Textanalyse

Medienlinguistische Methodik

Arne Rubehn
arne.rubehn@uni-passau.de

02.11.2025

Zusammenfassung

In dieser Sitzung verschaffen wir uns einen Überblick über Methoden in der Analyse von Texten und Korpora, welche Methoden wofür geeignet sind und wie diese Methoden praktisch umzusetzen sind.

1 Einleitung

Es gibt sehr vielfältige Formen der computergestützten Datenanalyse, die man anwenden kann. Welche Methoden man verwendet, hängt dabei häufig nicht nur von der Fragestellung ab, sondern auch vom Umfang des Korpus, das man zugrunde legt. Im Folgenden schauen wir uns einige Formen an, die relativ einfach zu realisieren sind und sowohl bei großen, als auch bei kleineren Korpora sinnvoll angewendet werden können.

Korpora sind im weitesten Sinne Sammlungen verschiedener Texte. Es besteht der Anspruch, dass die einzelnen Texte untereinander vergleichbar sind, sowohl technisch (einheitliche Formatierung), als auch linguistisch (spezifische Forschungsfrage). Diese einzelnen Texte werden im Bezug auf Korpora auch häufig Dokumente genannt. Bei einem Korpus zu Rilkes Gedichten wäre also jedes Gedicht ein einzelnes Dokument.

Tokens sind die kleinste „Recheneinheit“ in einem Korpus (oder generell, in der maschinellen Sprachverarbeitung). In Korpusanalysen entsprechen Tokens i.d.R. einzelnen Wörtern. Im Folgenden wird immer wieder von „Wörtern“ die Rede sein, technisch gesehen sind hier aber Tokens gemeint. N-Gramme sind die Verbindung von N aufeinanderfolgenden Tokens. Meistens wird mit Bigrammen (N = 2) und Trigrammen (N = 3) gearbeitet, aber auch größere N-Gramme sind möglich (ab N = 4 spricht man meistens einfach von 4-Grammen, 5-Grammen, ...)

Im Folgenden verschaffen wir uns einen Überblick über verschiedene Methoden, mit deren Hilfe wir Korpora analysieren können. Die meisten dieser Methoden geben uns einen grundlegenden Überblick über Worthäufigkeiten und Wortzusammenhänge, die wiederum als Anhaltspunkte für eine genauere Betrachtung dienen können. Diese Herangehensweise lässt sich als „quantitativ informierte qualitative Analyse“ (Bubenhofer, 2013, 129) verstehen.

2 Wortwolken

Wortwolken (auch *tag cloud* genannt) sind ein sehr beliebtes Format zur Darstellung von Textinhalten, dessen Ursprung anscheinend im Unklaren liegt, aber schon recht weit zurückreicht. In Wortwolken werden Wörter nach Häufigkeit in unterschiedlicher Schriftgröße angeordnet und dabei in unterschiedlichen Farben so nebeneinander arrangiert, dass

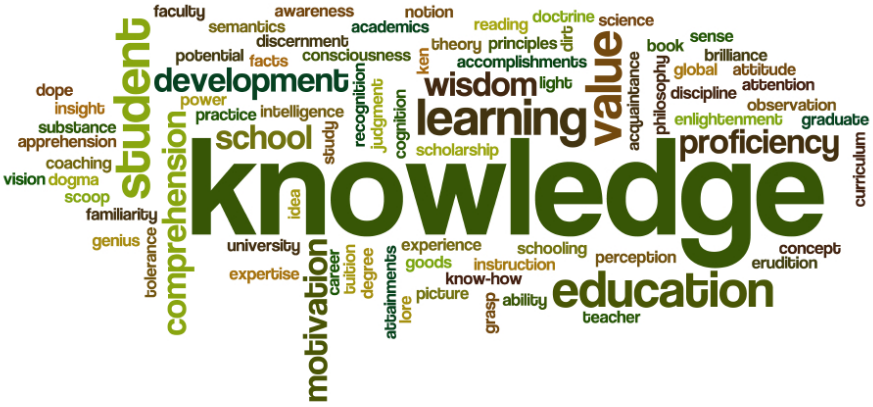


Abbildung 1: Beispiel einer Wortwolke.

man direkt sehen kann, wie häufig und wie selten welche Wörter in einer bestimmten Textsammlung auftauchen (Abb. 1). Auf diese Art kann man rasch sehen, welchen Wörtern eine wichtige Bedeutung in Texten zukommt und auch konkret verschiedene Wortwolken aus verschiedenen Texten miteinander vergleichen (vgl. die Beispielanalyse in Jackson et al., 2022).

Die Berechnung der Wortwolken beruht dabei auf Frequenzdaten, wobei für jeden Worttypen in einem Text die Anzahl an Worttokens berechnet und für deren Häufigkeit in Relation zu anderen Worttypen eine spezifische Schriftartengröße und mitunter auch eine spezifische Farbe berechnet, die diese Häufigkeit repräsentiert (oder randomisiert vergeben wird). Das Arrangement der Worttypen wird dann algorithmisch so festgelegt, dass die häufigsten Worttypen zentral positioniert werden.

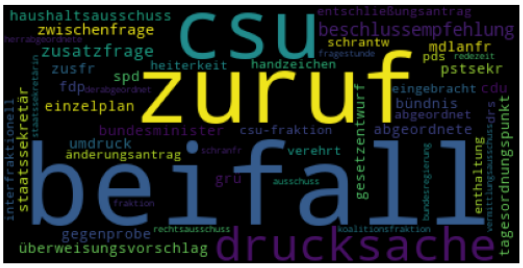
Die Frequenzdaten, die Wortwolken zugrundeliegen, können dabei absolut oder relativ zu einem Referenzkorpus sein – letzteres stellt dann dar, welche Wörter im Zielkorpus bedeutend häufiger vorkommen, als man es im Gesamtbestand der Sprache erwarten würde. Ferner können Wortwolken auch andere messbare Größen visualisieren, die Wörtern zugewiesen werden können.

3 Schlagwortanalyse

Als Schlagwörter (oder *Keywords*) werden solche Worte bezeichnet, die besonders prägnant für ein Dokument sind. Eine rechenbare Methode, um Schlagwörter aus einem Text zu extrahieren, ist der einfache Abgleich von Wortfrequenzen innerhalb des Dokuments mit Frequenzen in einem Referenzkorpus. Wörter, die in einem Dokument viel häufiger vorkommen, als allgemein zu erwarten wäre, sind demnach von besonderer Bedeutung (Abb. 2).

	Lemma
1	beifall
2	zuruf
3	csu
4	drucksache
5	beschlussempfehlung
6	zusatzfrage
7	staatssekretär
8	zwischenfrage
9	cdu
10	gegenprobe

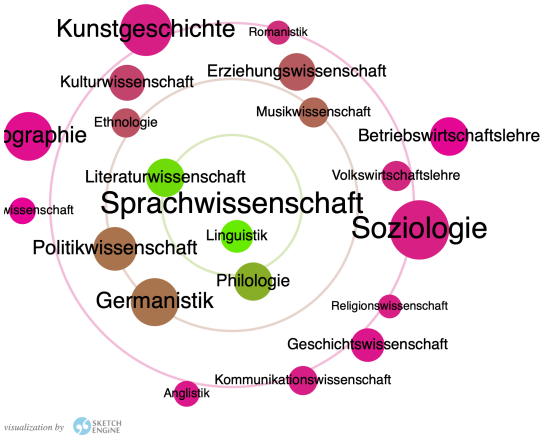
(a)



(b)

Abbildung 2: Darstellung von Schlagwörtern in Bundestagsdebatten als Liste (a) und Wortwolke (b).

4 Lexikographie



(a)

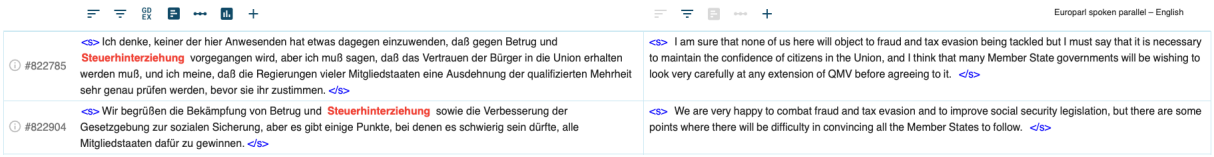
(b)

Abbildung 3: Die ähnlichsten Wörter zum Begriff „Sprachwissenschaft“ im deutschen Referenzkorpus deTenTen-23, dargestellt als Blasendiagramm (a) und Wortwolke (b).

Korpora eignen sich auch für lexikographische Analysen (insbesondere solche, die syntaktisch annotiert sind). Hierbei wird das typische Vorkommen eines Wortes in Bezug auf andere Worte geprüft: Frequenzen, Ähnlichkeiten, syntaktische Abhängigkeit. Hier können auch Modalitäten und Korpora miteinander verglichen werden: Korpusanalyses belegen z.B., dass wegen in der Schriftsprache vorwiegend mit Genitivobjekt, in der gesprochenen Sprache aber vor Allem mit Dativobjekt verwendet wird. Eine Visualisierung von Wortähnlichkeiten findet sich in Abb. 3.

Left context	KWIC	Right context
ıdes.</s><s>Ab jetzt sollten die politischen Agitatoren von	Klimakleber	über Attac bis Antifa mit Steuersubventionen – (Michael
.ebensschützern schleifen.</s><s>Es geht Ihnen nicht um	Klimakleber	, die Hunderttausende Autofahrer blockieren, oder um Z
i.</s><s>Jeder Kriminelle, jeder Linksautonome und jeder	Klimakleber	wird behaupten können, es gebe eine strukturelle Fehler
_and im Ehrenamt einbringen – (Kay Gottschalk [AfD]: Als	Klimakleber	mit Begleitschutz!) oder einen sehr guten Berufsabschlu
.fts- und Energiepolitik mit Sinn und Verstand ein, statt auf	Klimakleber	und sonstige Ökoterroristen zu hören!</s><s>(Beifall bei
›m 8. Oktober 2023, Seite 15, Titel: "Lufthansa fordert von	Klimaklebern	120.000 Euro Schadensersatz"), und, falls die beiden Bu
ch volle Straßen.</s><s>(Karsten Hilse [AfD]: Oder durch	Klimakleber	!) Statt unabhängig fühlen sich viele Menschen vom Aut
nergiewende.</s><s>Die wird ja logischerweise nicht von	Klimaklebern	umgesetzt, sondern von Handwerkern in unserem Land,
ssische CDU ist, deren Justizminister Roman Poseck die	Klimakleber	ernsthaft auf eine Stufe mit der RAF stellt –</s><s>Herr
· Arbeit quälen und dann noch, wenn sie Pech haben, von	Klimaklebern	aufgehalten werden.</s><s>Die, die auf maroden Bahnh
Anwälte der Nichtleister, Schützer der Gesetzesbrecher –	Klimakleber	und Fridays for Future lassen grüßen –, die Anwälte der
Till Steffen [BÜNDNIS 90/DIE GRÜNEN]) Hochkriminelle	Klimakleber	werden nicht verurteilt, auf der anderen Seite werden Ve
n!) – Das zeigen vor allen Dingen die Prozesse gegen die	Klimakleber	.</s><s>Immer öfter werden die weisungsgebundenen S
llweg oder Oliver Janich, (Leni Breymaier [SPD]: Sind das	Klimakleber	, oder was?) ohne dass am Ende irgendetwas von den T
ssing ist als Digitalminister fast unsichtbar.</s><s>Anstatt	Klimakleber	einzuladen, sollten Sie sich darauf konzentrieren, Deuts
ıektieren!</s><s>Das tun Sie wahrscheinlich nur bei Ihren	Klimaklebern	und Ihren Kraftwerksgegnern einer linken Klientel, die in

Abbildung 4: Beispiel einer Konkordanzlinie um das Lemma „Klimakleber“ in deutschen Bundestagsdebatten.



	Europarl spoken parallel – English
#822785	<s> Ich denke, keiner der hier Anwesenden hat etwas dagegen einzuwenden, daß gegen Betrug und Steuerhinterziehung vorgegangen wird, aber ich muß sagen, daß das Vertrauen der Bürger in die Union erhalten werden muß, und ich meine, daß die Regierungen vieler Mitgliedstaaten eine Ausdehnung der qualifizierten Mehrheit sehr genau prüfen werden, bevor sie ihr zustimmen. </s>
	<s> I am sure that none of us here will object to fraud and tax evasion being tackled but I must say that it is necessary to maintain the confidence of citizens in the Union, and I think that many Member State governments will be wishing to look very carefully at any extension of QMV before agreeing to it. </s>
#822904	<s> Wir begrüßen die Bekämpfung von Betrug und Steuerhinterziehung sowie die Verbesserung der Gesetzgebung zur sozialen Sicherung, aber es gibt einige Punkte, bei denen es schwierig sein dürfte, alle Mitgliedstaaten dafür zu gewinnen. </s>
	<s> We are very happy to combat fraud and tax evasion and to improve social security legislation, but there are some points where there will be difficulty in convincing all the Member States to follow. </s>

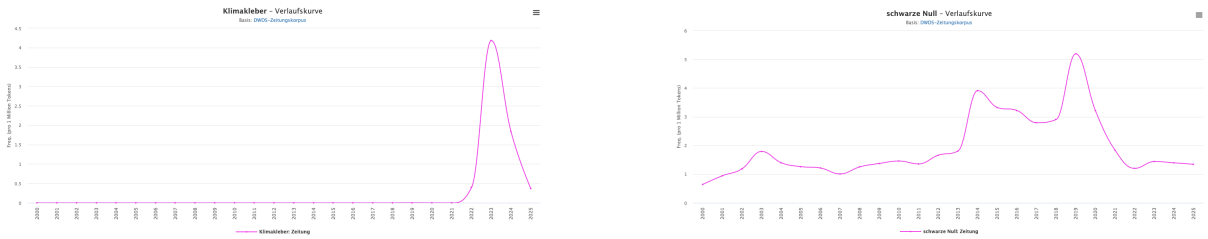
Abbildung 5: Beispiel einer parallelen Konkordanz aus mehrsprachigen Protokollen des Europaparlaments.

5 Konkordanzen

Konkordanzen sind ursprünglich „alphabetische Zusammenstellung von Worten […] und Stellen […] eines Werkes oder eines Schriftstellers, meist mit Angabe der Seiten […]” (Corsten, Füssel, Pflug & Schmidt-Künsemüller, 2017). In der Computerlinguistik versteht man darunter inzwischen aber auch eine visuelle Anordnung von Belegstellen, die jeweils auch den Kontext in dem ein Wort erscheint, visuell hervorheben, indem man das gesuchte Wort in die Mitte einer Matrix (Tabelle) setzt und die dem Wort vorangehenden und dem Wort folgenden Wörter links und rechts tabellarisch anordnet (Abb. 4). Diese Visualisierung wird auch Konkordanzlinie (*concordance line*, vgl. Hunston, 2022, 47) genannt.

Eine Sonderform der Konkordanzen sind parallele Konkordanzen, die sich aus alinierten parallelen Korpora ergeben. Hier liegt also der selbe Text in verschiedenen Sprachen vor, wobei annotiert ist, welche Sätze (oder teilweise sogar Wörter) einander direkt entsprechen 5. Parallele Konkordanzen bieten also direkte Belege für Übersetzungen. Solche parallelen Korpora existieren beispielsweise für Reden aus dem Europaparlament (Koehn, 2005) oder die Bibel (Mayer & Cysouw, 2014).

6 Wortverlaufskurven



(a) (b)

Abbildung 6: Beispiele für Wortverlaufskurven im DWDS anhand eines einzelnen Wortes (a) und eines Bigramms (b).

Mit der Hilfe von Wortverlaufskurven kann man versuchen, über einen Zeit- raum hinweg zu ermitteln, wie häufig bestimmte Wörter in einem Korpus verwendet worden sind. Wichtig ist dabei, dass die Texte im Korpus so an- notiert sind, dass man auf die Jahreszahlen zurückgreifen kann. Wichtig ist ferner auch, die Wortfrequenzen zu normalisieren, was normalerweise getan wird, indem man das Aufkommen eines Wortes pro Million Wörter in einem bestimmten Zeitraum misst. Die Zeitabschnitte müssen dabei auch festgelegt werden, wobei unterschiedliche Abschnitte auch zu unterschiedlichen Visualisierungen führen können.

Das Deutsche Textarchiv (BBAW 2024) erlaubt es, sehr schöne Zeitver- laufskurven aus den frei zur Verfügung gestellten Korpusdaten zu erzeu- gen. Noch mehr Möglichkeiten bietet das Digitale Wörterbuch der Deutschen Sprache (Geyken, 2025), da man hier auch N-Gramma visualisieren kann (siehe Abb. 6) und verschiedene Wortverlaufskurven direkt miteinander vergleichen kann. Ähnliche Funktionalitäten bietet auch der große und weit bekannte Google Books Ngram Corpus (Michel et al., 2011).

7 Kollokationen

In Kollokationsanalysen wird betrachtet, welche Wörter besonders häufig miteinander vor- kommen. Hierbei werden die Nachbarn eines Zielwortes gezählt, wobei ein Kontextfenster einer festen Größe zugrundegelegt wird. Berechnet wird, ob ein anderes Wort übermä- ßig häufig im Kontext des Zielwortes vorkommt (also Kookkurrenzhäufigkeit, normalisiert durch generelle Worthäufigkeit). Im Bezug auf ein Zielwort können somit Kollokations- profile erstellt werden. Aus solchen Profilen lässt sich zum Beispiel ableiten, dass das Wort „Steuer“ häufig mit den Verben „sparen“ oder „hinterziehen“ verwendet wird (Meier- Vieracker, 2025).

Literatur

Bubenhofer, N. (2013). Quantitativ informierte qualitative Diskursanalyse: Korpuslinguistische Zugänge zu Einzeltexten und Serien. In K. S. Roth & C. Spiegel (Hrsg.), *Angewandte diskurslinguistik*. Akademie Verlag. doi: https://doi.org/10.1524/9783050061054.109

Corsten, S., Füssel, S., Pflug, G. & Schmidt-Künsemüller, A. (Hrsg.). (2017). *Lexikon des gesamten Buchwesens Online*. Leiden: Brill.

Geyken, A. (2025). *Digitales Wörterbuch der deutschen Sprache DWDS. Das Wort- auskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. Zugriff auf `https://dwds.de` (aufgerufen am 01.12.2025)

Hunston, S. (2022). *Corpora in applied linguistics* (2. Aufl.). Cambridge: Cambridge University Press. doi: 10.1017/9781108616218

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R. & Lindquist, K. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17* (3), 805-826. doi: https://doi.org/10.1177/17456916211004899

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit*.

Mayer, T. & Cysouw, M. (2014, Mai). Creating a massively parallel Bible corpus. In N. Calzolari et al. (Hrsg.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (S. 3158–3163). Reykjavik, Iceland: European Language Resources Association (ELRA). Zugriff auf `https://aclanthology.org/L14-1215/`

Meier-Vieracker, S. (2025). *Wie erstellt man Kollokationsprofile?* Zugriff auf `https://www.tiktok.com/@fussballinguist/video/7555470919637126403` (aufgerufen am 01.12.2025)

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., ... others (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331* (6014), 176–182.