Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

# Korpus- und Computerlinguistik

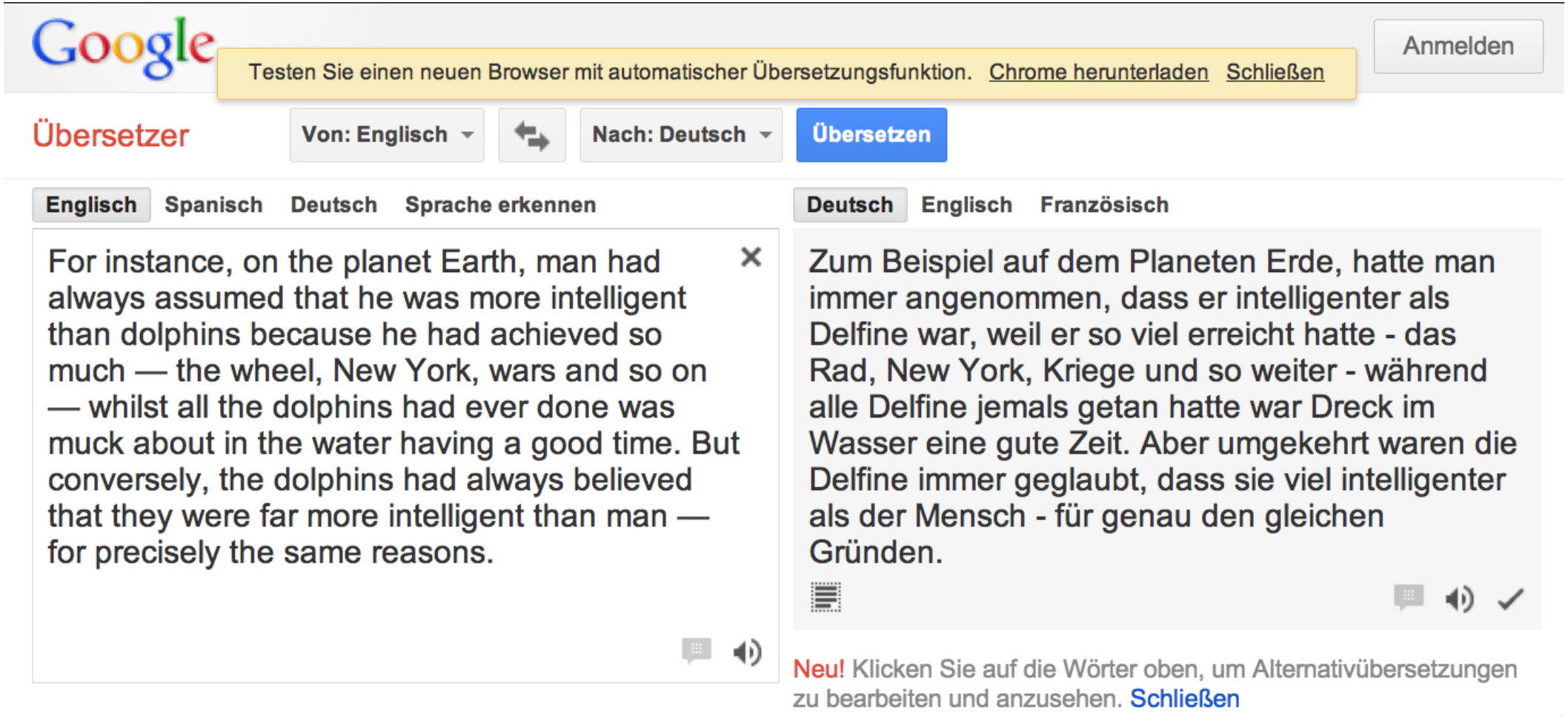
Vorlesungsreihe *Recht durch Maschinen und Künstliche Intelligenz*

**Philipp Heinrich**

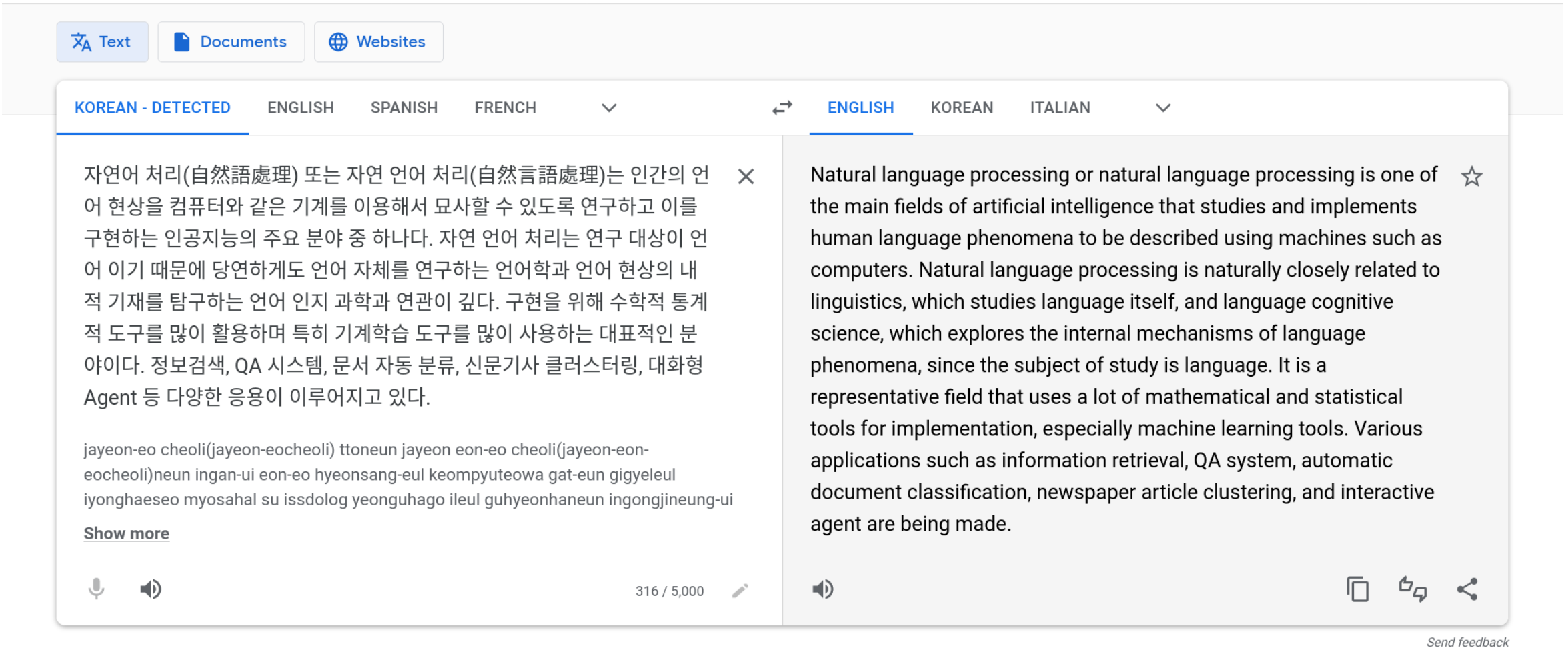
Lehrstuhl für Korpus- und Computerlinguistik

17. November 2022

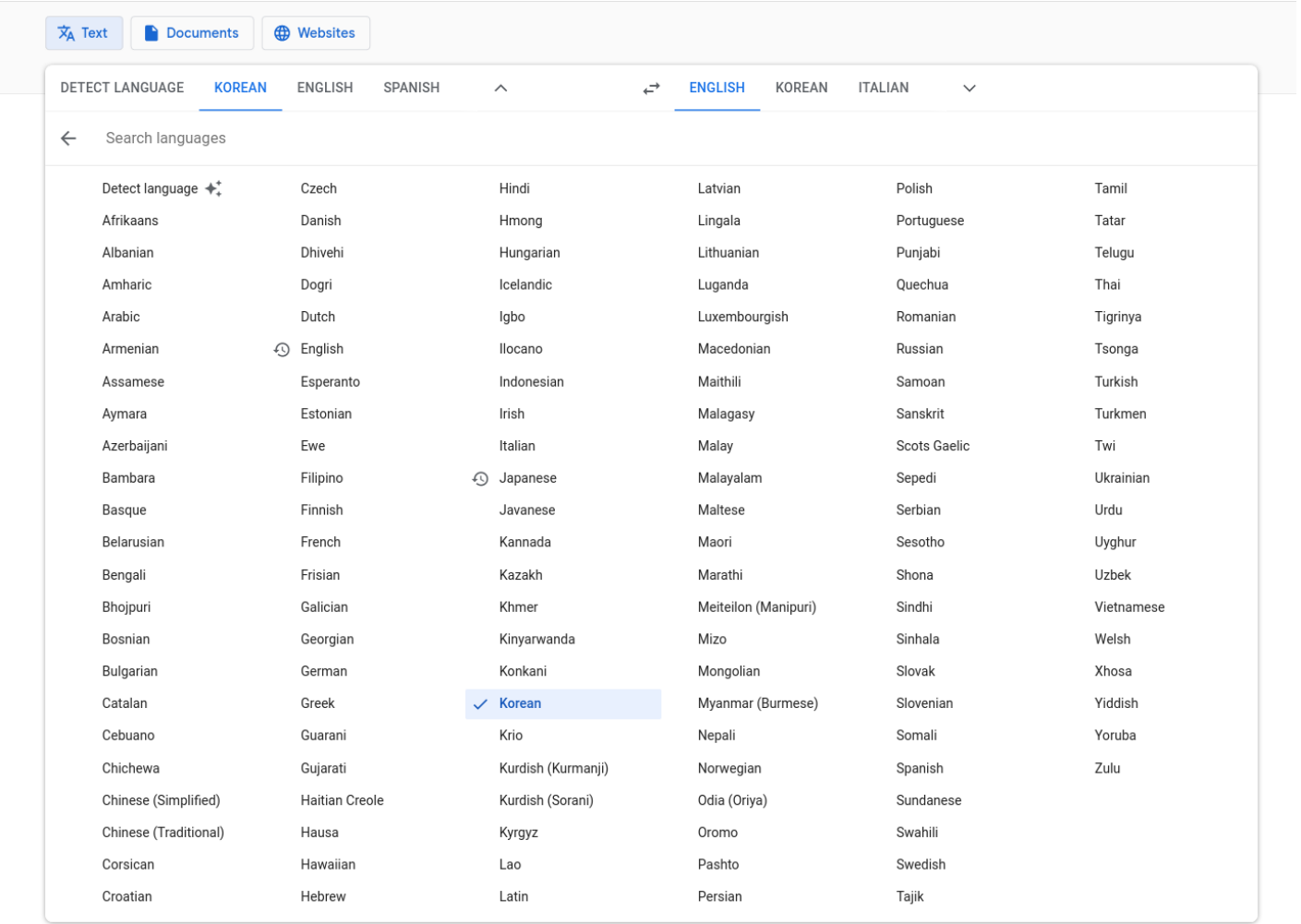
# Maschinelle Übersetzung



# Maschinelle Übersetzung



# Maschinelle Übersetzung

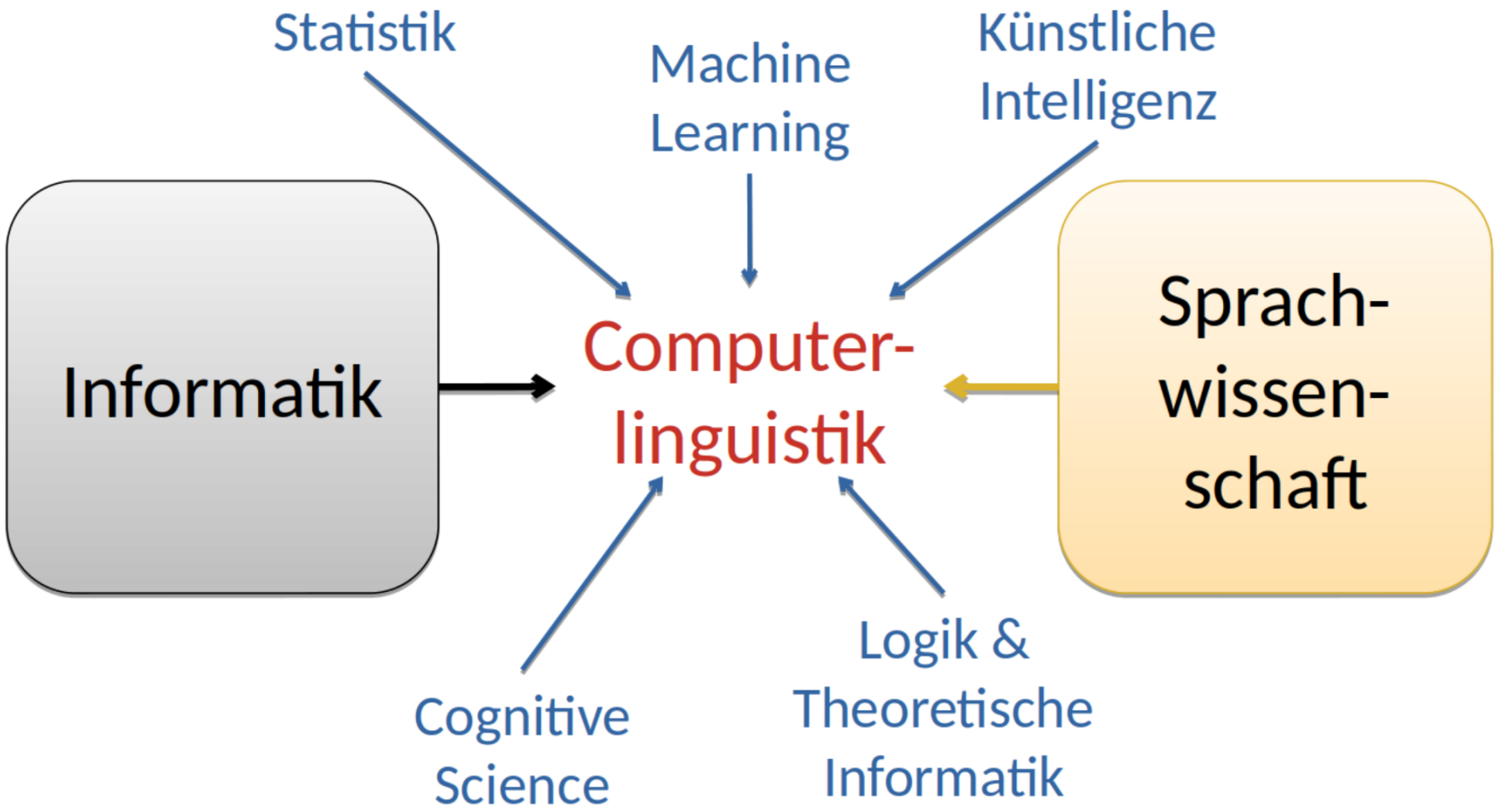


# Anwendungen der Computerlinguistik

= NLP, linguistische Datenverarbeitung, . . .

- Maschinelle Übersetzung (Google, DeepL, . . . )
- Rechtschreibkorrektur und Grammatikprüfung
- Virtuelle Keyboards
- Spamerkennung
- (Semantische) Websuche
- Diktieren & Sprachsteuerung
- Sprachausgabe (z.B. Navi)
- Sprachdialogsystem (z.B. im Auto, Fahrplanauskunft, . . . )
- Virtuelle Assistenten (z.B. Siri, Alexa, Cortana, . . . )
- KI (Watson, IBM Project Debater, . . . )
- Meinungs- & Marktforschung
- Wörterbücher (zweisprachige, Lerner-WB, Kollokations-WB)
- Information Retrieval & Question Answering
- Text / Knowledge Mining (z.B. Biomedizin, Patentrecherche)
- Social Bots & Fake News
- Kontrollierte Terminologie
- Plagiaterkennung
- Automatische Bewertung von Klausuren und Hausarbeiten
- Spam verfassen
- Forensik & Aufklärung (z.B. linguistische Steganographie)

# Was ist Computerlinguistik?



# Was ist ein Korpus?

## Corpus

OXFORD UNIVERSITY PRESS Views **1,883,913** Updated **Jun 11 2018**

**cor·pus** / 'kôrpəs/

- n. (pl. **-po·ra** / -pərə/ or **-pus·es**) **1.** a collection of written texts, esp. the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus.*
  - a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.

  **2.** Anat. the main body or mass of a structure.
    - the central part of the stomach, between the fundus and the antrum.

The Oxford Pocket Dictionary of Current English

- Korpus[a] als repräsentative Stichprobe der Sprache
- Korpus als maschinenlesbare Textsammlung

[a] Achtung: *das Korpus, die Korpora*

# Was ist ein Korpus?

Korpus im engeren Sinn = repräsentative Stichprobe

- repräsentativ für Sprachvarietät, Register, Spezialsprache, Gruppe von Sprechern, ...
- statistische Inferenz generalisiert Beobachtungen
- Brown Corpus: AmE Schriftsprache der 1960er Jahre
  - 500 Textausschnitte à 2000 Wörtern = 1M Wörter
  - 15 Genres (Zeitungen, Akad., Sachbücher, Belletristik)
- British National Corpus: BrE der 1990er Jahre
  - 90M Schriftsprache + 10M gesprochene Sprache
  - umfangreiche Metadaten zu Texten und Sprechern

Korpus im weiteren Sinn = elektronische Textsammlung

- Korpussuche, statistische Auswertung, Training von maschinellen Lernverfahren, Wissensextraktion
- Digital Humanities: Digitalisierung von Kulturgütern
- Computerlinguistik (NLP): bigger is better
- Webkorpora: opportunistische Sammlung v. Webseiten
  - WaCky (2007): je 2G Wörter Englisch, Deutsch, ...
  - COW (2014): > 10G Wörter Englisch, Deutsch, ...
- N-Gramm-Datenbanken: Häufige Wortsequenzen
  - Google Web1T5 aus ca. 1T (Billion) Wörter Webseiten
  - Google Books: bis zu 900G Wörter aus gescannten Büchern

# Bestandteile eines Korpus

- 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

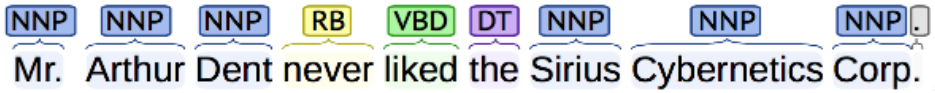
# Linguistische Annotation von Korpora

## Annotation auf Wortebene

- Jedem (laufenden) Wort wird eine Kategorie zugeordnet
  - sog. 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) starkeHinweise darauf, dass(wie auch imRealLife) diverseFaktoren die sprVariation beeinflussen: http://tinyurl.com/3umxkuh

- Wortartenannotierung = POS-Tagging
  - Substantiv (noun), Adjektiv, Verb, Adverb, Pronomen, Präposition, Konjunktion, Zahl, Satzzeichen, . . .
  - engl. POS = part of speech
  - Tagset = Kategorienschema (fein vs. grob)




- zusätzlich:
  - Lemmatisierung
  - semantische Kategorien
  - emotionale Valenzen
  - Schwierigkeitsgrad

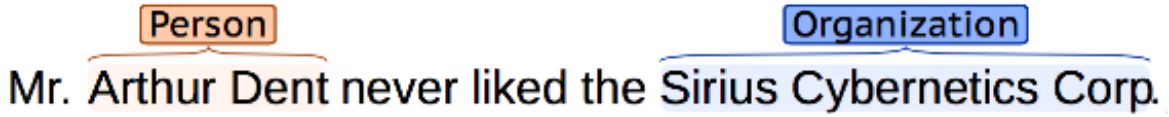
# Linguistische Annotation von Korpora

## POS-Tagset: STTS (Deutsch)

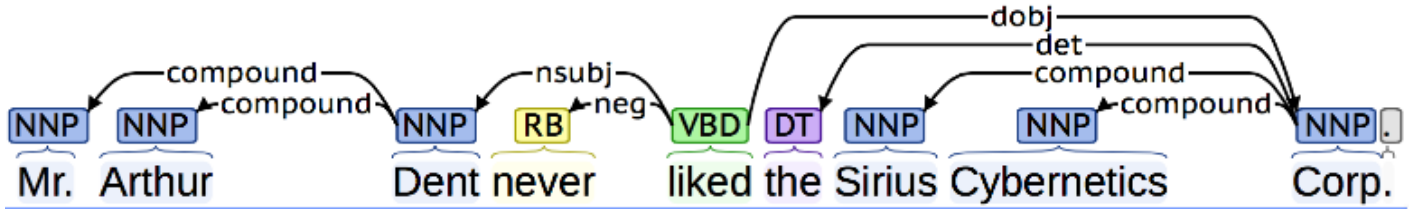
| | | |
|---|---|---|
| ADJA | attributives Adjektiv | |
| ADJD | adverbiales / prädikatives Adjektiv | |
| ADV | Adverb | *schon, bald, doch* |
| APPR | Präposition / Zirkumposition links | |
| APPRART | Präposition mit Artikel fusioniert | *zum* |
| APPO | Postposition | *zufolge, wegen* |
| APZR | Zirkumposition rechts | *von ... an* |
| ART | bestimmter oder unbestimmter Artikel | |
| CARD | Kardinalzahlen (Ordinalzahl = ADJA) | |
| FM | Fremdsprachliches Material | |
| ITJ | Interjektion | *mhm, ach, tja* |
| KOUI | unterordnende Konj. mit *zu* + Inf | |
| KOUS | unterordnende Konjunktion mit Satz | |
| KON | nebenordnende Konjunktion | *und, oder* |
| KOKOM | Vergleichskonjunktion | *als, wie* |
| 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) | |
| PPOSS | 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 | *dafür, deswegen* |
| PTKZU | *zu* vor Infinitiv | |
| PTKNEG | Negationspartikel | *nicht* |
| PTKVZ | abgetrennter Verbzusatz | *kommt ... an* |
| PTKANT | Antwortpartikel | *ja, nein, danke* |
| PTKA | Partikel bei Adjektiv/Adverb | *am, zu* |
| TRUNC | Kompositions-Erstglied | *Unter- und ...* |
| VVFIN | finites Verb, voll (= lexikalisch) | |
| VVIMP | Imperativ, voll | |
| VVINF | Infinitiv, voll | |
| VVIZU | Infinitiv mit *zu*, 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 | *3:7, H2O* |
| $, | Komma | *,* |
| $. | Satzbeendende Interpunktion | *. ? ! ; :* |
| $( | sonstige Satzzeichen (intern) | *– [ ] ( )* |

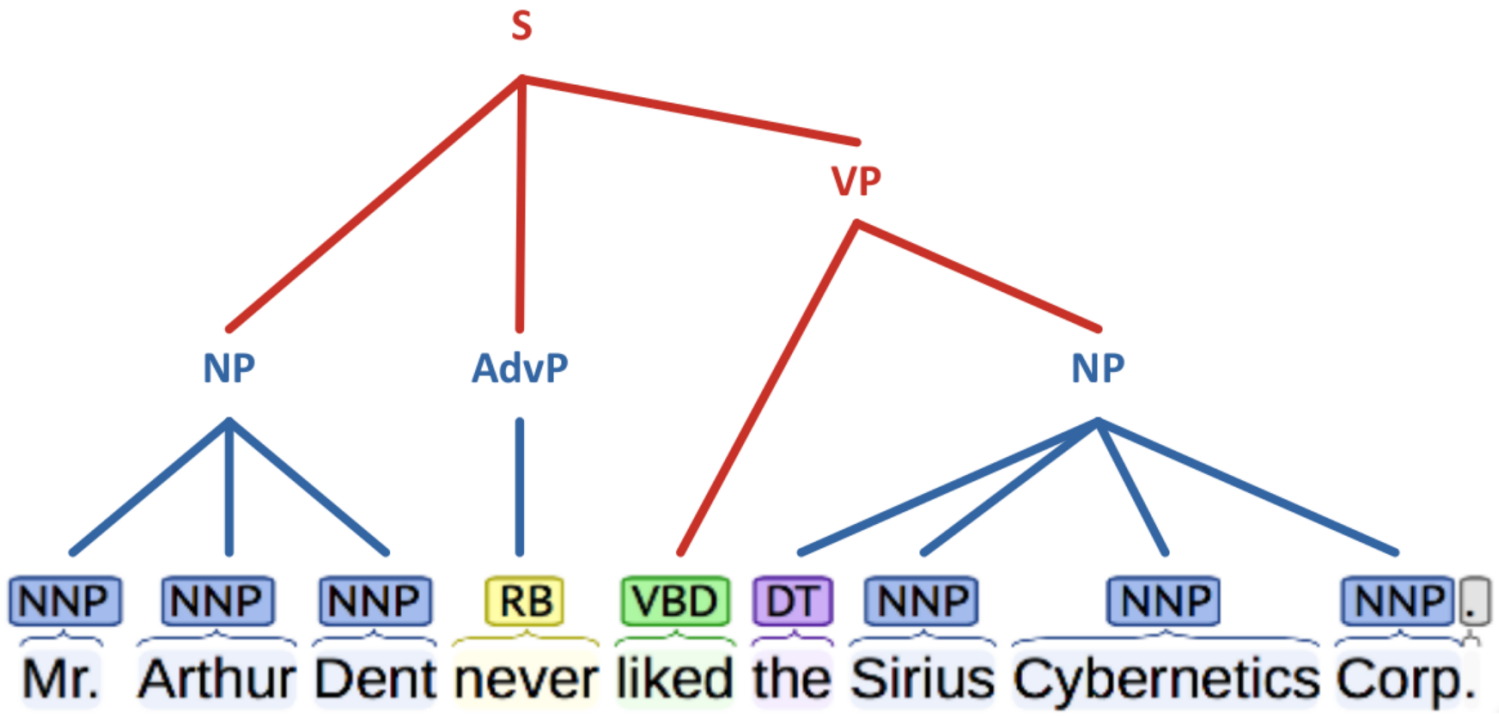
- Erkennung von speziellen Wortfolgen (Segmenten) und ihre Kategorisierung
  - z.B. Eigennamen (NER = named entity recognition)




- Erkennung der Satzstruktur = Parsing
  - z.B. Abhängigkeiten zwischen Wörtern (Dependenz-Graph)




  - alternativ: Phrasenstruktur als baumförmige Hierarchie
  - „minimale“ Phrasen können auch als Segmente interpretiert werden (Chunk-Parsing)

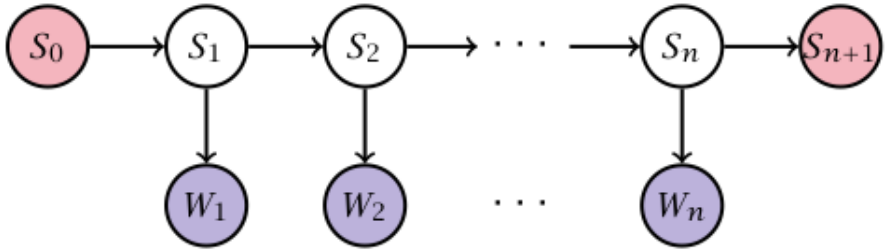

S
VP
NP
AdvP
NP
NNP
NNP
NNP
RB
VBD
DT
NNP
NNP
NNP
.
Mr.
Arthur
Dent
never
liked
the
Sirius
Cybernetics
Corp.

# Manuelle vs. automatische Annotation

- kleine Korpora werden oft manuell annotiert
  - z.B. digitale Editionen, Reden eines Präsidenten, ...
- Annotationsschema und -kategorien
- Richtlinien (Guidelines)
  - zusätzlich: Beispielsammlung für schwierige Einzelfälle
- Annotationswerkzeuge (meist Web-basiert)
  - z.B. brat (https://brat.nlplab.org/)
- Inter-Annotator Agreement (IAA)
  - überprüft Reliabilität und Validität der Annotation
  - Flüchtigkeitsfehler vs. systematische Differenzen
- für größere Korpora ist eine manuelle Annotation zu teuer und zeitaufwändig
  - 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 Verfahren zur Annotation von großen Datenmengen
- erfolgreichster Ansatz: maschinelle Lernverfahren
- 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)



# XML/VRT als Repräsentationsformat

```
<article year="2010" month="2010-01" date="2010-01-01" fname="litauenreaktor100" rubrik="ausland">
<p type="topline">
<s>
Forderung       NN
der     ART
EU      NE
erfüllt VVPP
</s>
</p>
<p type="headline">
<s>
Litauen NE
schaltet        VVFIN
sein    PPOSAT
letztes ADJA
Atomkraftwerk   NN
ab      PTKVZ
</s>
</p>
<p type="base">
<s>
Litauen NE
hat     VAFIN
sein    PPOSAT
einziges        ADJA
Atomkraftwerk   NN
endgültig       ADJD
abgeschaltet    VVPP
.       $.
</s>
```

| corpus position | word form | ID | part of speech | ID | lemma | ID |
|---|---|---|---|---|---|---|
| (0) | <text> value = "id=42 lang="English"" | | | | | |
| (0) | <text_id> value = "42" | | | | | |
| (0) | <text_lang> value = "English" | | | | | |
| (0) | <s> | | | | | |
| 0 | An | 0 | DET | 0 | a | 0 |
| 1 | easy | 1 | ADJ | 1 | easy | 1 |
| 2 | example | 2 | NN | 2 | example | 2 |
| 3 | . | 3 | PUN | 3 | . | 3 |
| (3) | </s> | | | | | |
| ⋮ | ⋮ | | | | | |
| (13) | </text_lang> | | | | | |
| (13) | </text_id> | | | | | |
| (13) | </text> | | | | | |

# Korpusabfragen

- Abfrage von Korpora via Korpusabfragesprachen (CQL)
- Softwarelösungen nutzen unterschiedliche Abfragesprachen:
  - CQP (CQPweb, SketchEngine)
  - AQL (ANNIS)
  - COSMAS II QL (DeReKo)
- manche Resourcen sind in der Praxis nur über Software verfügbar, die *eine spezielle* CQL verwendet
- Abfragesprachen sind unterschiedlich ausdrucksstark:
  - finde alle Wörter, die auf *ität* enden
  - finde alle Sätze, die sowohl *Covid* als auch *Bayern* enthalten
  - finde alle Nomina, die durch *gut* modifiziert werden
- CCL Erlangen: CQP als Abfragesprache
  - vereinfachte Version in CQPweb
  - Arbeit an einer Corpus Query Lingua Franca (CQLF)

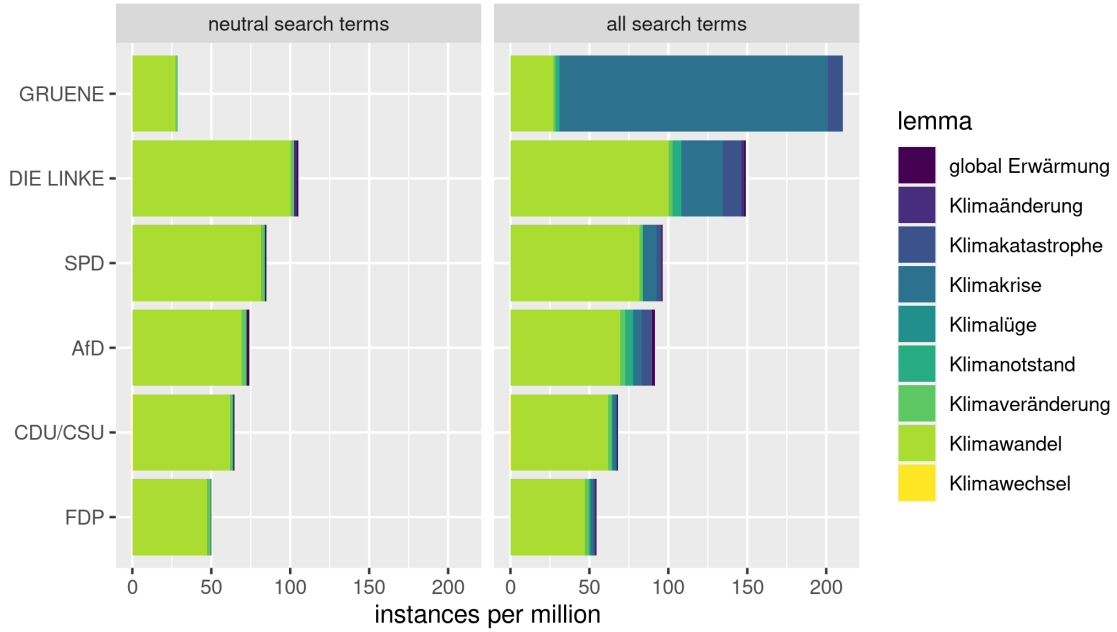



Abbildung: Frequency Breakdown der Query (GermaParl).

```
[lemma="global"] [lemma="Erwärmung"] | [lemma="Klimawandel"] | [lemma="Klimalüge"] | ...
```

# Konkordanz

| No. | Text | Solution 1 to 50 Page 1 / 38 | | |
|---|---|---|---|---|
| 1 | p_19_111_56 | dass es in Zukunft Kriege gibt , die etwas mit dem | Klimawandel | zu tun haben , der muss heute den Klimawandel bekämpfen . Daher |
| 2 | p_19_231_49 | Welche habe ich vorhin gesagt ? ) – Sie ignorieren den menschengemachten | Klimawandel | . Und Sie reden hier mit jemandem , der Biologie studiert hat |
| 3 | p_19_176_151 | Flucht . Schon heute werden innerstaatlich mehr Menschen durch die Folgen der | Klimakrise | vertrieben als durch Gewalt und Konflikte . Am härtesten trifft es den |
| 4 | p_19_48_35 | Braun und Gelb dort , wo eigentlich Grün sein müsste . Die | Klimakrise | selbst ist das größte Problem , die größte Zumutung , die wir |
| 5 | p_19_63_82 | darauf eingegangen : Im Bundeshaushalt geschieht jetzt schon einiges , um den | Klimawandel | zu stoppen bzw. wenigstens zu verlangsamen . Im vergangenen Jahr 2017 sagte |
| 6 | p_19_24_161 | wenn er mich fragt : Warum habt ihr nichts gegen den menschengemachten | Klimawandel | getan ? ( Beifall bei der FDP , der SPD und dem |
| 7 | p_19_111_154 | denn nun will auch das BMZ verstärkt in die weltweite Bekämpfung des | Klimawandels | einsteigen . Herr Minister , Sie sagten vor wenigen Wochen in der |
| 8 | p_19_81_11 | Ihnen , dass Sie auch etwas vorlegen ; denn in Zeiten der | Klimakrise | können wir nicht über KI reden , ohne diese Themen ganz klar |
| 9 | p_19_175_406 | Das Geld fehlt für Investitionen und befeuert vor allen Dingen massiv die | Klimakrise | . Wenn wir also eine nachhaltige Finanzpolitik machen wollen , können wir |
| 10 | p_19_184_239 | sich beliebig fortsetzen , von den immensen Herausforderungen der Coronapandemie und der | Klimakrise | ganz zu schweigen . Und bei all diesen Problemen richtet sich der |
| 11 | p_19_238_96 | Beifall bei der AfD ) und : „ Die Behauptung , der | Klimawandel | ist schuld , ist … nicht haltbar . “ Das war ein |
| 12 | p_19_54_323 | der gegen RWE klagt , weil sein Bergdorf in den Anden vom | Klimawandel | bedroht ist . Wie es jetzt aussieht , hat seine Klage sogar |
| 13 | p_19_182_233 | jetzt umso mehr ; denn wir haben erlebt , wie sich die | Klimakrise | verschärft , wir haben doch erlebt , was in Australien los war |
| 14 | p_19_205_335 | dass die Landwirtschaft einen Beitrag leisten muss im Kampf gegen den | Klimawandel | . Weltweit steht immer weniger Fläche zur Verfügung . Und wir haben |
| 15 | p_19_44_121 | einfacher Kaufmann auch nicht wie Sie hier ein pseudowissenschaftliches Seminar zum Thema | Klimawandel | halten . Ich probiere es erst einmal versöhnlich . ( Karsten Hilse |
| 16 | p_19_179_129 | mit denen wir konfrontiert sind – Stichwort „ Klimawandel “ . Der | Klimawandel | – wir haben das vorhin auch schon angesprochen – ist ein sicherheitsrelevanter |
| 17 | p_19_135_96 | wie viel Geld wir , die Industrienationen und somit die Hauptverursacher des | Klimawandels | , für von Naturkatastrophen betroffene Länder beisteuern , wenn wir unsere Hausaufgaben |
| 18 | p_19_99_138 | Aber natürlich gehört zu den Herausforderungen auch der Kampf gegen den | Klimawandel | . Wir wollen das Klimaschutzabkommen erfüllen , damit die Temperatur auf der |
| 19 | p_19_80_201 | denn die Welt schaut , wie wir den Kampf gegen den | Klimawandel | aufnehmen . Ich hatte das Vergnügen und die Ehre , an der |
| 20 | p_19_185_119 | wenn man auf der einen Seite sagt : „ Die Bekämpfung des | Klimawandels | und das Gelingen der Energiewende sind von essenzieller Bedeutung für die Zukunft |

# Kollokation

- *Kollokate* eines Wortes (dem *Knoten*, oder engl. *node*) sind Wörter, die häufig in dessen Umgebung auftreten (Ko-Okkurrenz)

- Einblick in die Semantik des Wortes, vgl. Firth's (1957) *distributional hypothesis*:

  *You shall know a word by the company it keeps!*

- Anwendung bspw. in der *Diskursanalyse*, *Lexikographie*, ...
- Kollokation als Phänomen, das in Korpora empirisch beobachtbar ist
  - Kollokate von „Klimawandel“ in GermaParl
  - Kollokate von „Impfung“ auf Twitter
- Kollokate $\neq$ Mehrworteinheiten, Idiome, Phraseologismen, ...

- Parameter: Kollokationsart, Fenstergröße, Assoziationsmaß, ...

| rank | lemma | LRC |
|---|---|---|
| 1 | menschengemacht | 11.74 |
| 2 | Artensterben | 9.20 |
| 3 | Ressourcenknappheit | 7.76 |
| 4 | Anpassungsstrategie | 7.74 |
| 5 | Menschheitsherausforderung | 6.85 |
| 6 | Erderwärmung | 6.37 |
| 7 | Dürre | 6.19 |
| 8 | drohend | 5.98 |
| 9 | Wetterextrem | 5.89 |
| 10 | Hauptverursacher | 5.66 |
| 11 | Wassermangel | 5.62 |
| 12 | Artenvielfalt | 5.48 |
| 13 | Bevölkerungswachstum | 5.37 |
| 14 | Anpassungsmaßnahme | 5.34 |
| 15 | Biodiversität | 5.27 |
| 16 | Herausforderung | 5.24 |
| 17 | Naturkatastrophe | 5.23 |
| 18 | Ernährungskrise | 5.22 |
| 19 | Treibhauseffekt | 5.16 |
| 20 | leugnen | 5.08 |

Tabelle: 10L-10R-S-Kollokate zu *Klimawandel* (GermaParl).

# Kollokation

## Assoziationsmaße

| rank | lemma | LLR |
|---|---|---|
| 1 | menschengemacht | 2406.84 |
| 2 | Herausforderung | 1848.52 |
| 3 | die | 1425.40 |
| 4 | Folge | 1400.30 |
| 5 | global | 1099.30 |
| 6 | Auswirkung | 854.99 |
| 7 | Anpassung | 850.94 |
| 8 | Bekämpfung | 826.61 |
| 9 | Kampf | 790.89 |
| 10 | drohend | 703.85 |
| 11 | gegen | 652.04 |
| 12 | Artensterben | 616.13 |
| 13 | bekämpfen | 545.56 |
| 14 | leugnen | 526.50 |
| 15 | , | 492.06 |
| 16 | dass | 482.90 |
| 17 | Digitalisierung | 461.30 |
| 18 | aufhalten | 402.87 |
| 19 | Klimaschutz | 389.67 |
| 20 | und | 376.72 |

Tabelle: 10L-10R-S-Kollokate zu *Klimawandel* (GermaParl).

| rank | lemma | MI |
|---|---|---|
| 1 | menschengemacht | 3.38 |
| 2 | Menschheitsherausforderung | 3.08 |
| 3 | Artensterben | 3.04 |
| 4 | Ressourcenknappheit | 3.03 |
| 5 | Extremwetterlage | 3.00 |
| 6 | Ernährungskrise | 2.99 |
| 7 | Biodiversitätsverlust | 2.97 |
| 8 | Anpassungsstrategie | 2.89 |
| 9 | Leugner | 2.85 |
| 10 | Wetterextrem | 2.83 |
| 11 | Nicholas | 2.70 |
| 12 | Wassermangel | 2.68 |
| 13 | Extremwetterereignis | 2.65 |
| 14 | Eisbär | 2.64 |
| 15 | Klimaflüchtling | 2.62 |
| 16 | Hungerkrise | 2.58 |
| 17 | Dürreperiode | 2.56 |
| 18 | Menschheitsfrage | 2.56 |
| 19 | Wüstenbildung | 2.52 |
| 20 | Erderwärmung | 2.50 |

Tabelle: 10L-10R-S-Kollokate zu *Klimawandel* (GermaParl).

# Konkordanz

| ID | ... context | keyword | context ... |
|---|---|---|---|
| 0 | Dieses Beispiel zeigt einmal mehr , welche gravierenden Auswirkungen der | Klimawandel | gerade auf die Meere und die marinen Ökosysteme hat . |
| 1 | Angesichts von Überfischung , Vermüllung der Meere und | Klimawandel | ist dieses Argument mehr als fragwürdig . |
| 2 | Grund dafür ist der | Klimawandel | , der den Meeresspiegel ansteigen lässt . |
| 3 | Mit Einsatz im Indischen und Pazifischen Ozean soll die FS „ Sonne “ dazu beitragen , relevante Forschungsfragen hinsichtlich des | Klimawandels | , der Versorgung mit Rohstoffen aus dem Meer und der Folgen des Eingreifens in die Ökosysteme zu beantworten . |
| 4 | Die Temperaturen steigen , der Meeresspiegel steigt , Wetterextreme häufen sich : Der | Klimawandel | ist da . |
| 5 | Der globale | Klimawandel | führt zu einer Erhöhung des Meeresspiegels , zur Häufung von Wetterextremen wie Stürmen oder Starkregen und in manchen Regionen z... |
| 6 | Dennoch kommen wir nicht an der Tatsache vorbei , dass der | Klimawandel | natürlich auch Auswirkungen auf die Meere und die Biodiversität dort hat . |
| 7 | ...elber werden immer stärker durch die Begleiterscheinungen unserer modernen Welt gefährdet : Meeres - und Umweltverschmutzung , | Klimawandel | , Beifänge in der Fischerei , Schiffsverkehr , Unterwasserlärm und Offshoreaktivitäten . |
| 8 | Angesichts der drohenden | Klimakatastrophe | , der fortschreitenden Zerstörung der Wälder und Meere , angesichts des Aussterbens zahlreicher Tier - und Pflanzenarten muß die Erha... |
| 9 | | Klimawandel | , Plastikmüll in den Meeren , Artensterben , das alles sind weltweite Probleme , die wir nur gemeinsam in der Staatengemeinschaft werden l... |
| 10 | Es gibt tatsächlich immer wieder welche , die nicht wahrhaben wollen , welche Bedrohung für die Bevölkerung mit dem | Klimawandel | einhergeht : millionenfach Klimaflüchtlinge , Anstieg der Meeresspiegel , Erderwärmung , die auch Ernteschäden nach sich zieht ; ic... |
| 11 | Die Meere sind wichtige Seismographen des | Klimawandels | und der Biodiversität . |
| 12 | ...erung von Ökosystemen nennen , auf die Zerstörung der Regenwälder hinweisen , die Überfischung der Meere anprangern , die Folgen des | Klimawandels | analysieren sowie auf das größte Artensterben seit der Zeit der Dinosaurier eingehen . |
| 13 | Wir stehen vor riesigen Herausforderungen : | Klimakrise | , Artensterben , Vermüllung der Meere mit Plastik . |
| 14 | Hinzu kommen - als wäre das alles noch nicht bedrohlich genug - die Erwärmung und Versauerung der Meere durch den | Klimawandel | . |
| 15 | Aber wir haben große Gebiete auf der Welt , die noch nicht so zugänglich sind – das sind vor allen Dingen die Meere – , die wir aber über | Klimawandel | und vieles andere schon sehr verändern . |
| 16 | Absurd ist , dass wir das Weddellmeer und andere Teile der Meere deshalb schützen müssen , weil es den | Klimawandel | gibt , weil heute bestimmte Teile der Meere eben ganz anders zugänglich sind , als sie es in der Vergangenheit waren , und wir in der Lage ... |
| 17 | ...innung von Bodenschätzen und Energie aus dem Meer , Verschmutzung und Vermüllung sowie die Erwärmung der Meere infolge des | Klimawandels | sind nur einige Stichworte . |
| 18 | Viel lesen können wir auch jeden Tag über die Auswirkungen des | Klimawandels | und die Erwärmung der Meere . |
| 19 | Die Kommission geht davon aus , dass infolge des | Klimawandels | die Niederschlagsmengen und der Meeresspiegel steigen und wetterbedingte Naturkatastrophen häufiger werden . |
| 20 | ...t es wichtig , dass wir dem Meeresschutz als Naturschutz mehr Raum geben , dass wir Programme starten , dass wir eben auch , je mehr der | Klimawandel | den Meeren zusetzt , umso mehr dafür sorgen , dass es Räume gibt , wo Meeresschutz sich entsprechend entfalten kann . |
| 21 | Wer sich mit dem Bericht des Weltklimarates auseinandergesetzt hat , weiß , dass der | Klimawandel | zu einem Anstieg der Meeresspiegel führen wird . |
| 22 | Erzählen Sie doch mal einer Familie auf Langeoog , deren Haus im Meer versinkt , es gäbe keinen | Klimawandel | ! |
| 23 | So können genau diese Bewohner der Küsten auch vor dem | Klimawandel | und den daraus resultierenden katastrophalen Folgen des Anstiegs des Meeresspiegels geschützt werden . |
| 24 | Betroffen ist das Wattenmeer in dramatischer Form – das ist gerade schon angesprochen worden – vom | Klimawandel | , insbesondere vom Anstieg des Meeresspiegels , der diesen Raum , der sich über einen Zeitraum von mehr als 7 000 Jahren dort gebil... |

# Keywords

- *Keywords* sind Wörter, die in einem gegebenen Korpus überdurchschnittlich oft vorkommen – im Vgl. zur Häufigkeit in einem *Referenzkorpus*

- *Keyness* ist ein textuelles, kein sprachliches Feature
  - gesprochene Sprache vs. geschriebene Sprache
  - soziale Medien vs. Zeitungen
  - Hochliteratur vs. Groschenromane
  - links-liberale Zeitungen vs. rechts-konservative Zeitungen
  - Grüne vs. AfD
  - ...

- Anwendung bspw. in der *Diskursanalyse*, *Indexerstellung*, ...

| rank | lemma | LRC |
|---|---|---|
| 1 | Klimahysterie | 6.96 |
| 2 | Altpartei | 6.95 |
| 3 | Masseneinwanderung | 6.44 |
| 4 | 0,000653 | 6.06 |
| 5 | Lockdown | 5.78 |
| 6 | Klimareligion | 5.62 |
| 7 | Timon | 5.61 |
| 8 | Windindustrieanlage | 5.58 |
| 9 | Gremmels | 5.53 |
| 10 | Nullzinspolitik | 5.45 |
| 11 | AfD | 5.44 |
| 12 | Altfraktion | 5.34 |
| 13 | Coronamaßnahmen | 5.33 |
| 14 | Antifa | 5.32 |
| 15 | Shutdown | 5.27 |
| 16 | Massenmigration | 5.25 |
| 17 | Coronapolitik | 5.24 |
| 18 | globalistisch | 5.17 |
| 19 | Lockdowns | 5.16 |
| 20 | Greta | 5.15 |

Tabelle: Keywords der AfD (GermaParl).

# Keywords

## GermaParl (1949–2021) vs. SZ

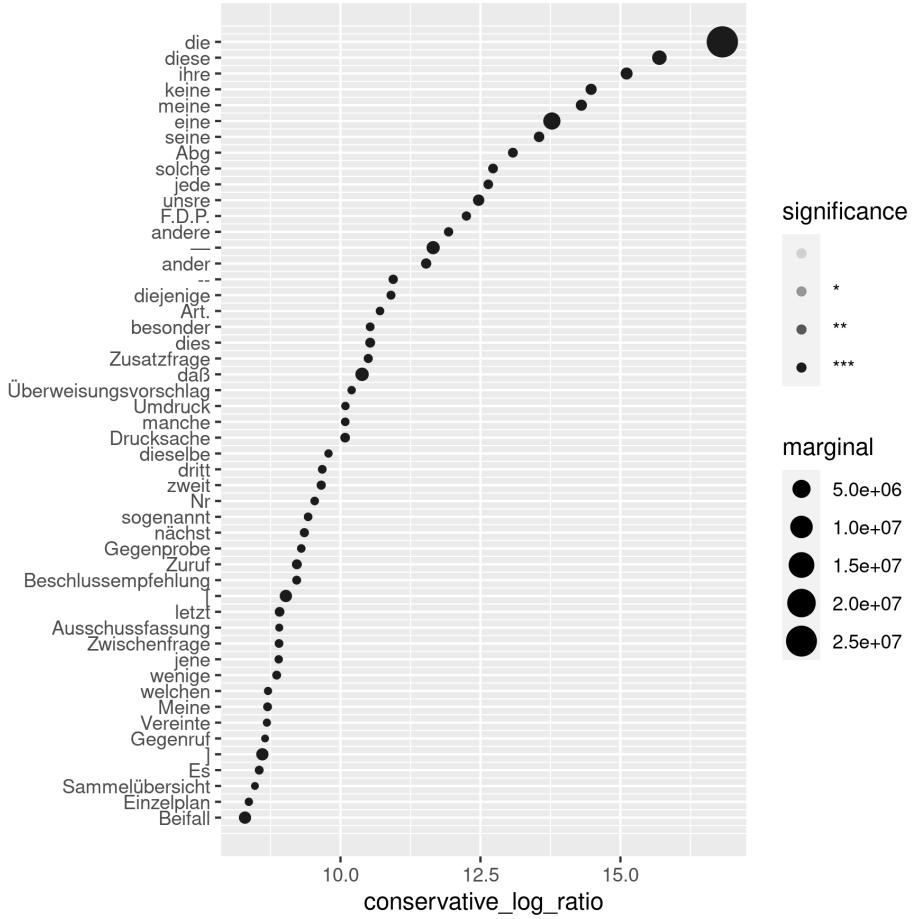



Abbildung: Keywords bei Referenz SZ 2011–2014

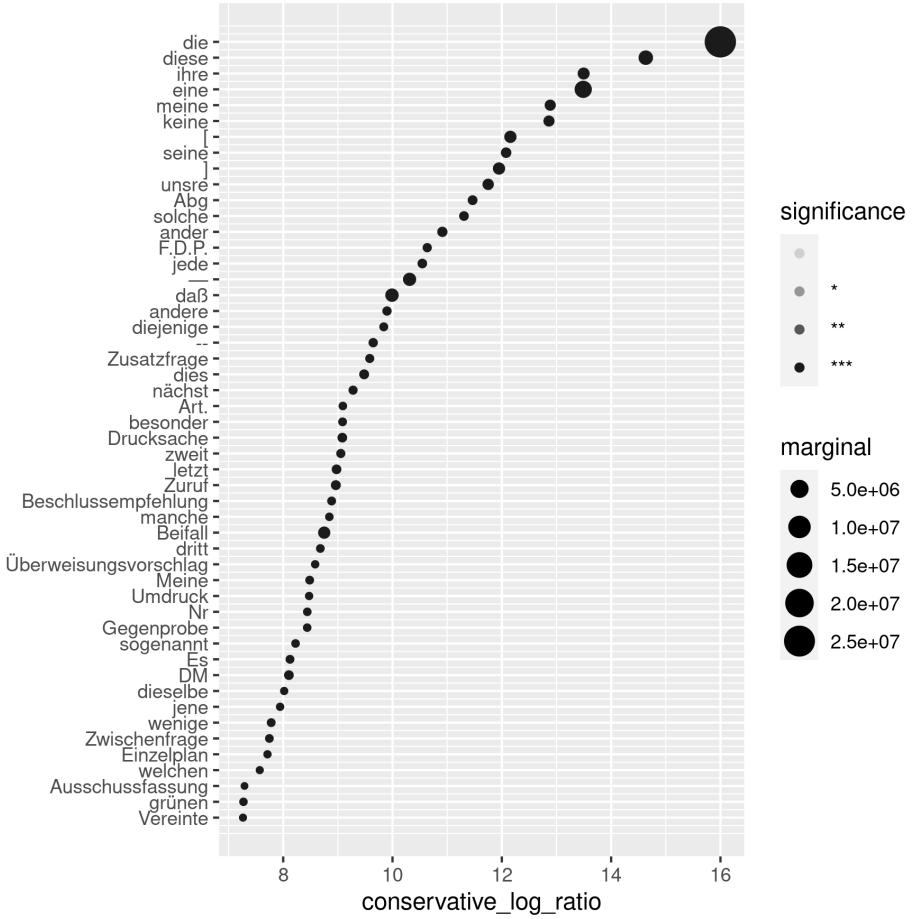
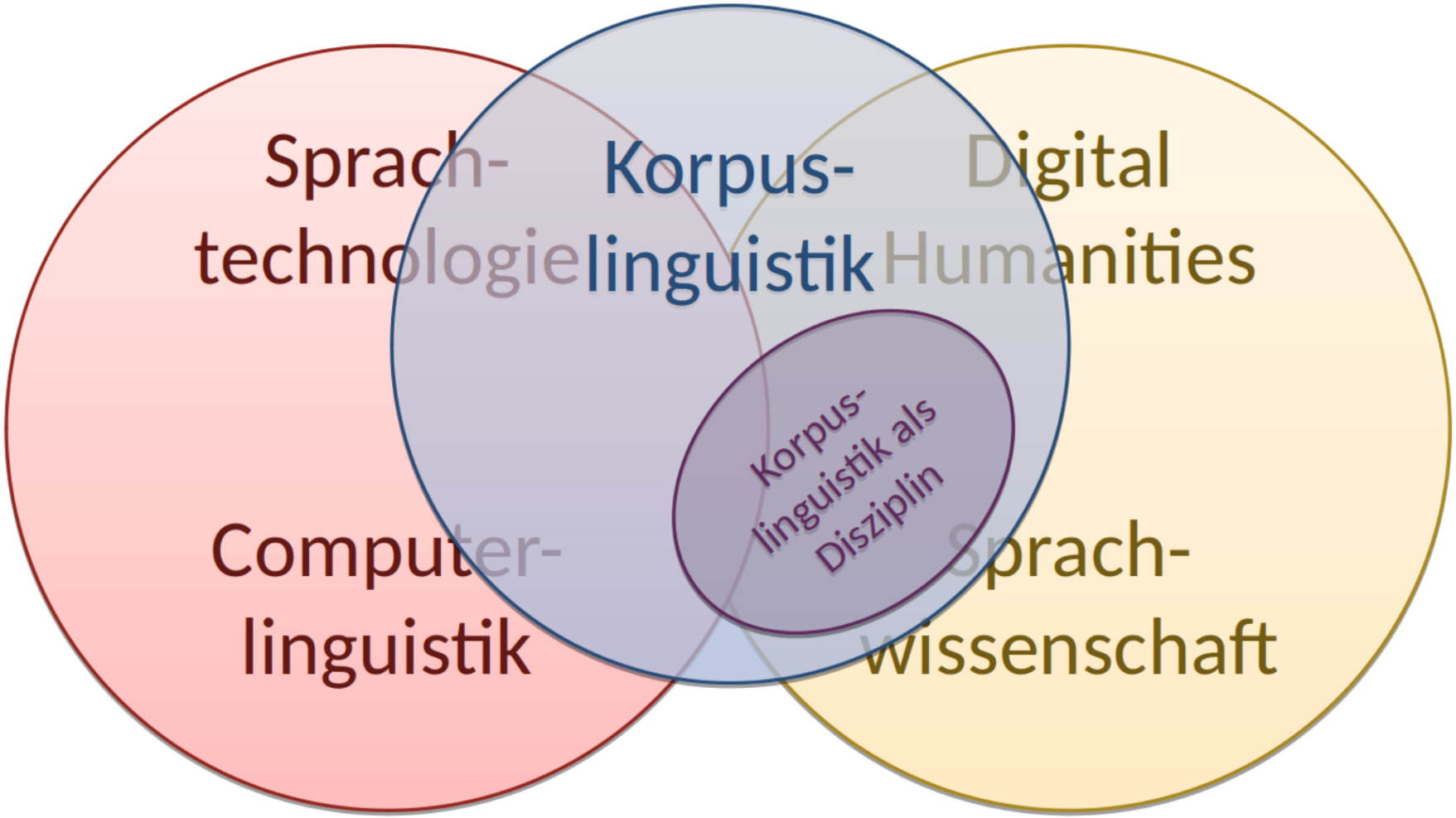



Abbildung: Keywords bei Referenz SZ 2019–2021

# Was ist Korpuslinguistik?



# Was ist Korpuslinguistik?

Korpuslinguistik als Methodenwissenschaft
(corpus-based research)

- Erstellung von Korpora als Stichproben
  (Grundgesamtheit, Stichprobenplanung, . . . )
- Automatische linguistische Annotation
  (schwierig z.B. für literarische oder historische Texte)
- Werkzeuge für Query & Visualisierung
- Quantitative Auswertung mit statistischen Tests
  $\rightarrow$ methodologische Schwierigkeiten
- Statistische Modellierung & explorative Verfahren

Korpuslinguistik als Teildisziplin der Sprachwissenschaft
(corpus-driven research)

- Britische Forschungstradition seit 1950er Jahren
  (Firth, Sinclair, Halliday, Leech, McEnery, . . . )
- Keywords, Kollokationen, semantische Prosodie, . . .
- Auch qualitative Analyse von Konkordanzen
  (John Sinclair: *trust the text!*)
- Vorwiegend deskriptiv, Fokus auf Anwendungen
- Themen: Sprachvarietäten, soziologische und politische
  Aspekte, Spracherwerb und -unterricht, . . .

# Anwendungsgebiete der Korpuslinguistik

- Dialektologie, kontrastive Linguistik, Kulturwissenschaft
- Fremdsprachenunterricht & Spracherwerb ($\rightarrow$ CALL)
- Grammatik & Syntax ($\rightarrow$ Quirk/Greenbaum CGEL)
- Historische Sprachwissenschaft & Sprachwandel
- Lexikalische Semantik ($\rightarrow$ semantische Prosodien)
- Lexikographie & Lexikologie ($\rightarrow$ COBUILD, Kollokations-WB)
- Morphologie ($\rightarrow$ Produktivitätsbegriff)
- Phonologie ($\rightarrow$ Sprachatlas)
- Pragmatik & Diskursanalyse ($\rightarrow$ Rhetorik, Soziologie, Politik)
- Soziolinguistik ($\rightarrow$ Gender Studies, Sprachideologie)
- Sprachdokumentation ($\rightarrow$ bedrohte Sprachen)
- Sprachvariation ($\rightarrow$ z.B. Registeranalyse nach Biber)
- Stilometrie & Literaturwissenschaft ($\rightarrow$ Autorenerkennung)
- Übersetzungswissenschaft ($\rightarrow$ „Translationese“)

# Tracking the Infodemic

- Tracking the Infodemic: Automatische Identifikation von Verschwörungstheorien in sozialen Medien

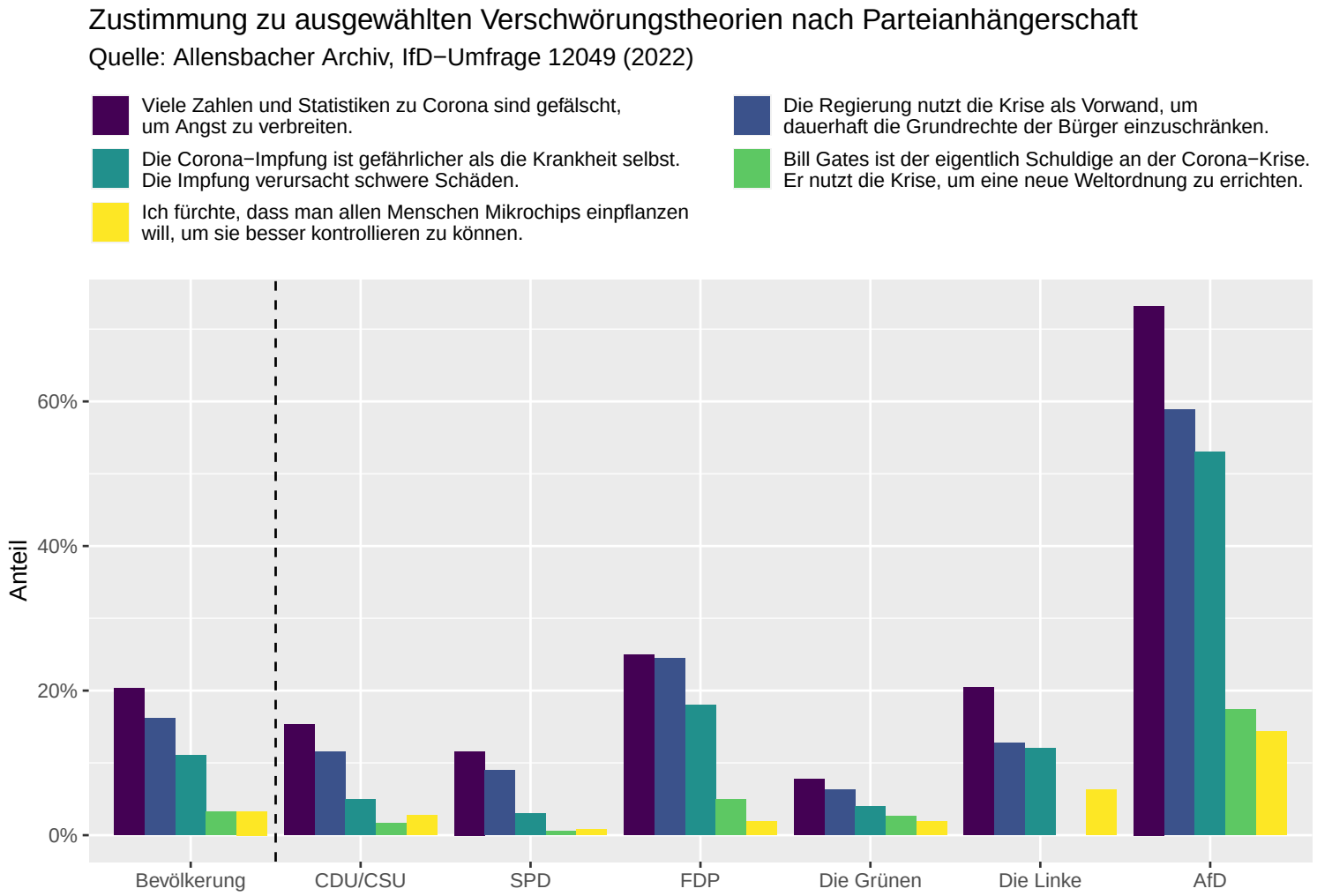
- WHO warnt 2020: „Infodemie“ neben der Pandemie
  - Verschwörungstheorien
  - Misrepräsentation von medizinischen Fakten
  - Desinformation

- vorhersehbar:
  - Debatten finden unmoderiert online statt
  - Glaube an Verschwörungstheorien ist stärker in Zeiten von Unsicherheit („Wissensvakuum“)

- in Deutschland glaubt 1 von 5, dass die
  - Gefahren von SARS-CoV-2 absichtlich übertrieben wurden und (Infratest dimap)
  - Zahlen und Statistiken gefälscht sind (Institut für Demoskopie Allensbach)

# Tracking the Infodemic

## Eine Infodemie während einer Pandemie



# Tracking the Infodemic

## Daten

### Datensammlung

- Identifikation bekannter Verschwörungstheoretiker (KenFM, Eva Herman, Attila Hildmann, Bodo Schiffmann, Oliver Janich usw.)
- Scraping ausgewählter, öffentlicher Kanäle und verknüpfter Chat-Gruppen
- Ergänzung des Korpus um häufig genannte/zitierte Kanäle (mehrmals)

### Korpus: Infodemic-Telegram-v1

- über 130 Kanäle, 20 öffentliche Gruppenchats
- ca. 4,5 Mio. Beiträge
- ca. 150 Mio. Token

| | | |
|---|---|---|
| 1 | Demokratur & Systemmedien | 130 |
| 2 | Pseudopandemie | 78 |
| 3 | Rechtsaußen | 77 |
| 4 | QAnon | 75 |
| 5 | Corona-Verschwörungsnarrative | 73 |
| 6 | Maßnahmenkritik | 43 |
| 7 | Impfängste | 40 |
| 8 | Esoterik & Medizingeschwurbel | 39 |
| 9 | allgemeine Verschwörungsnarrative | 31 |
| 10 | Schlafschafe | 19 |
| 11 | Millenarismus / Tag der Abrechnung | 14 |
| 12 | Alternative Gegenmittel | 10 |

Tabelle: Häufigkeiten von Narrativgruppen in 500 zufälligen Texten

## Zero-Shot-Klassifikation: Satzähnlichkeiten

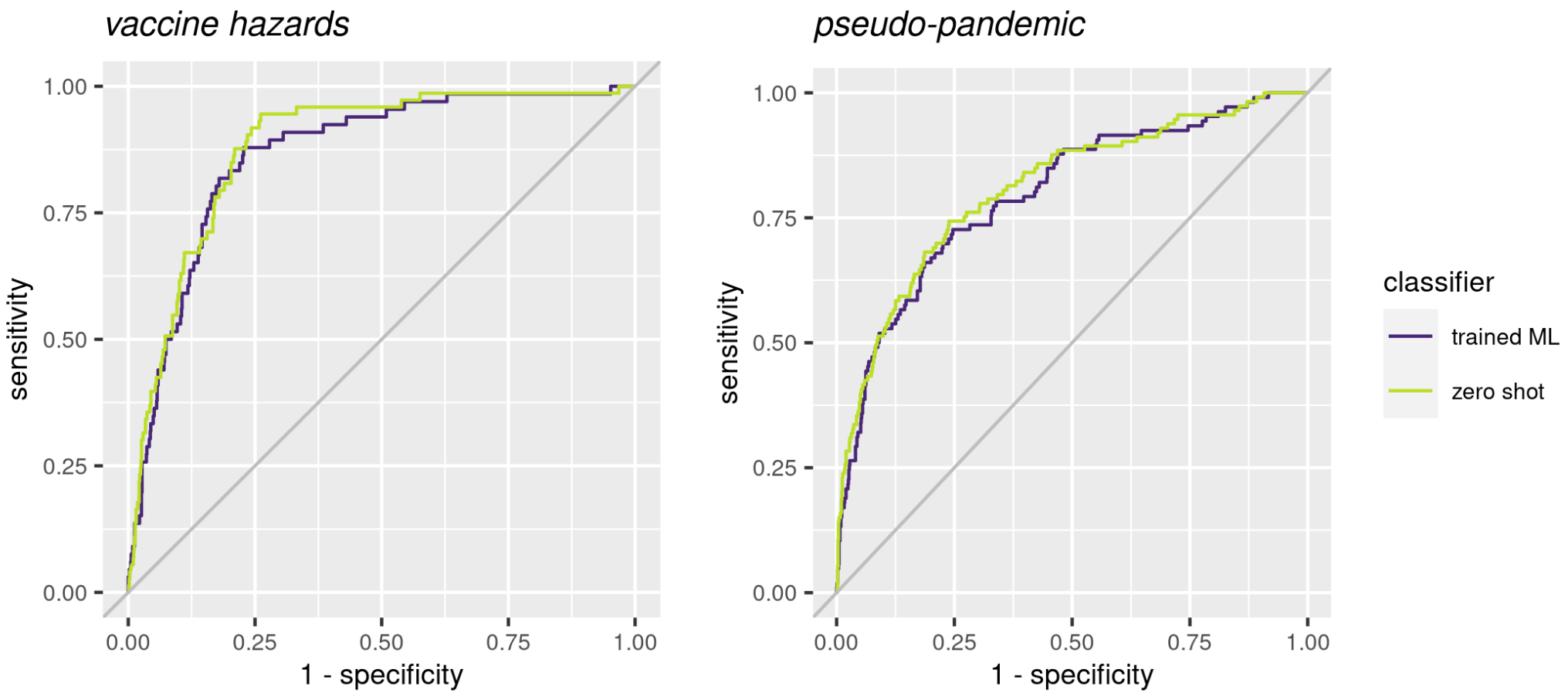
- Idee: Vergleiche Sätze der Posts mit Narrativ-Beschreibungen

| Satz | Narrativ | Ähnlichkeit |
|---|---|---|
| Glaubt mir es gibt kein Coronavirus, das ist kompletter Schwachsinn. | Corona gibt es nicht. | 0.82 |
| Die Impfungen sind keine Viren Schutzimpfungen sondern tödlich! | Impfung ist gefährlich. | 0.76 |
| Impfstoffbedingte Schäden an unseren Blutgefäßen sind häufig. | Impfung ist gefährlich. | 0.74 |
| Die neue Weltordnung - Great Reset wartet. | Pandemie ist geplant. | 0.53 |
| ... | ... | ... |

- Satzähnlichkeiten via Kosinusähnlichkeit von Satz-Embeddings
  - Embedding = hochdimensionaler Vektor, der die Semantik des Satzes anhand von Ko-Okkurrenz-Mustern einfängt

# Tracking the Infodemic

## Zero-Shot-Klassifikation: Evaluation

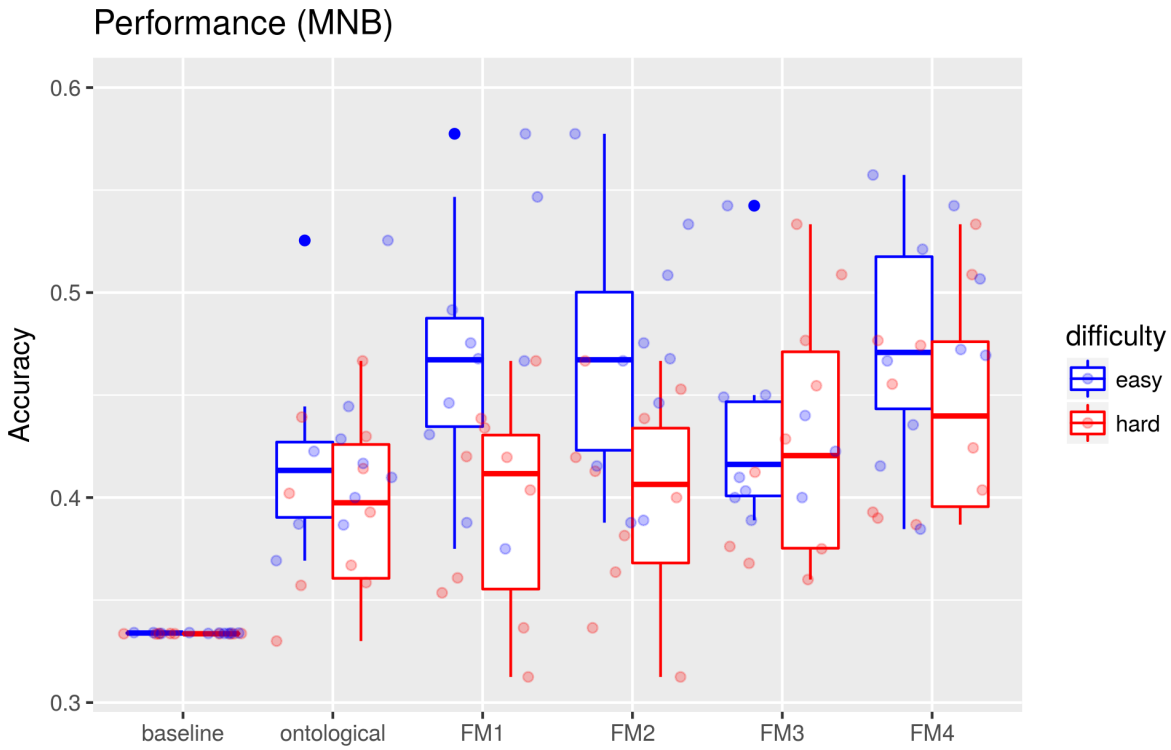


## Show Case: Analyse von Bundestagsdebatten

# Financial Narrative Processing

## Analyse von AdHoc-Mitteilungen

- Untersuchungsgegenstand: AdHoc-Mitteilungen
  - Unternehmensberichte bei Ereignissen, die den Aktienkurs beeinflussen können
- Daten
  - gut 30,000 Mitteilungen von deutschen Unternehmen
  - kumulative abnormale Rendite bei Veröffentlichung
- Vorhersage des Aktienkurses
  1. Ontologie-basierte Identifikation von Themen
  2. Integration der Merkmale in klassische MLV



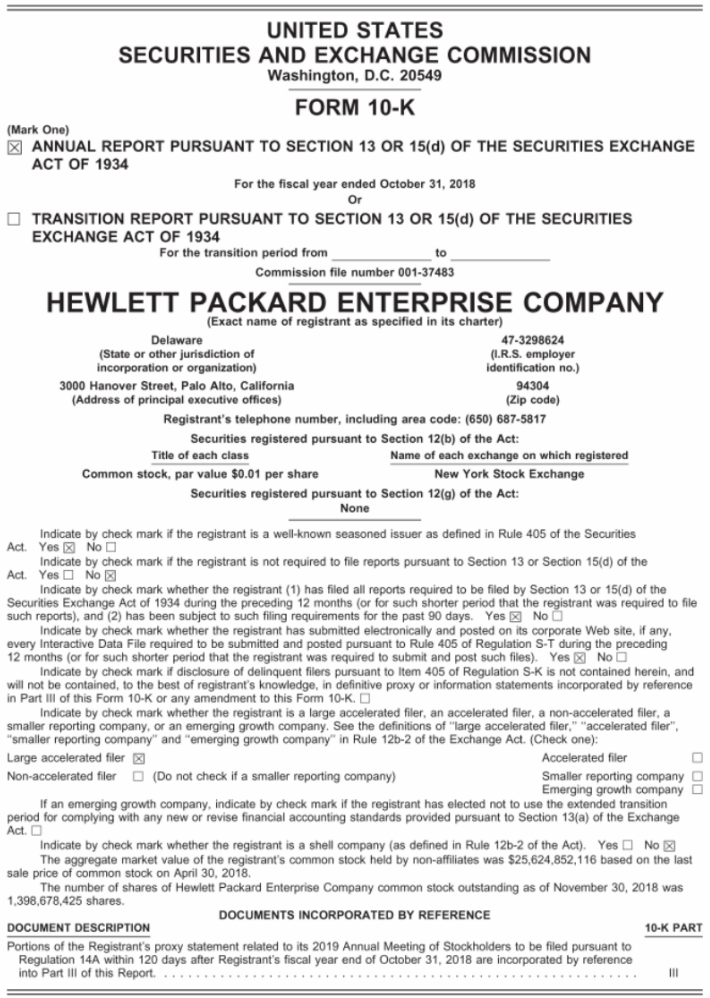
# Financial Narrative Processing

## Analyse von Jahresberichten

- Untersuchungsgegenstand: Jahresberichte in den USA
  - eingereicht via SEC's EDGAR filing system ("form 10-k")
  - Format: XBRL + Fließtext (ohne semantisches Mark-Up)

- Daten
  - knapp 80,000 Berichte (2006 – 2015)
  - Informationen über Unternehmen (bspw. Industrie-Kategorisierung)

- Identifikation von *red flags*
  - „zwischen den Zeilen lesen"
  - stilometrische Merkmale der einzelnen Abschnitte



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)
☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended October 31, 2018
Or

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from _____ to _____

Commission file number 001-37483

HEWLETT PACKARD ENTERPRISE COMPANY
(Exact name of registrant as specified in its charter)

| Delaware | 47-3298624 |
|---|---|
| (State or other jurisdiction of incorporation or organization) | (I.R.S. employer identification no.) |
| 3000 Hanover Street, Palo Alto, California | 94304 |
| (Address of principal executive offices) | (Zip code) |

Registrant's telephone number, including area code: (650) 687-5817

Securities registered pursuant to Section 12(b) of the Act:

| Title of each class | Name of each exchange on which registered |
|---|---|
| Common stock, par value $0.01 per share | New York Stock Exchange |

Securities registered pursuant to Section 12(g) of the Act:
None

Indicate by check mark if the registrant is a well-known seasoned issuer as defined in Rule 405 of the Securities Act. Yes ☒ No ☐

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act. Yes ☐ No ☒

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes ☒ No ☐

Indicate by check mark whether the registrant has submitted electronically and posted on its corporate Web site, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T during the preceding 12 months (or for such shorter period that the registrant was required to submit and post such files). Yes ☒ No ☐

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K is not contained herein, and will not be contained, to the best of registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K. ☐

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer", "smaller reporting company" and "emerging growth company" in Rule 12b-2 of the Exchange Act. (Check one):

| | |
|---|---|
| Large accelerated filer ☒ | Accelerated filer ☐ |
| Non-accelerated filer ☐ (Do not check if a smaller reporting company) | Smaller reporting company ☐ |
| | Emerging growth company ☐ |

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revise financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. ☐

Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Act). Yes ☐ No ☒

The aggregate market value of the registrant's common stock held by non-affiliates was $25,624,852,116 based on the last sale price of common stock on April 30, 2018.

The number of shares of Hewlett Packard Enterprise Company common stock outstanding as of November 30, 2018 was 1,398,678,425 shares.

DOCUMENTS INCORPORATED BY REFERENCE

| DOCUMENT DESCRIPTION | 10-K PART |
|---|---|
| Portions of the Registrant's proxy statement related to its 2019 Annual Meeting of Stockholders to be filed pursuant to Regulation 14A within 120 days after Registrant's fiscal year end of October 31, 2018 are incorporated by reference into Part III of this Report. | III |

### Part I

- Item 1 **Business**
- Item 1A **Risk Factors**
- Item 1B Unresolved Staff Comments
- Item 2 **Properties**
- Item 3 **Legal Proceedings**
- Item 4 Mine Safety Disclosures

### Part II

- Item 5 **Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases**
- Item 6 Selected Financial Data
- Item 7 **Management's Discussion and Analysis of Financial Condiftion and Results of Operations**
- Item 7A **Quantitative and Qualitative Dislosures About Market Risk**
- Item 8 **Financial Statements and Supplementary Data**
- Item 9 Changes in and Disagreements With Accountants on Accounting and Financial Disclosure
- Item 9A **Controls and Procedures**
- Item 9B Other Information

### Part III

- Item 10 **Directors, Executive Officers and Corporate Governance**
- Item 11 Executive Compensation
- Item 12 Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
- Item 13 Certain Relationships and Related Transactions, and Director Independence
- Item 14 Principal Accounting Fees and Services

### Part IV

- Item 15 Exhibits, Financial Statement Schedules
- Item 16 Form 10-K Summary

# Quantitative linguistische Merkmale

## Sentiment & Subjectivity

- meist Wörterbuch-basiert
- zusätzlich: WSD, Negationserkennung
- viele frei verfügbare Module

## Readability

- verschiedene Maße, basierend auf
  - Anzahl Wörter pro Satz
  - Anzahl Silben pro Wort
- bspw. Flesh-Kincaid grade-level (Kincaid 1975):

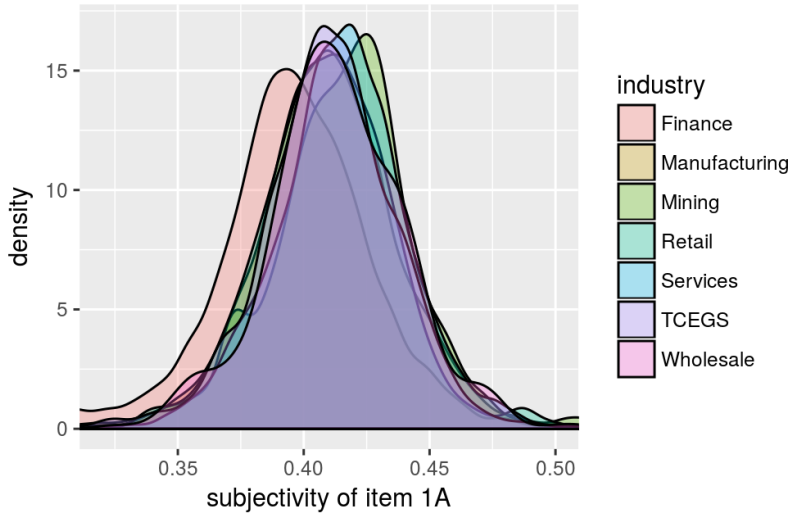
$$\ell_{\text{fk}} = 0.39\,(n_w/n_s) + 11.8\,(n_y/n_w) - 15.59$$

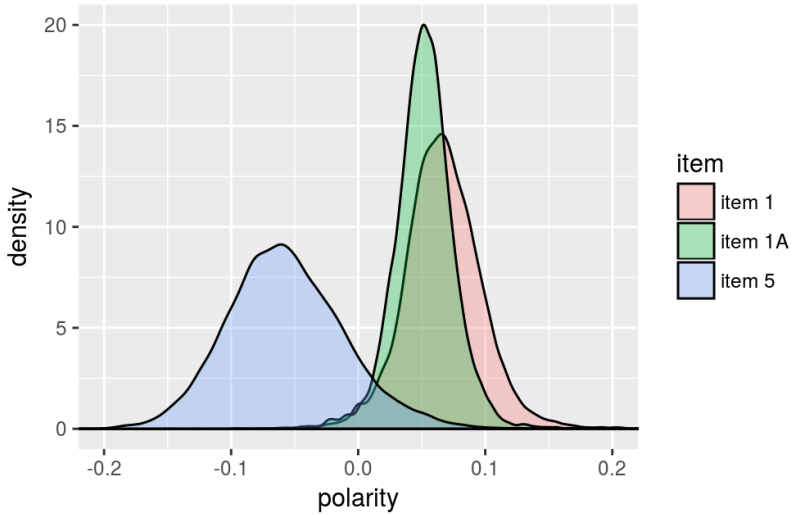
## Stilistische Merkmale

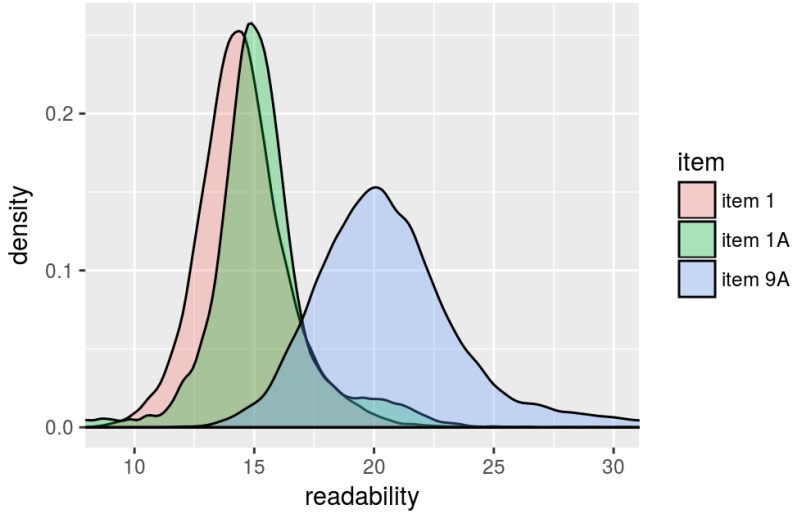
- Biber (1998): *Variation across Speech and Writing: The Multi-dimensional Approach to Linguistic Analyses*
- Dimensionen:
  1. the opposition between **involved** and **informational** discourse
  2. the opposition between **narrative** and **non-narrative** concerns
  3. the opposition between **context-independent** and **context-dependent** discourse
  4. overt expression of **persuasion**
  5. opposition between **abstract** and **non-abstract** information
  6. **on-line informational elaboration**
- Implementierungen frei verfügbar (z.B. Andrea Nini)
- basierend auf Stanford Tagger (Merkmale = Sequenzen)

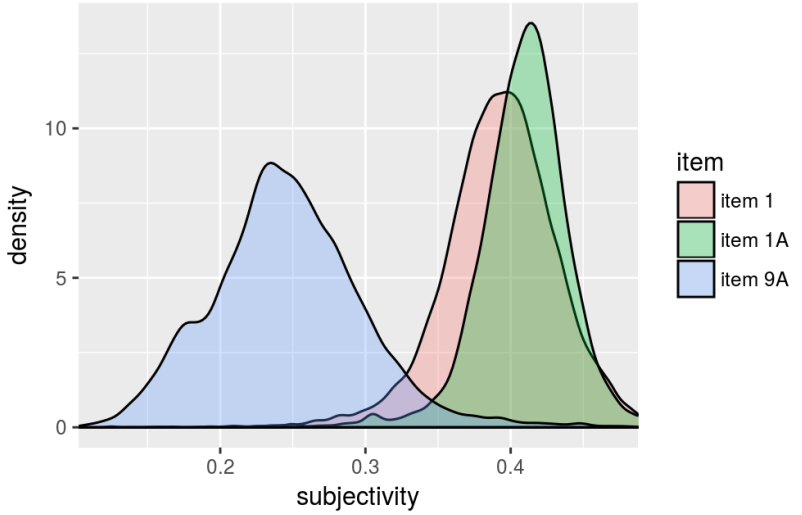
# Financial Narrative Processing

## Deskriptive Statistiken: Sentiment & Readability



# Argumentation Mining

## Reconstructing Arguments from Noisy Texts

- Argumentation Mining
  - automatische Extraktion von Argumentationsstrukturen aus natürlichsprachlichem Text
  - Strukturen: Prämissen, Folgerungen, . . . : logische Schemata

- hier: Identifikation von Argumentfragmenten auf Twitter
  - Fallstudie: Brexit-Tweets (bis zur Abstimmung am 23.06.2016)
  - Ziel: Erkundung und Formalisierung der argumentativen Landschaft auf Twitter
  - linguistische Oberflächenrealisierung $\rightarrow$ logische Muster
  - Precision wichtig für automatische Extraktion (Fehlerpropagation)

- höchst informelle Argumentation in sozialen Medien (*defeasible logic*):
  - keine klassischen Schemata (Modus Ponens, Modus Tollens)
  - „objektive Wahrheit“ scheint irrelevant (*anything goes*)
  - stattdessen: alltägliche Kommunikation (argumentation by authority, ad hominem, etc.)
  - Twitter: zu kurz für volle Argumente (d. h. Prämissen und Konklusionen)
  - hier: *implizite Konklusion*, dass die UK die EU verlassen soll (oder nicht)

# Textbeispiele

- we European knows it and see our culture disappearing
- the #UN says that most of them are NOT ’refugees’ at all

- Today cast a vote that you can make a "that’s what she said" joke about. Don’t pull out #BetterIn #Brexit
- Brexit is still better than being in the EU
- pensionists stand to lose from Brexit

- this lie that young people can’t travel, it’s another stay lie
- more bullshit from the in camp, next they’ll be claiming that Brexit ’could’ cause Mars to crash in to Venus lol

- argumentation by *authority*
  - quotes
  - ascribing knowledge or expertise to authorities
  - etc.

- statements of *evaluation*
  - as good or bad
  - either for some or for all
  - with or without hedging and / or comparison to other statements

- *ad-hominem*, i. e. discrediting others by alleging
  - incompetence
  - conspiracy
  - scaremongering or plain lying

# Logische Muster

- *desire*: truth of some formula 1 is desirable for some entity 0 $D_{?0:entity}(?1 : formula)$
- *bad for concept*: formula 1 is/would be bad for concept 0 $badFor(?0 : concept, ?1 : formula)$
- *warning*: entity 0 warns of formula 1 being bad $Q_{?0:entity}bad(?1 : formula)$
- *membership*: entity 0 is part of entity 1 $?0 : entity \implies ?1 : entity$
- *lying*: entity 0 lies about formula 1 $Q_{?0:entity}(?1 : formula) \wedge \neg B_{?0}(?1)$
- *ad-hominem*: entity 0 is morally bad and hence discredited $\neg moral(?0 : entity)$
- ...

# Manuelle Annotation

- Erarbeitung *detaillierter Guidelines* für alle Muster
- Annotation von Tweets bzw. Sätzen bzgl. des Vorkommens der Muster (und der realisierten *Slots*)
- Recall: *zuällige Auswahl* von Sätzen aus dem Gesamtkorpus
- Precision: Annotation von Treffern des Verfahrens (*Matches*)

| | | **desire** | | | **lying** | | | **warning** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | gold | PH | MG | gold | PH | MG | gold | PH | MG |
| random | ND | 0.673 | 0.456 | 0.521 | 0.722 | 0.511 | 0.588 | 0.902 | 0.667 | 0.682 |
| | MG | 0.661 | 0.523 | | 0.659 | 0.713 | | 0.713 | 0.747 | |
| | PH | 0.487 | | | 0.701 | | | 0.761 | | |
| matches | ND | 0.564 | 0.463 | 0.229 | 0.611 | 0.648 | 0.415 | 0.430 | 0.449 | 0.337 |
| | MG | 0.439 | 0.311 | | 0.777 | 0.713 | | 0.961 | 0.430 | |
| | PH | 0.527 | | | 0.765 | | | 0.500 | | |

Tabelle: Pairwise kappa scores for inter-annotator agreement and agreement with the adjudicated gold data.

# Extraktion mittels Corpus Queries

- maßgeschneiderte Queries versuchen je eine spezielle Gruppe von Oberflächenrealisierungen eines logischen Musters abzudecken
  - Korpora in CWB indexiert (CQP-Syntax)
  - Modularität: wiederverwendbare Wortlisten und Macros mit domänenspezifischen Wörtern
  - grammatische Muster via POS-Sequenzen
- Grundgedanke:
  - Nutzung linguistischen Wissens hilft bei der Analyse von *Noisy Text*
  - Queries können explizit so konzipiert werden, dass sie Phänomene der internetbasierten Kommunikation abdecken
  - keine Nutzung bspw. von Depenzenparsing nötig (funktioniert nicht auf Tweets!)
- Queries werden iterativ entwickelt (Fallbeispieldiskussion)
- *high precision*, *low recall*
- Recall
  - `random-1000`
  - nur auf Ebene der logischen Muster möglich
- Precision
  - `query-matches`
  - qualitativ via Konkordanzen
- sehr arbeitsaufwändig
  - insb. bei veränderlichem Muster-Repertoire

# Example Query

```
<np>
  @0:[::]
  /actor_np_all[]
  @1:[::]
</np>
<vp>
  [lemma=$verbs_prefer]
  ("to" "have|get")?
</vp>
<np>
  @2:[::]
  []+
</np>
@3:"back"
within tweet;
```

Beschreibung:

- `/actor_np_all[]`: macro mit optionalen *determiners* and *modifiers*, gefolgt von einer *Entität*, bspw. Eigennamen oder spez. Wortliste(n)
- `$verbs_prefer`: Liste von Verben, die *desire* (*want, wish, prefer, hope*...) ausdrücken
- Tokenspannen @0...@1 und @2...@3: *Slots* des Musters (*entity*, *formula*)

Beispieltreffer:

- Yes *we do want our fish back*. They take our money and our fish and then tells us what to do! #brexit
- *We want our borders back*, *we want our democracy back*, *we want our country back* ! #BrexitBusTour
- don't think *most #Brexiteers want an empire back* just their freedom and democracy #brexit

# Evaluation

| | TP | FP | precision | $n_{\text{matches}}$ | $n_{\text{estimated}}$ |
|---|---|---|---|---|---|
| **pattern #3** | | | **0.95** | 19709 | 18723 |
| – entity supports x | 47 | 3 | 0.94 | 14386 | 13522 |
| – entity favours 1 implicit 2 | 72 | 0 | 1.00 | 9570 | 9570 |
| – ... | ... | ... | ... | ... | ... |
| **pattern #16** | | | **0.73** | 116 | 84 |
| – 2 is better than 0 implicit entity | 40 | 1 | 0.98 | 41 | 40 |
| – entity favours policy over 2 | 14 | 18 | 0.44 | 32 | 14 |
| – ... | ... | ... | ... | ... | ... |
| **pattern #20** | | | **0.91** | 2908 | 2646 |
| – entity's lie about | 162 | 17 | 0.91 | 2802 | 2549 |
| – another entity lie | 35 | 16 | 0.69 | 87 | 60 |
| – ... | ... | ... | ... | ... | ... |
| **pattern #38** | | | **0.99** | 6133 | 6071 |
| – entity warns that 1 | 50 | 0 | 1.00 | 4800 | 4800 |
| – entity says X harms y | 47 | 3 | 0.94 | 671 | 630 |
| – ... | ... | ... | ... | ... | ... |

Tabelle: Precision of queries and patterns estimated on random selections of matches for each query. $n_{\text{matches}}$ is the number of matches in the whole corpus, $n_{\text{estimated}}$ corresponds to this number corrected by precision.

| | TP | FP | TN | FN | recall |
|---|---|---|---|---|---|
| **pattern #3** | 11 | 3 | 922 | 64 | **0.147** |
| – adj. | 4 | 1 | 471 | 24 | 0.143 |
| – maj. | 7 | 2 | 451 | 40 | 0.149 |
| **pattern #16** | 2 | 0 | 975 | 23 | **0.080** |
| – man. | 2 | 0 | 481 | 17 | 0.105 |
| – aut. | 0 | 0 | 494 | 6 | 0.000 |
| **pattern #20** | 2 | 0 | 971 | 27 | **0.069** |
| – adj. | 1 | 0 | 492 | 7 | 0.125 |
| – maj. | 1 | 0 | 479 | 20 | 0.048 |
| **pattern #38** | 9 | 0 | 935 | 56 | **0.138** |
| – adj. | 4 | 0 | 478 | 18 | 0.182 |
| – maj. | 5 | 0 | 457 | 38 | 0.116 |

Tabelle: Estimation of recall of logical patterns on random-1000.

# Spheroscope

- *off-the-shelf* keine passenden Werkzeuge verfügbar
  → spheroscope: a web app for argumentation mining via corpus queries

- Features
  - Definition und Management der *Queries*, *Wordlisten* und *Macros*
  - Query-Kategorisierung logischen Mustern
  - erweiterte *Konkordanzen*
  - halb-automatische Erweiterung von *Wortlisten*

- *Query Refinement*
  - welche Tweets werden durch Update entfernt / kommen hinzu?
  - detaillierter Vergleich auf Tokenebene
  - quantitative Query-Evaluation

# Spheroscope

| ID | Name | #Q | Formula |
|---|---|---|---|
| 0 | quotation | 7 | $Q_{?0:entity}(?1 : formula)$ |
| 1 | explanation | 2 | $(?0 : formula) \square\!\!\Longrightarrow (?1 : formula)$ |
| 2 | causal implication | 0 | $(?0 : formula) \overset{c}{\Longrightarrow} (?1 : formula)$ |
| 3 | desire | 18 | $D_{?0:entity}(?1 : formula)$ |
| 4 | ought | 0 | $O(?0 : formula)$ |
| 5 | possibility | 1 | $\diamond(?0 : formula)$ |

Abbildung: Pattern Overview

| Item | Frequency | Similarity ▲ |
|---|---|---|
| useful | 958 | 0.684856 |
| important | 7143 | 0.658682 |
| difficult | 1977 | 0.650458 |
| challenging | 42 | 0.636462 |
| consistent | 272 | 0.629459 |
| sensible | 1763 | 0.625485 |
| valuable | 199 | 0.620347 |
| satisfactory | 18 | 0.615404 |
| interesting | 9016 | 0.612765 |
| profitable | 106 | 0.612587 |

Abbildung: Similarity View

# Spheroscope: Concordancing

| Index ▾ | whole | 0 | 1 | tweet | TP |
|---|---|---|---|---|---|
| ('pattern3_entity_wants_their_concept_back', 30776292, 30776297) | From brexit to @realDonaldTrump everyone wants their country back | @realDonaldTrump | their country back | t746036667706916864 | ? |
| ('pattern3_entity_wants_their_concept_back', 4216663, 4216667) | I want my country back . So I'm voting remain . #VoteRemain https://t.co /q6vKo77x6T | I | my country back | t743734737106141184 | True |
| ('pattern3_entity_wants_their_concept_back', 14983065, 14983073) | I just wish #ukip would stop taking everything back to #brexit the NHS is another subject this isn't an EU Debate ! #bbcqt #EUreferendum | I | #ukip would stop taking everything back | t733426508254023680 | True |
| ('pattern3_entity_wants_their_concept_back', 1600174, 1600178) | @Col_Irrelevant resolve with the EU ? Don't make me laugh . Do you even know those in charge ? We want our sovereignty back | We | our sovereignty back | t745527503217795076 | True |
| ('pattern3_entity_wants_their_concept_back', 912895, 912902) | @Iron_Spike as I understand it , Brexit is mainly supported by 60-somethings longing back to an era of imperial britain | Brexit | by 60-somethings longing back | t744846007075209216 | False |

Abbildung: Concordance View

# Deutsche Daten

```
<np>
  @0:[::]
  /actor_np_all[]
  @1:[::]
</np>
<vp>
  [lemma=$verbs_prefer]
</vp>
<np>
  @2:[::]
  []+
</np>
<vp>
  "wieder.*|zurück.*"
  []*
  @3:[::]
</vp>
within s;
```

Beispiele aus der SZ:

- *Er* will *sein altes Leben zurück*
- *Der Staat* will *die Konsummaschine wieder anwerfen*
- deshalb wollen *wir nächste Saison wieder Erstligaspiele haben*
- Als unabhängiger *Küstenstaat* will *das Vereinigte Königreich wieder voll und ganz die Kontrolle über seine Gewässer haben*

Schwierigkeiten bei der Übertragung:

- einige Queries sind sehr stark an das Zielkorpus angepasst
- Macros und Wortlisten müssen komplett neu geschrieben werden
- Wortstellung im Deutschen viel freier
- Zeitungstexte komplizierter als Tweets
- weniger NLP-Werkzeuge für Deutsch verfügbar

# Fin

Vielen Dank für Ihre Aufmerksamkeit!

. . . Zeit für Diskussion . . .