sprechern
 - Registervariation (z.B. formell vs. informell), multivariate Analyse (Biber 1988)
- Wie hat sich eine Sprache entwickelt?
 - Sprachwandel (Grammatik, Vokabular, Bedeutungen)
 - historische Linguistik, Sprachevolution
- Wozu wird Sprache verwendet?
 - korpusbasierte Diskursanalyse (z.B. Populismus)
 - digitale Literaturwissenschaft (Stilometrie, Autorschaft, ...)
 - Übersetzungswissenschaft („Translationese“)



Eine kurze Geschichte der Korpuslinguistik



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Vorläufer der Korpuslinguistik

- Erste „korpus“basierte Untersuchungen ab Ende des 19. Jhd.
- Rechtschreibung, Stenographie & Lexikographie
 - Kaeding (1897): dt. Häufigkeitswörterbuch auf Basis eines Korpus von ca. 11 M Wörtern (von über 600 Mitarbeitern in über 100 Zählstellen im Deutschen Reich händisch erstellt!)
 - Murray (1879–1928): Belegsammlung (Karteikarten!) für das *Oxford English Dictionary*
- Spracherwerb
 - erste Longitudinalstudien ca. 1876–1926 (Elterntagebücher)
 - große Querschnittsstudien ca. 1927–1957
- Sprachdokumentation im Strukturalismus
 - Boas (1940), Firth (1930–1955), ...
- Fremdsprachenunterricht
 - Grundwortschatz, Sprachniveaus, Kollokationen (z.B. Palmer 1933)
- Vergleichende Sprachwissenschaft
 - Eaton (1940) zur Häufigkeit von Wortbedeutungen in NL, FR, DE, IT

Chomsky (1957): Generative Linguistik

- Rationalismus (Introspektion) gegen Empirismus (datenbasiert)
 - Korpuslinguistik: empirische Beschreibung sprachlicher Phänomene
 - Chomsky: Theorie muss Phänomene erklären & kognitiv plausibel sein
- Kompetenz vs. Performanz
 - Chomskys Kritik: Korpus gibt nur Performanz wieder, empirische Häufigkeiten sind für die Kompetenz nicht relevant
 - Gegenkritik: „armchair linguistics“ auf Basis zweifelhafter Beispiele
- Repräsentativität
 - Menschen können unendlich viele Sätze bilden, von denen jedes Korpus nur einen kleinen, endlichen Ausschnitt erfassen kann
 - Chomsky : „Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, ...“
- Lernbarkeit: Poverty of Stimulus
 - Korpus genügt nicht zum Spracherwerb, insb. fehlen Negativbeispiele

Korpuslinguistik als eigenständige Teildisziplin

- Humanities Computing ab 1950 (→ Digital Humanities)
 - Index Thomisticus (1949–) von Robert Busa & IBM
- Julland: Mechanolinguistik (kontrastive Korpora ab 1956)
 - → quantitative / mathematische Linguistik (Harris 1968)
- Korpuslinguistik ca. 1960–1980 als Gegenbewegung zu Chomsky
- Korpusbasierte Grammatiken
 - Survey of English Usage (SEU) ab 1960, Brown Corpus 1961–1963
- Britischer Kontextualismus (Firth & Sinclair)
 - Firth (1957) nach Malinowski & Jones, Sinclair (1991), COBUILD
 - Kollokation und Kolligation, „trust the text“
- Ab 1990 als Teildisziplin anerkannt, Methodologie wird zunehmend auch von anderen Teildisziplinen übernommen und umgekehrt



Korpora



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Arten von Korpora

- Textkorpora (geschrieben) **vs.** Sprachkorpora (gesprochen)
vs. multimodale/multimediale Korpora (z.B. NewsScape)
- Referenzkorpus **vs.** spezialisierte Korpora
- synchrone **vs.** diachrone Korpora
- statische **vs.** dynamische Korpora (→ Monitorkorpus)
- unannotiert **vs.** annotiert (Metadaten, linguistische Analysen)
- einsprachig **vs.** Parallelkorpus **vs.** vergleichbare Korpora
- Unterscheidung nach Größe: 10k – 1M – 100M – 10G – 1T
→ statistische Auswertung, Abwägung: Größe vs. Datenqualität

Wichtige Korpora (Englisch)

- Brown Corpus (Francis & Kucera 1964)
 - Amerikanisches Englisch, geschriebene Sprache, 1961
 - 500 Stichproben à 2000 Wörter aus 15 Textsorten
- Brown Family
 - Brown (AmE, 1961), LOB (BrE, 1961) – Frown (AmE, 1991),
FLOB (BrE, 1991) – BLOB (BrE, 1931), BE2006 (BrE, 2006)
- Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993)
 - ca. 3 Millionen Wörter AmE mit syntaktischen Analysen ([parse trees](#))
<http://www.natcorp.ox.ac.uk/>
- British National Corpus (Aston & Burnard 1998)
 - Britisches Englisch, 90% geschrieben / 10% gesprochen, um 1991
 - ca. 100 Millionen Wörter in 4048 Dateien (= Texte / Sammlungen)
- Gigaword Corpus (Linguistic Data Consortium (LDC), 2003)
 - über 1 Milliarde Wörter Zeitungstext (5th ed.: 4 Milliarden Wörter)

Wichtige Korpora (Deutsch)

<https://www.dwds.de/>

- DWDS (Digitales Wörterbuch der deutschen Sprache)
 - Kernkorpus des 20. Jahrhunderts (100 Mio. Wörter): balanciertes Korpus (Zeitabschnitte und Genres, aber nicht unproblematisch)
 - Zeitungskorpora (insb. *Zeit*): > 1,5 Mrd. Wörter
 - Web-Korpus (nach Anmeldung): > 3 Mrd. Tokens
 - diverse Spezialkorpora (Blogs, Untertitel, DDR-Korpus, ...)
- Korpora des Instituts für deutsche Sprache (IdS)
 - > 50 Mrd. Wörter, davon aber nur ca. 3 Mrd. auch getaggt
 - sehr viele deutsche, österr. und schweizer Regionalzeitungen
 - überreg. Zeitungen (*Zeit, SZ, taz, ...*), Zeitschriften (*Spiegel, Focus, Gala, ...*)
 - Wikipedia-Artikel und -Diskussionen
 - historische und literarische Texte, diverse kleinere Spezialkorpora
- **Problem:** Quelldaten nicht verfügbar, nur eingeschränkter Online-Zugriff

Sprachübergreifende Korpora

<http://wacky.sslmit.unibo.it/doku.php?id=corpora>

- WaCky-Korpora (Baroni et al. 2009)
 - jeweils ca. 2 Milliarden Wörter aus automatisch gecrawlten Webseiten für Deutsch, Englisch, Französisch und Italienisch
- COW-Webkorpora (Schäfer & Bildhauer 2012)
 - aktuelle Fassung von 2014/2016
 - je 5–10 Milliarden Wörter für Französisch, Spanisch, Niederländisch und Schwedisch; für Deutsch und Englisch über 20 bzw. über 16 Mrd. Wörter
- Google Books N-Grams (Lin et al. 2012)
 - Häufigkeitszählungen für N-Gramme (bis N=5) aus gescannten Büchern in Englisch, Deutsch, Französisch, Spanisch, Chinesisch, ...
- Europarl (Koehn 2005)
 - Parallelkorpus aus Debatten des Europäischen Parlaments
 - aktuell: Release 7 mit je 10–60 Millionen Wörtern in 21 Sprachen
 - auf Satzebene aligniert, ursprünglich für statistische MÜ gedacht

Spezialkorpora (kleine Auswahl)

- DGD: Datenbank für Gesprochenes Deutsch
 - multimodal: Originalaufnahmen und Transkripte
- Referenzkorpus Altdeutsch
- CHILDES
 - aufgezeichnete und transkribierte Interaktionen aus Studien zum kindlichen Spracherwerb; diverse Sprachen (Englisch, Deutsch, Japanisch, Chinesisch, ...)
- Trinity Lancaster Corpus (Gablasova, Brezina & McEnery 2019)
 - gesprochene Sprache von Englischlernern aus versch. Ländern, ca. 4 M Wörter
- GermaParl (Blätte & Blessing 2018)
 - Plenarprotokolle des Bundestags (1996–2016, bald 1949–2021)
- DROC: Deutsches Romankorpus (Krug et al. 2018)
- DraCor: Drama Corpora Project
 - Figuren annotiert, dadurch z.B. für Netzwerkanalyse geeignet



Korpusstudien



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Grundprinzip empirischer Forschung

- am Anfang: **Fragestellung** – z.B. soll irgendeine **Anfangsbeobachtung** untersucht und erklärt werden
- **Theoriebildung** (unter Einbezug verfügbarer Literatur)
- Ableitung überprüfbarer **Hypothesen** aus dieser Theorie
- Identifizierung der **Variablen** und geeigneter Messmethoden
- **Datensammlung** zur Überprüfung der Hypothese(n):
Messen der Variablen
- **Datenanalyse**
- **Ablehnung oder Bestätigung** der Hypothese(n)
- ggf. Anpassung der Theorie oder neue Theorie



Arbeitsschritte einer Korpusstudie

1. Operationalisierung
 - Hypothesen, Festlegung der Grundgesamtheit (*sampling frame*)
2. Korpuserstellung
 - Auswahl von Texten, Digitalisierung / Konvertierung
 - Erfassung von Metadaten, juristische Fragen
3. Linguistische Annotation
 - manuelle Annotation: graphische Editoren, *annotator agreement*
 - automatische Annotation mit NLP¹-Werkzeugen: Qualität?
4. Repräsentationsformat (→ Standards)
 - Unicode, XML, TEI, XCES, ISO 24610–24627 (TC 37/SC 4), ...

¹ Natural Language Processing (dt.: Computerlinguistik, maschinelle Sprachverarbeitung)

Arbeitsschritte einer Korpusstudie

5. Indexierung und Abfrage

- Indexierung: effiziente Spezialformate
- Suche nach Stichwörtern, lexiko-grammatischen Mustern, ... ([query](#))
- Darstellung als Konkordanz („[kwic](#)“)
- Sortieren & Gruppierung

Your query "[word="what"%c & !bound(s)] "a"%c [pos="JJ.*"]+ "night"%c" returned 18 matches in 15 different texts (in 98,511,777 words [9,802 texts]; frequency: 0.18 instances per million words)
[4.616 seconds]

| Solution 1 to 18 Page 1 / 1 | | |
|---|--|--|
| My dearJarmila , | what a great night | for you . |
| oes crazy on a hot night , and maybe that's | what a hot night | is for . |
| Wow , | what a miserable night | . |
| tee that you 'll have a ball Come one and all | What a great night | you 've got in store You 'll wanna keep comi |
| e Hey , everyone , let 's go on with the show | What a great night | you 've got in store I 'll bet you 'll wanna kee |
| I am glad on ^ t. | What a fearful night | is this ! |
| Oh , dear , | what a terrible night | . |
| it 's gone Love goes on and on Oh , Robin , | what a beautiful night | . |
| Crimson morning skyline Whoa oh | What a weird night | , huh ? |
| It 's a wonder | what a good night | 'sleep will do for you . |
| When they think it 's sunset and see | what a nice night | it is , they 'll muster in the lobby . |
| God , | what a beautiful night | , Jack . |
| God , | what a beautiful night | , huh ? |

6. Quantitative Auswertung

- statistische Verfahren für Häufigkeitsvergleich (auch: Keywords, Kollokationen)
- explorative Datenanalyse & Visualisierung

7. Interpretation

Brainstorming

Wie würden Sie eine Korpusstudie durchführen?

- über Romane aus dem 19. Jhd.
 - z.B. stilistische Ähnlichkeiten von Autoren oder Gattungen
 - z.B. Verwendung bestimmter Metaphern
- oder über Liedtexte in der Populärmusik
 - z.B. Untersuchung von Formelhaftigkeit



Annotation



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Ein Korpus besteht aus ...

- **Objektdaten** = Texte 
 - primärer Untersuchungsgegenstand
- **Metadaten** = Informationen über die Texte
 - Titel, Autor/in, Veröffentlichungsdatum, Textsorte, ...
 - Alter, Geschlecht, Bildungsstand, ... der Autoren
- **Typographie** & Textstruktur
 - Abschnitte, Überschriften, Schriftarten, Listen, ...
- **Annotation** = linguistische Interpretation 
 - einfach (Wortebene) vs. strukturiert (z.B. Syntax)
 - Voraussetzung für die Erschließung großer Korpora

Korpusannotation: Wortebene

- Jedem (laufenden) Wort wird eine Kategorie zugeordnet
→ **Tagging** (= Etikettierung)
 - Voraussetzung: Text muss in Wörter zerlegt sein
- **Tokenisierung**
 - **Token** = Wort, Zahl, Symbol (😎), Satzzeichen, ...
 - im Gegensatz zu **Typen** = verschiedene Wörter
- Kann schwieriger sein, als man vermuten würde ...

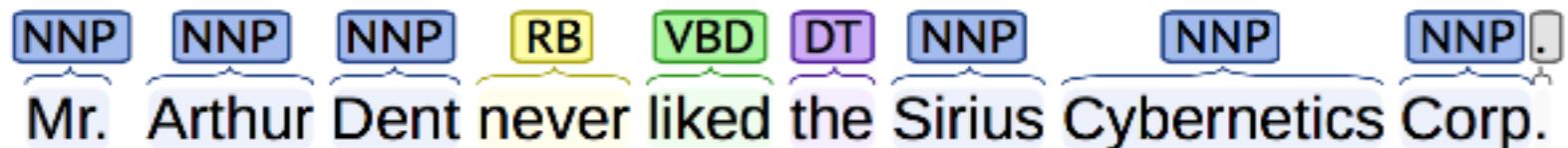
@Mia1234 #semibk [1] Das schließt direkt an die vorige Frage von @DieMaJa22 an. In jedem Fall gibt es (wie auch in der Sitzung ...

@Mia1234 #semibk [2] am BspChats gezeigt) starke Hinweise darauf, dass (wie auch imRealLife) diverse Faktoren die sprVariation beeinflussen: <http://tinyurl.com/3umxkuh>

<https://sites.google.com/site/empirist2015/> (Beißwenger et al. 2016)

Korpusannotation: Wortebene

- Zentral: Wortartenannotierung = **POS-Tagging**
 - Substantiv (**noun**), Adjektiv, Verb, Adverb, Pronomen, Präposition, Konjunktion, Zahl, Satzzeichen, ...
 - engl. POS = **part of speech**
- Tagset = Kategorienschema
 - meist feinere Unterschiede: Sg/Pl, inf./fin./imp., ...



- Weitere Annotationen: Lemmatisierung (CCL *to appear*), semantische Kategorien, emotionale Valenz, Schwierigkeitsgrad (CEFR), ...

Deutsch: STTS-Tagset

| | |
|----------------|---|
| ADJA | attributives Adjektiv |
| ADJD | adverbiales / prädikatives Adjektiv |
| ADV | Adverb <i>schon, bald, doch</i> |
| APPR | Präposition / Zirkumposition links |
| APPRART | Präposition mit Artikel fusioniert <i>zum</i> |
| APPO | Postposition <i>zufolge, wegen</i> |
| APZR | Zirkumposition rechts <i>von ... an</i> |
| ART | bestimmter oder unbestimmter Artikel |
| CARD | Kardinalzahlen (Ordinalzahl = ADJA) |
| FM | Fremdsprachliches Material |
| ITJ | Interjektion <i>mhm, ach, tja</i> |
| KOUI | unterordnende Konj. mit <i>zu</i> + Inf |
| KOUS | unterordnende Konjunktion mit Satz |
| KON | nebenordnende Konjunktion <i>und, oder</i> |
| KOKOM | Vergleichskonjunktion <i>als, wie</i> |
| NN | normales Nomen |
| NE | Eigenname |
| PDS | substituierendes Demonstrativpron. |
| PDAT | attribuierendes Demonstrativpron. |
| PIS | substituierendes Indefinitpron. |
| PIAT | attrib. Indefinitpron. ohne Determiner |
| PIDAT | attrib. Indefinitpron. mit Determiner |
| PPER | Personalpronomen (nicht reflexiv) |
| PPOS | substituierendes Possessivpronomen |
| PPOSAT | attribuierendes Possessivpronomen |
| PRELS | substituierendes Relativpronomen |
| PRELAT | attribuierendes Relativpronomen |

| | |
|---------------|--|
| PRF | reflexives Personalpronomen |
| PWS | substituierendes Interrogativpron. |
| PWAT | attribuierendes Interrogativpronomen |
| PWAV | adverbiales Interrogativ-/Relativpron. |
| PAV | Pronominaladverb <i>dafür, deswegen</i> |
| PTKZU | zu vor Infinitiv |
| PTKNEG | Negationspartikel <i>nicht</i> |
| PTKVZ | abgetrennter Verbzusatz <i>kommt ... an</i> |
| PTKANT | Antwortpartikel <i>ja, nein, danke</i> |
| PTKA | Partikel bei Adjektiv/Adverb <i>am, zu</i> |
| TRUNC | Kompositions-Erstglied <i>Unter- und ...</i> |
| VVFIN | finites Verb, voll (= lexikalisch) |
| VVIMP | Imperativ, voll |
| VVINF | Infinitiv, voll |
| VVIZU | Infinitiv mit <i>zu</i> , voll |
| VVPP | Partizip Perfekt, voll |
| VAFIN | finites Hilfsverb |
| VAIMP | Imperativ, Hilfsverb |
| VAINF | Infinitiv, Hilfsverb |
| VAPP | Partizip Perfekt, Hilfsverb |
| VMFIN | Finites Modalverb |
| VMINF | Infinitiv, Modalverb |
| VMPP | Partizip Perfekt, Modalverb |
| XY | Nichtwort mit Sonderzeichen <i>3:7, H2O</i> |
| \$, | Komma , |
| \$. | Satzbeendende Interpunktions . ? ! ; : |
| \$() | sonstige Satzzeichen (intern) - [] () |

Englisch: Penn-Tagset (modifiziert)



| | |
|----------------|---|
| CC | Coordinating conjunction |
| CD | Cardinal number |
| DT | Determiner |
| EX | Existential <i>there</i> |
| FW | Foreign word |
| IN | Preposition / subordinating conjunction |
| IN/that | Subordinating conjunction <i>that</i> |
| JJ | Adjective (positive) |
| JJR | Adjective (comparative) |
| JJS | Adjective (superlative) |
| LS | List item marker |
| MD | Modal verb |
| NN | Noun, singular or mass |
| NNS | Noun, plural |
| NP | Proper noun, singular |
| NPS | Proper noun, plural |
| PDT | Predeterminer |
| POS | Possessive ending ('s) |
| PP | Personal pronoun |
| PP\$ | Possessive pronoun |
| RB | Adverb |
| RP | Particle |
| SYM | Symbol (mathematical/scientific) |
| TO | <i>to</i> (any usage) <i>fly to Paris, ready to go, ...</i> |
| UH | Interjection |
| # | Pound sign £ |
| \$ | Dollar sign \$ |

| | | |
|-------------|--|-----------|
| VB | Verb <i>be</i> , base form | |
| VBD | Verb <i>be</i> , past tense | |
| VBG | Verb <i>be</i> , gerund/progressive | |
| VBN | Verb <i>be</i> , past participle | |
| VBP | Verb <i>be</i> , non-3rd pers. sg. present | |
| VBZ | Verb <i>be</i> , 3rd pers. sg. present tense | |
| VH | Verb <i>have</i> , base form | |
| VHD | Verb <i>have</i> , past tense | |
| VHG | Verb <i>have</i> , gerund/progressive | |
| VHN | Verb <i>have</i> , past participle | |
| VHP | Verb <i>have</i> , non-3rd pers. sg. present | |
| VHZ | Verb <i>have</i> , 3rd pers. sg. present tense | |
| VV | Lexical verb, base form | |
| VVD | Lexical verb, past tense | |
| VVG | Lexical verb, gerund/progressive | |
| VVN | Lexical verb, past participle | |
| VVP | Lexical verb, non-3rd pers. sg. present | |
| VVZ | Lexical verb, 3rd pers. sg. present tense | |
| WDT | Wh-determiner | |
| WP | Wh-pronoun | |
| WP\$ | Possessive wh-pronoun | |
| WRB | Wh-adverb | |
| SENT | Sentence-final punctuation | . ! ? |
| , | Comma | , |
| : | Colon, semi-colon | : ; |
| () | Comma | ([]) |
| `` '' | Comma | “ ” ‘ ’ ” |

| Open class words | Closed class words | Other |
|-----------------------|-----------------------|-----------------------|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

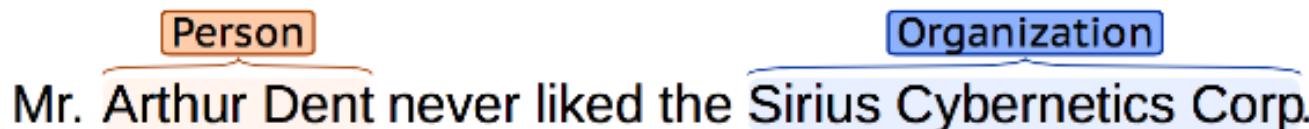
- [ADJ](#): adjective
- [ADP](#): adposition
- [ADV](#): adverb
- [AUX](#): auxiliary
- [CCONJ](#): coordinating conjunction
- [DET](#): determiner
- [INTJ](#): interjection
- [NOUN](#): noun
- [NUM](#): numeral
- [PART](#): particle
- [PRON](#): pronoun
- [PROPN](#): proper noun
- [PUNCT](#): punctuation
- [SCONJ](#): subordinating conjunction
- [SYM](#): symbol
- [VERB](#): verb
- [X](#): other

Korpusannotation: Segmente und Strukturen

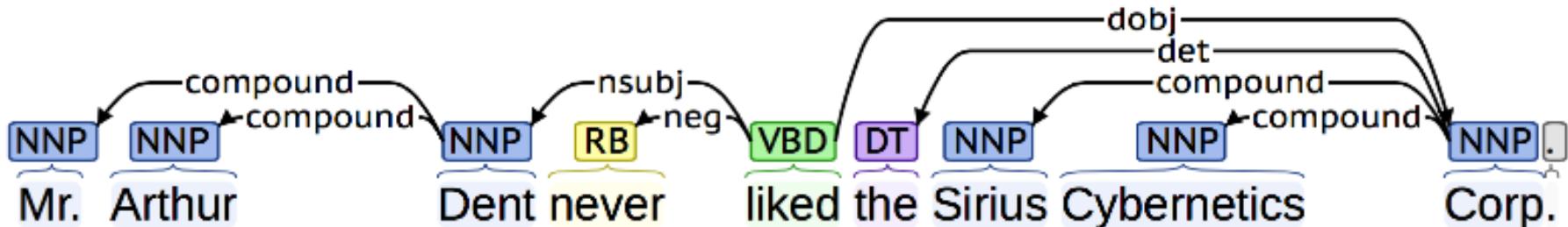
- Erkennung von speziellen Wortfolgen (Segmenten) und ihre Kategorisierung
- z.B. Eigennamen (**NER** = named entity recognition)

Mr. Arthur Dent never liked the Sirius Cybernetics Corp.

Person Organization

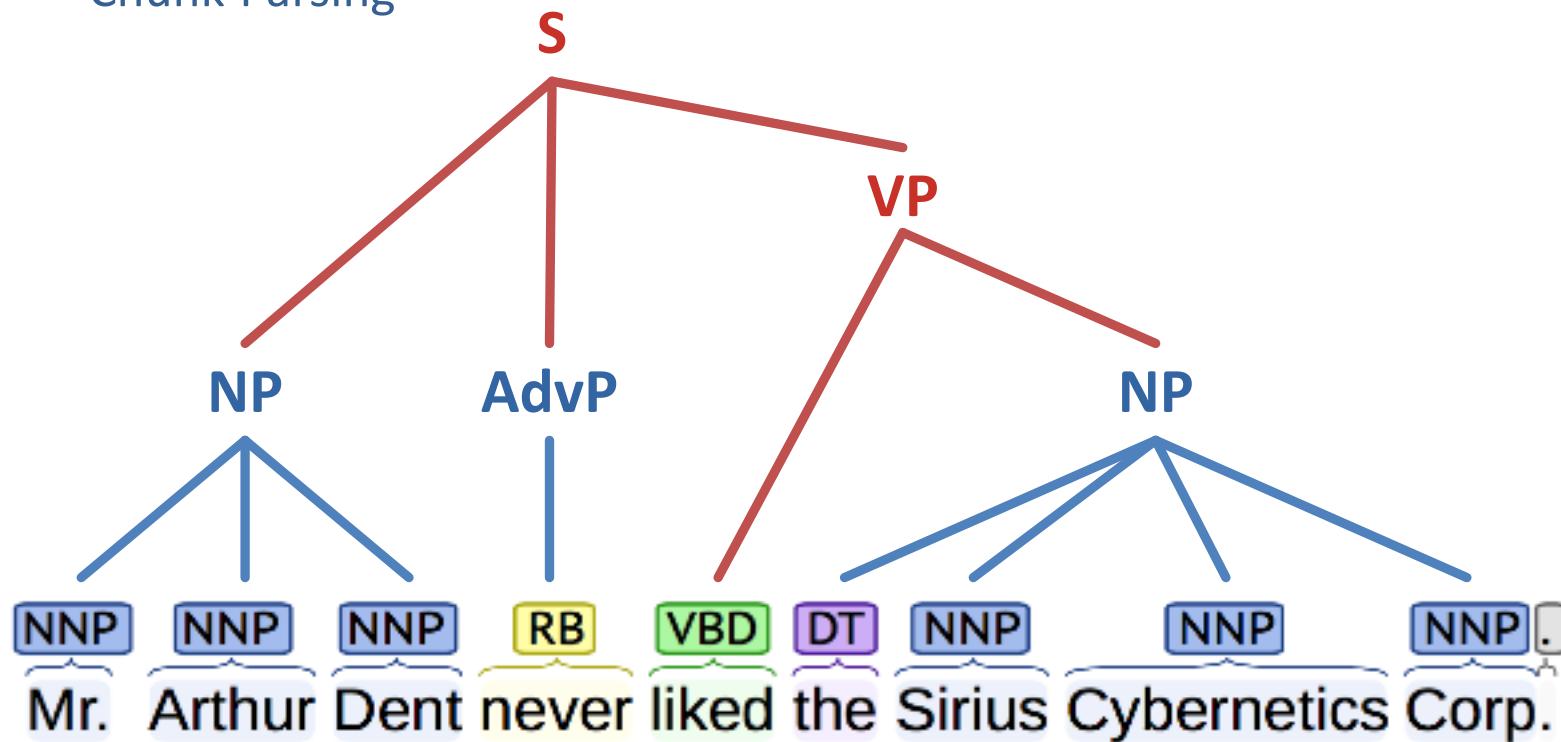


- auch wichtig: Erkennung zu anonymisierender Textstellen
- Erkennung der Satzstruktur = **Parsing**
 - z.B. Abhängigkeiten zwischen Wörtern → Dependenz-Graph



Korpusannotation: Satzstruktur als Baum

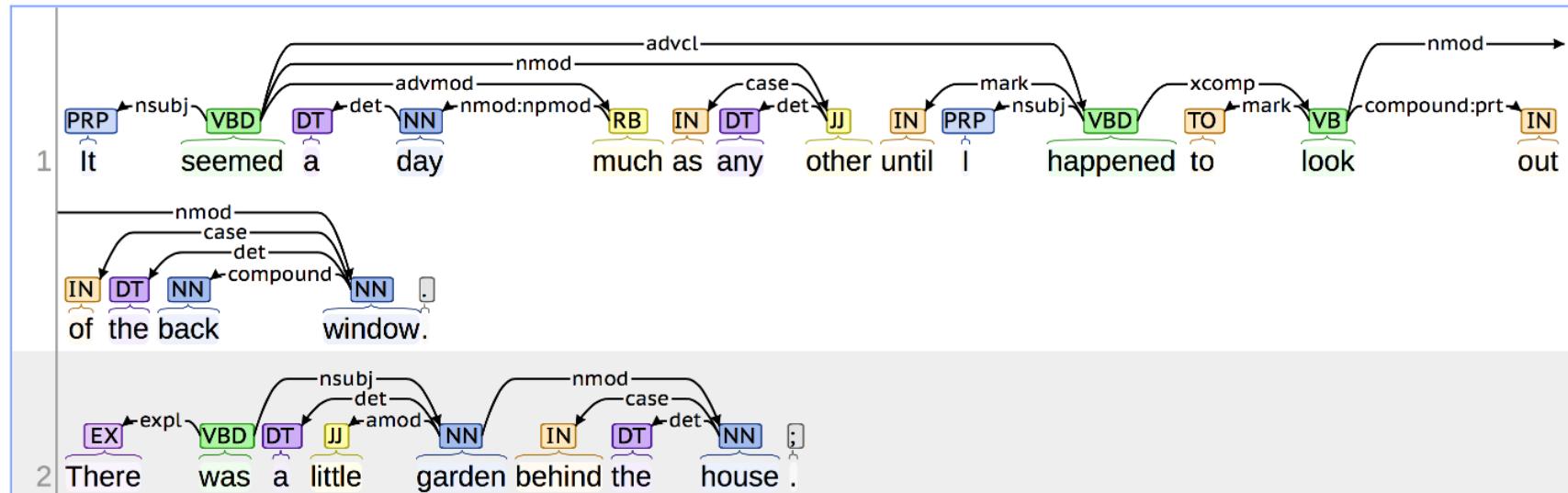
- Phrasenstruktur als baumförmige Hierarchie von Konstituenten
- „minimale“ Phrasen können auch als Segmente interpretiert werden
→ Chunk-Parsing



Korpusannotation: Satzalignierung für Parallelkorpora

| | |
|---|--|
| Das stört mich keineswegs, ich halte das für eine gute Initiative, aber wiederum ist Europa nicht zur Stelle. | That is no problem for me. I think it is a good initiative, but again Europe is absent. |
| Es darf nicht wieder geschehen! | It should not happen again, Mr President. |
| Meine Fraktion verlangt, daß die italienische Präsidentschaft hier vor uns erklärt, welche Rolle sie spielt. | My Group wants the Italian presidency to come here and explain what its role is. |
| Herr Präsident, liebe Kolleginnen und Kollegen! | Mr President, ladies and gentlemen, I think it is important that we should discuss the situation in the Middle East this week. |
| Ich halte es für wichtig, daß wir diese Woche über die Situation im Nahen Osten reden. | We all agree on that. |

Kurze Spielpause ...



... zum Ausprobieren (wie gut & nützlich sind die Analysen?)

- <http://corenlp.run/>
- <https://explosion.ai/demos/displacy>

Manuelle Annotation

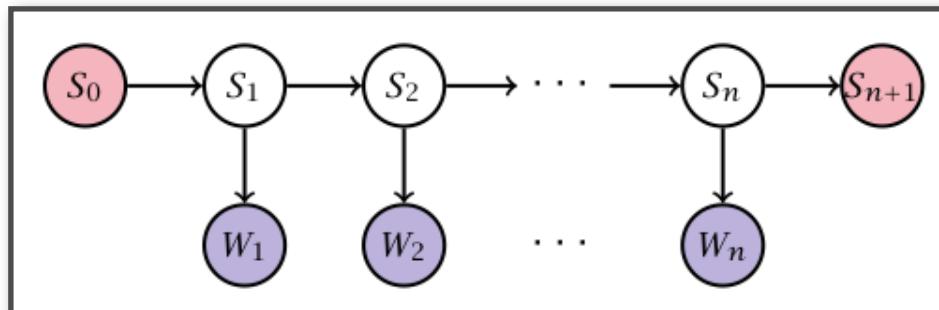
- Kleine Korpora werden oft manuell annotiert
 - z.B. digitale Editionen, Reden eines Präsidenten, ...
- **Annotationsschema** und -kategorien (**Tagset**)
- **Richtlinien (Guidelines)**
 - detaillierte Beschreibung und Abgrenzung der Zielkategorien
 - zusätzlich: Beispielsammlung für schwierige Einzelfälle
- Annotationswerkzeuge (meist Web-basiert)
 - z.B. INCEpTION (<https://inception-project.github.io>), Prodigy (<https://prodi.gy>)
- **Inter-Annotator Agreement (IAA)**
 - wichtig! – überprüft Reliabilität und Validität der Annotation
 - Flüchtigkeitsfehler vs. systematische Differenzen
 - Adjudikation für Endfassung der Annotation

Automatische Annotation

- Für größere Korpora ist eine manuelle Annotation zu teuer und zeitaufwendig
- Auch in den Digital Humanities ...
 - Romane von Charles Dickens ca. 4 Mio. Wörter
 - Deutsches Gutenberg-Archiv > 100 Mio. Wörter
 - Early English Books (EEBO) > 500 Mio. Wörter
 - Times Online 1780–1900 ca. 4.000 Mio. Wörter

Automatische Annotation

- Erfolgreichster Ansatz: **maschinelle Lernverfahren**
 - ab ca. 1990 Einsatz von statistischen Modellen („statistical revolution“)
 - aktuell große Fortschritte mit **Deep Learning**
- **Trainingskorpus** (manuell annotiert)
 - wichtig: Konsistenz der Annotationen
 - Flüchtigkeitsfehler scheinen weniger problematisch
- Evaluation auf separatem Testkorpus
 - Gefahr der Überanpassung an das Trainingskorpus
- Beispiel: Tagging mit **Hidden Markov Model** (HMM)



(Evert et al. 2009)

XML als Repräsentationsformat

```
● H9C.xml*  
1 <bncDoc xml:id="H9C">  
2   <teiHeader> ← TEI header enthält Metadaten  
3     <fileDesc>  
4       <titleStmt>  
5         <title> The prince of darkness. Sample containing about 44223 words from a book  
6           (domain: imaginative) </title>  
7         <respStmt>  
8           <resp> Data capture and transcription </resp>  
9             <name> Oxford University Press </name>  
10            </respStmt>  
11        </titleStmt>  
12        <editionStmt> ← Beispiel aus dem British National Corpus  
13          <edition>BNC XML Edition, December 2006</edition>  
14        </editionStmt>  
15        <extent> 44223 tokens; 44797 w-units; 3933 s-units </extent>  
16        <publicationStmt>  
17          <distributor>Distributed under licence by Oxford University Computing Services on  
18            behalf of the BNC Consortium.</distributor>  
19          <availability> This material is protected by international copyright laws and may  
20            not be copied or redistributed in any way. Consult the BNC Web Site at  
21              http://www.natcorp.ox.ac.uk for full licencing and distribution  
22            conditions.</availability>  
23          <idno type="bnc">H9C</idno>  
24          <idno type="old"> PDarkn </idno>  
25        </publicationStmt>  
26        <sourceDesc>  
27          <bibl> ← Informationen über den Text  
28            <title>The prince of darkness. </title>  
29            <author domicile="Epping" n="DoherP1">Doherty, P C</author>  
30            <imprint n="HEADLI1">  
31              <publisher>Headline Book Publishing plc</publisher>  
32              <pubPlace>London</pubPlace>  
33              <date value="1992">1992</date>  
34            </imprint>  
35          </bibl>  
36        </sourceDesc>  
37      </fileDesc>  
38      <encodingDesc>  
39        <tagsDecl>  
40          <namespace name="">  
41            <tagUsage ci="c" occurs="8764" />
```

XML als Repräsentationsformat

```
80 <wtext type="FICTION">  
81   <pb n="69"/>  
82   <div level="1">  
83     <head>  
84       <s n="2">  
85         <w c5="NN1" hw="chapter" pos="SUBST">Chapter </w>  
86         <w c5="CRD" hw="5" pos="ADJ">5</w>  
87       </s>  
88     </head>  
89   <p>  
90     <s n="3">  
91       <w c5="VVB-NN1" hw="ranulf" pos="VERB">Ranulf </w>  
92       <w c5="CJC" hw="and" pos="CONJ">and </w>  
93       <w c5="NP0" hw="dame" pos="SUBST">Dame </w>  
94       <w c5="NP0" hw="agatha" pos="SUBST">Agatha </w>  
95       <w c5="VBD" hw="be" pos="VERB">were </w>  
96       <w c5="VVG" hw="wait" pos="VERB">waiting </w> ←  
97       <w c5="PRP" hw="for" pos="PREP">for </w>  
98       <w c5="PNP" hw="he" pos="PRON">him </w>  
99       <w c5="PRP" hw="near" pos="PREP">near </w>  
100      <w c5="AT0" hw="the" pos="ART">the </w>  
101      <w c5="NN1-NP0" hw="galilee" pos="SUBST">Galilee </w>  
102      <w c5="NN1" hw="gate" pos="SUBST">Gate</w>  
103      <c c5="PUN">, </c>  
104      <w c5="AT0" hw="the" pos="ART">the </w>  
105      <w c5="AJ0" hw="young" pos="ADJ">young </w>  
106      <w c5="NN1" hw="nun" pos="SUBST">nun </w>  
107      <w c5="AV0" hw="apparently" pos="ADV">apparently </w>  
108      <w c5="VVG" hw="enjoy" pos="VERB">enjoying </w>  
109      <w c5="AT0" hw="an" pos="ART">an </w>  
110      <w c5="NN1" hw="account" pos="SUBST">account </w>  
111      <w c5="PRF" hw="of" pos="PREP">of </w>  
112      <w c5="CRD" hw="one" pos="ADJ">one </w>  
113      <w c5="PRF" hw="of" pos="PREP">of </w>  
114      <w c5="DPS" hw="he" pos="PRON">his </w>  
115      <w c5="NN1" hw="manservant" pos="SUBST">manservant</w>  
116      <w c5="POS" hw="s" pos="UNC">'s </w>  
117      <w c5="DT0" hw="many" pos="ADJ">many </w>  
118      <w c5="NN2" hw="escapade" pos="SUBST">escapades </w>  
119      <w c5="PRP" hw="in" pos="PREP">in </w>  
120      <w c5="NP0" hw="london" pos="SUBST">London</w>  
121      <c c5="PUN">.</c>
```

TEI body enthält Objektdaten

Textstruktur & Darstellung

Token mit Annotationen

XML-Prinzip:
Durch Entfernen aller
XML-Tags kann der
ursprüngliche Objekttext
wiederhergestellt werden

Weiterführende Literatur (1)

- McEnery, Tony / Wilson, Andrew (2001): *Corpus Linguistics*. 2. Aufl. Edinburgh University Press.
- McEnery, Tony / Xiao, Richard / Tono, Yukio (2006): *Corpus-Based Language Studies: An advanced resource book*. London / New York: Routledge.
 - online: <https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm>
- McEnery, Tony / Hardie, Andrew (2012): *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- Hirschmann, Hagen (2019): *Korpuslinguistik: Eine Einführung*. Berlin: J.B. Metzler.
- Hoffmann, Sebastian / Evert, Stefan / Smith, Nicholas / Lee, David / Berglund Prytz, Ylva (2008): *Corpus Linguistics with BNCweb – a Practical Guide*. Frankfurt am Main: Peter Lang.

Weiterführende Literatur (2)

- Lemnitzer, Lothar / Zinsmeister, Heike (2015): *Korpuslinguistik: Eine Einführung*. 3. Aufl. Tübingen: Narr.
- Lüdeling, Anke / Kytö, Merja (Hrsg.) (2008/2009): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Stefanowitsch, Anatol (2020): *Corpus linguistics. A guide to the methodology*. Berlin: Language Science Press.
 - online: <https://langsci-press.org/catalog/book/148>
- Online-Kurs von Noah Bubenhofer:
<http://www.bubenhofer.com/korpuslinguistik/kurs/>