Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Was macht die Korpus- und Computerlinguistik?

Lehrstuhl für
Korpus- und
Computerlinguistik

Steffen Bothe • Bao Minh Doan Dang • Prof. Dr. Stephanie Evert • Philipp Heinrich • Mahdi Mantash • Naveed Unjum • Timm Weber
Lehrstuhl für Korpus- und Computerlinguistik | www.linguistik.fau.de

Korpusbasierte Diskursanalyse: Sprache, Macht und Gesellschaft

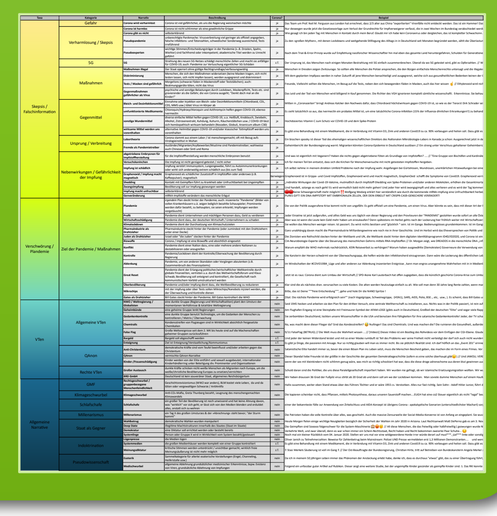
- Wie macht Sprache Machtverhältnisse sichtbar?
- Wie werden Meinungen und Positionen verhandelt?
- Themen: Gender & Diversity, Migration, Populismus, …
- Linguistische Muster: typische Wortwahl, Framing
- Methoden: Kollokationen, Keywords, Konkordanzen, Häufigkeitsverteilungen
- MMDA: Diskurseme und Konstellationen

ID		Context	Keyword	Context	
8,993,687		... Semrat Sengupta , Direktor für	Klimawandel	und erneuerbare Energie am Zentrum für ...	
6,772,680		... Energien eine sinnvolle Maßnahme ist , um den	Klimawandel	zu begrenzen . 90 Prozent halten strengere ...	
12,672,462		... " effizienter und schneller " im Kampf gegen den	Klimawandel	zu werden , müsse man insbesondere den Ausbau ...	
10,011,841		... Anteil der Erneuerbaren gesunken Angesichts des	Klimawandels	müssen die erneuerbaren Energien stark ...	
17,497,967		... Dollar waren zudem für den Kampf gegen die	Klimakrise	eingeplant , darunter Investitionen in erneuerbare ...	
16,641,056		... im Wahlkampf hatte Albanese versprochen , die	Klimakrise	anzugehen und in erneuerbare Energien zu ...	
7,259,503		... das so ist und die müssen uns beweisen , dass der	Klimawandel	nichts mit der Sonnenenergie und der ...	
15,073,047		... - die Universität in Niamey habe das Thema	Klimawandel	und erneuerbare Energien deshalb zu einem der ...	
16,890,110		... werden Im Kampf gegen die Energie- und die	Klimakrise	wollen die EU-Staaten den Ausbau erneuerbarer ...	

Verschwörungstheorien und Geschwurbel

- Zunehmende Desinformation und Unsachlichkeit u.a. durch soziale Medien
- Verschwörungstheorien, Geschwurbel, Populismus verzerren Meinungsbildung
- Ziel: linguistische Analyse & Stärkung der Medienkompetenz

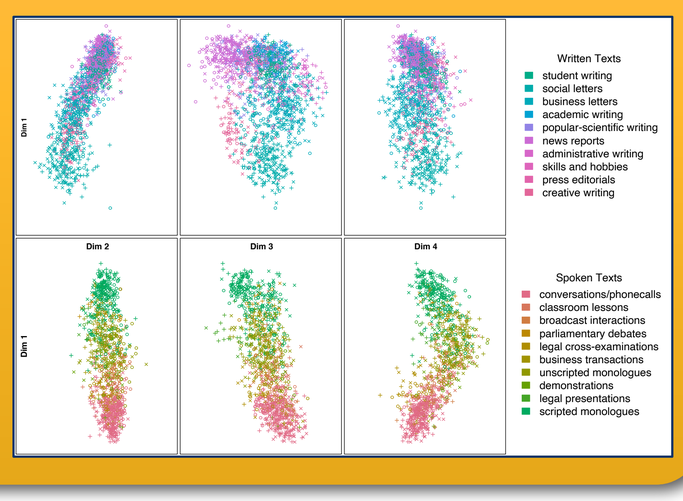
- Automatische Erkennung spezifischer Verschwörungsnarrative
- dient der Aufklärung, nicht der Zensur!
- benötigt transparente Modelle, die Muster zuverlässig erkennen
- Zero-shot-Ansatz mit KI auf Basis detaillierter Beschreibungen
- Few-shot-Ansatz mit LLMs auf Basis manuell annotierter Beispiele
- Untersuchung typischer sprachlicher Merkmale von „Schwurbeligkeit“



Korpuslinguistische Studien

- Sprachvariation: Geometric Multivariate Analysis (GMA)
- Kollokationen & Mehrwortausdrücke (MWE)
- Produktivität von Wortbildungsprozessen

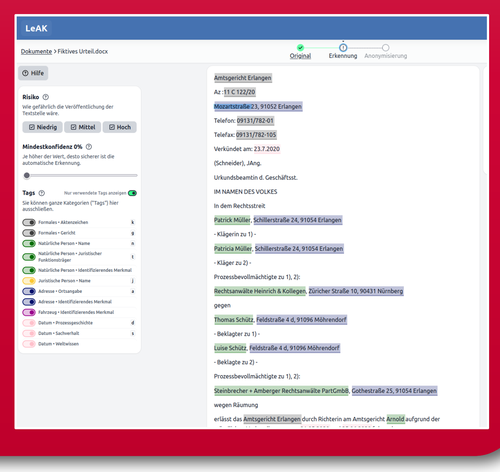
- Basis: Häufigkeiten von Wörtern & Wortklassen, lexikogrammatischen Konstruktionen, syntaktischen Mustern, Position im Text, …
- Auswertung: statistische Methoden, Visualisierung
- Hypothesentests, Assoziationsmaße, multivariate Statistik, Type-Token-Verteilung



Automatische Anonymisierung von Gerichtsurteilen (LeAK & AnGer)

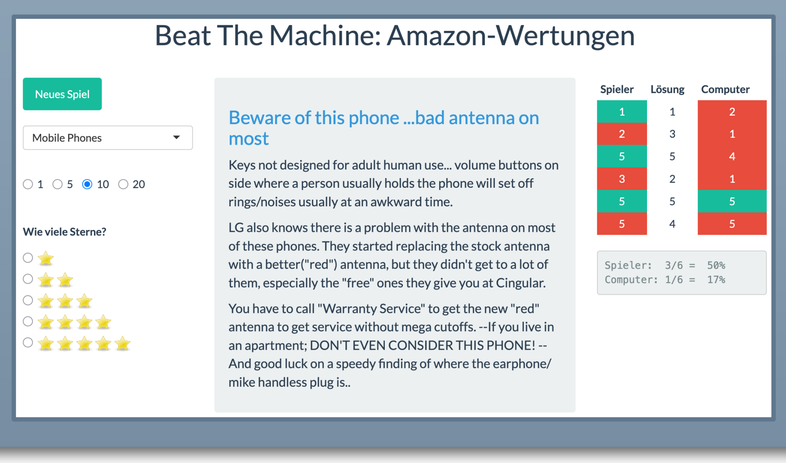
- Weniger als 3% aller Gerichtsurteile werden bisher veröffentlicht!
- Grund: manuelle Anonymisierung (teuer und fehleranfällig)

- Jeder direkte oder indirekte Bezug auf Personen muss erkannt werden (Kläger:innen, Zeug:innen, …)
- z.B. Vor- und Nachname, Adresse, Telefonnummer, Firmenname, …
- Forschungsfrage: Ist eine automatische Anonymsierung möglich?
- Antwort: Ja!
- mit speziell trainierten KI-Systemen (höchste Anforderungen nach KI-VO)
- Basis: hochwertiger Goldstandard



Beat the machine!

- Können Sie automatische maschinelle Lernverfahren schlagen?
- Aufgabe: Textklassifikation
- z.B. Sentiment, Produktwertung, Emotionen, Alter, Toxizität, Flausch, …
- Wir erklären gerne die computerlinguistischen Grundlagen!



Corpus Queries

zur Suche nach sprachlichen Mustern in großen Textsammlungen

Fragestellungen

- Wie wird Sprache verwendet? → Konkordanzanalyse
- Wie verändert sich Sprache? → diachrone Entwicklung
- Was ist „richtig“ und „falsch“? → Wörterbücher, Grammatiken
- Worüber wird gerade geredet? → z.B. Argumentation Mining

Korpora als Datengrundlage

- Korpus = (meist große) Datenbank maschinenlesbarer Texte
- angereichert mit linguistischen Informationen (Wortart, Lemma, Syntax, …)
- Arten von Korpora: Größe, Textsorte, Zeitperiode, Demographie, …

Werkzeuge

- Korpus-Plattformen: CQPweb, SketchEngine, KorAP, DWDS, Swiss-AL, …
- Suchanfragen (Corpus Queries): reguläre Ausdrücke, CQP/CQL, AQL, …
- Statistikprogramme und Visualisierung: R, Pandas, Matplotlib, Plotly, …