Medienlinguistische Methodik

Textanalyse

Arne Rubehn
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau

02.12.2025

Überblick

In dieser Sitzung verschaffen wir uns einen Überblick über Methoden in der Analyse von Texten und Korpora.

Welche Methoden gibt es?

Welche Methoden sind wofür geeignet?

Wie kann ich diese Methoden umsetzen?

Grundlagen

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.

Grundlagen

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, ...)

Grundlagen

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.

Wortwolken

Wortwolken sind eine sehr beliebte und einfache Methode zur **Visualisierung** von Worthäufigkeiten (oder anderen quantifizierbaren Eigenschaften von Wörtern).

Wortwolken können daher schnell und einfach darstellen, welchen Worten in einem Dokument eine besondere **Bedeutung** zukommt.

Als Grundlage können hierbei absolute Häufigkeiten dienen, aber auch **relative Häufigkeiten** (im Vergleich zu einem Referenzkorpus).

Wortwolken


knowledge
education
learning
development
student
comprehension
motivation
value
wisdom
proficiency
school
faculty
awareness
notion
semantics
academics
reading
doctrine
science
discernment
theory
principles
dirt
potential
facts
consciousness
accomplishments
book
sense
brilliance
global
attitude
attention
discipline
observation
enlightenment
graduate
curriculum
philosophy
acquaintance
dope
insight
substance
apprehension
coaching
vision
dogma
scoop
familiarity
genius
tolerance
power
practice
intelligence
recognition
judgment
cognition
study
idea
ken
light
scholarship
university
experience
schooling
perception
concept
erudition
goods
instruction
expertise
know-how
career
tuition
degree
attainments
lore
picture
grasp
ability
teacher

Schlagwortanalyse

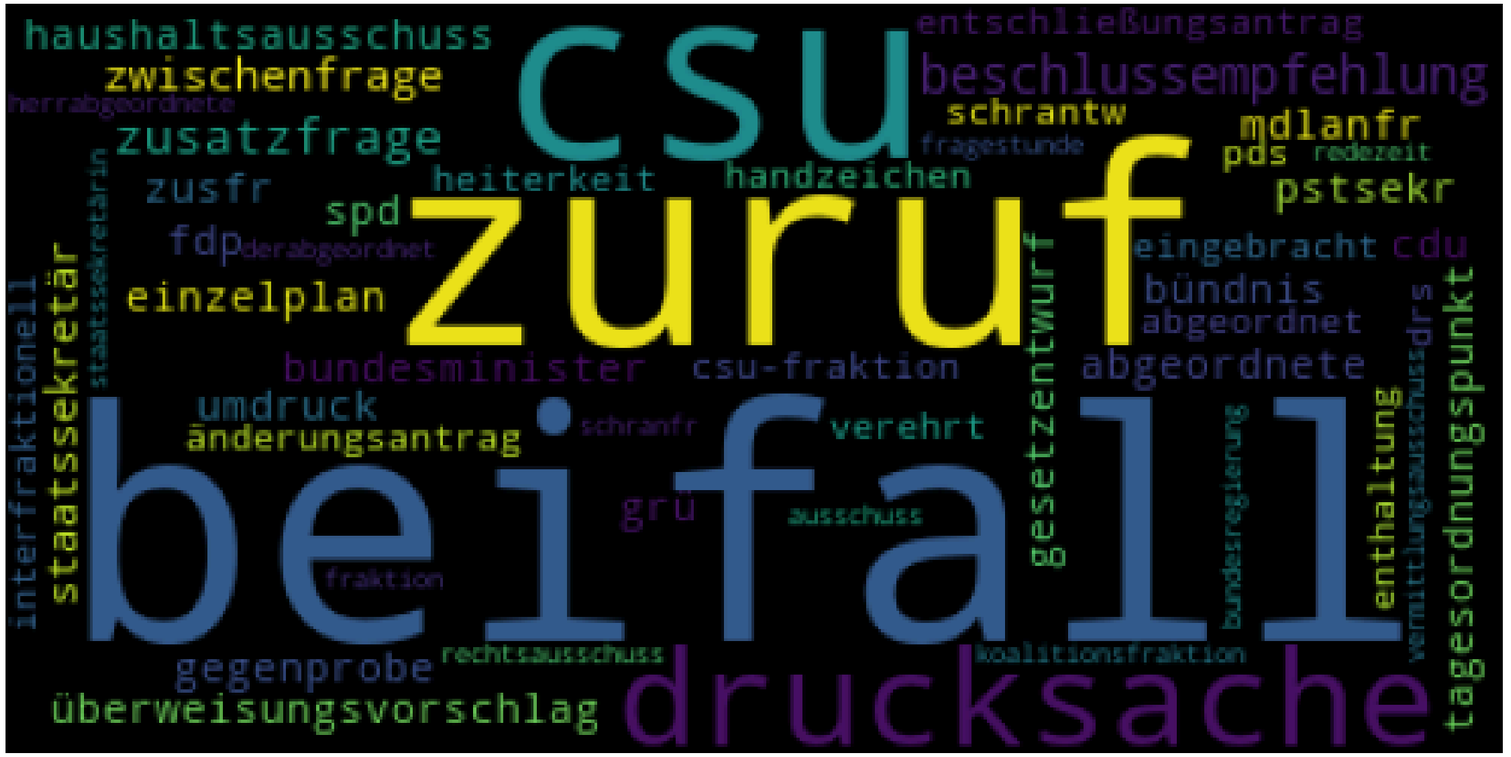
Als **Schlagwörter** (oder *Keywords*) werden solche Worte bezeichnet, die besonders prägnant für ein Dokument sind.

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.

Schlagwortanalyse

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



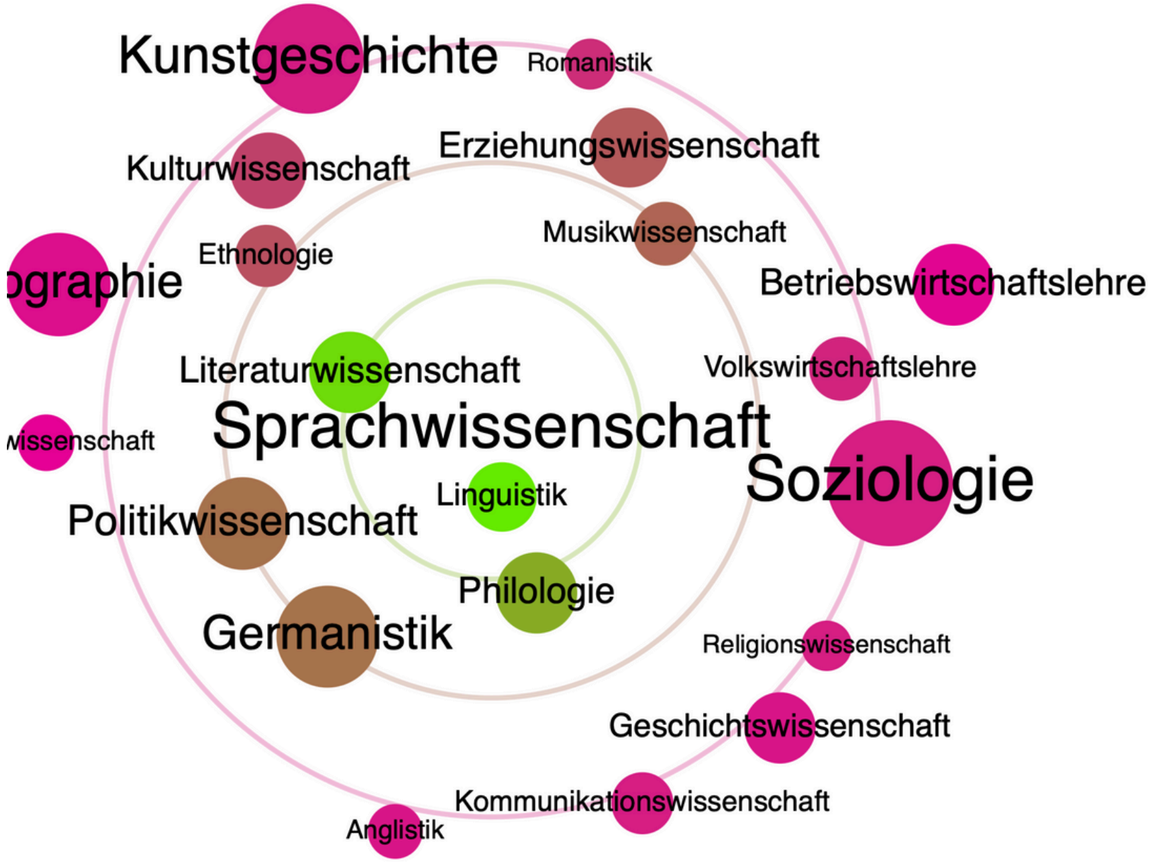
Lexikographie

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.

Lexikographie



visualization by SKETCH ENGINE

Lexikographie

Konkordanzen

Konkordanzen sind ursprünglich die “alphabetische Zusammenstellung von Worten […] und Stellen […] eines Werkes oder eines Schriftstellers, meistens mit Angabe der Seiten” (Corsten et al. 2017).

In der Korpuslinguistik bezeichnet der Begriff die Sammlung und visuelle Anordnung von **Belegstellen** eines bestimmten Wortes im Kontext.

Diese Form der Visualisierung wird auch **Konkordanzlinie** genannt.

Konkordanzen

Left context	KWIC	Right context
ıdes.</s><s>Ab jetzt sollten die politischen Agitatoren von	Klimaklebern	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
ɔm 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,
ɛssische 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

Parallele Konkordanzen

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.**

Parallele Konkordanzen bieten also direkte Belege für **Übersetzungen.**

Solche parallelen Korpora existieren beispielsweise für Reden aus dem Europaparlament oder die Bibel.

Parallele Konkordanzen

		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>

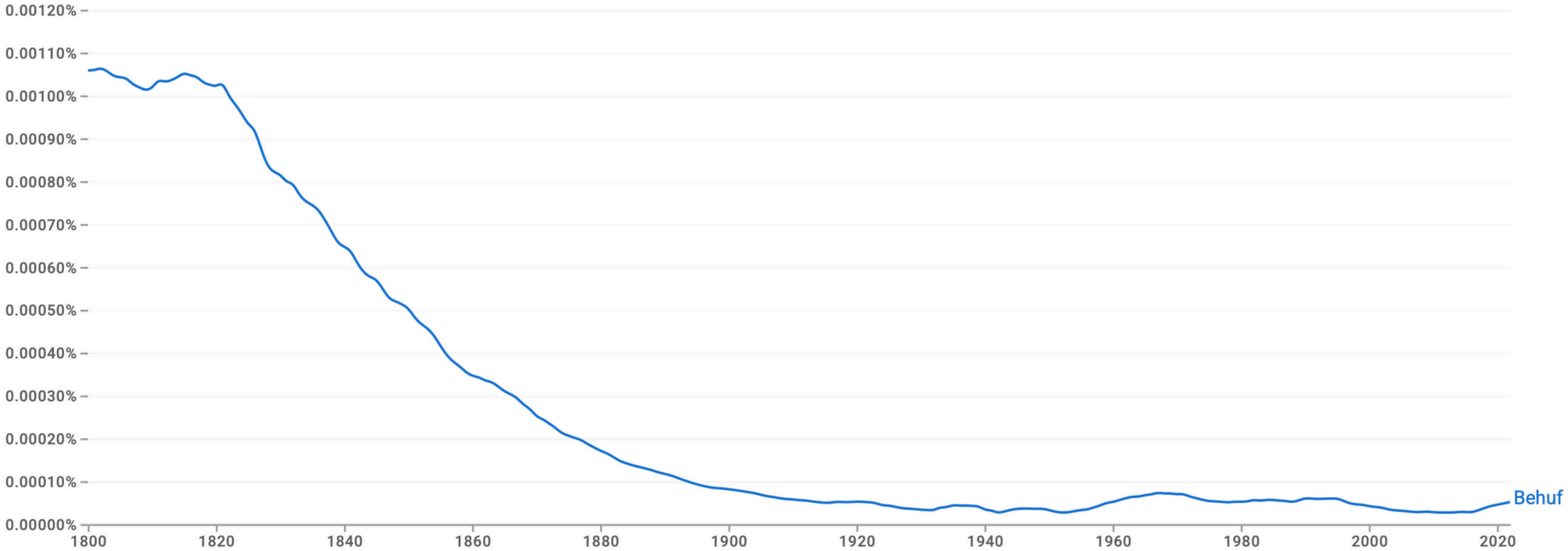
Wortverlaufskurven

Wortverlaufskurven stellen dar, wie häufig Worte **im Laufe der Zeit** verwendet werden (vgl. Google Ngram Viewer oder DWDS).

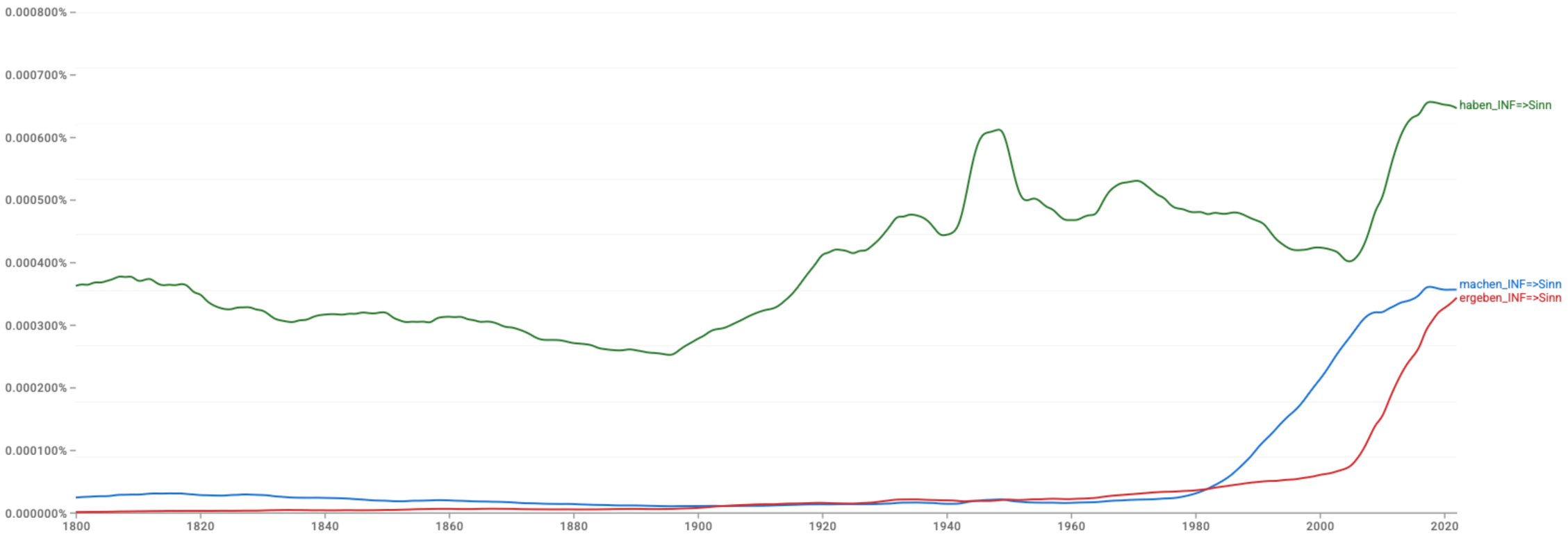
Für die Erstellung solcher Kurven sind **diachrone Korpora** notwendig, in denen Metainformationen zu den Jahreszahlen der einzelnen Dokumente vorliegen.

Wortfrequenzen müssen hierbei **normalisiert** werden, um aussagekräftig zu sein. Der einfachste Weg, das zu tun, ist pro Jahr (oder Dekade, …) die Frequenz des Wortes durch die Gesamtanzahl der Wörter zu teilen.

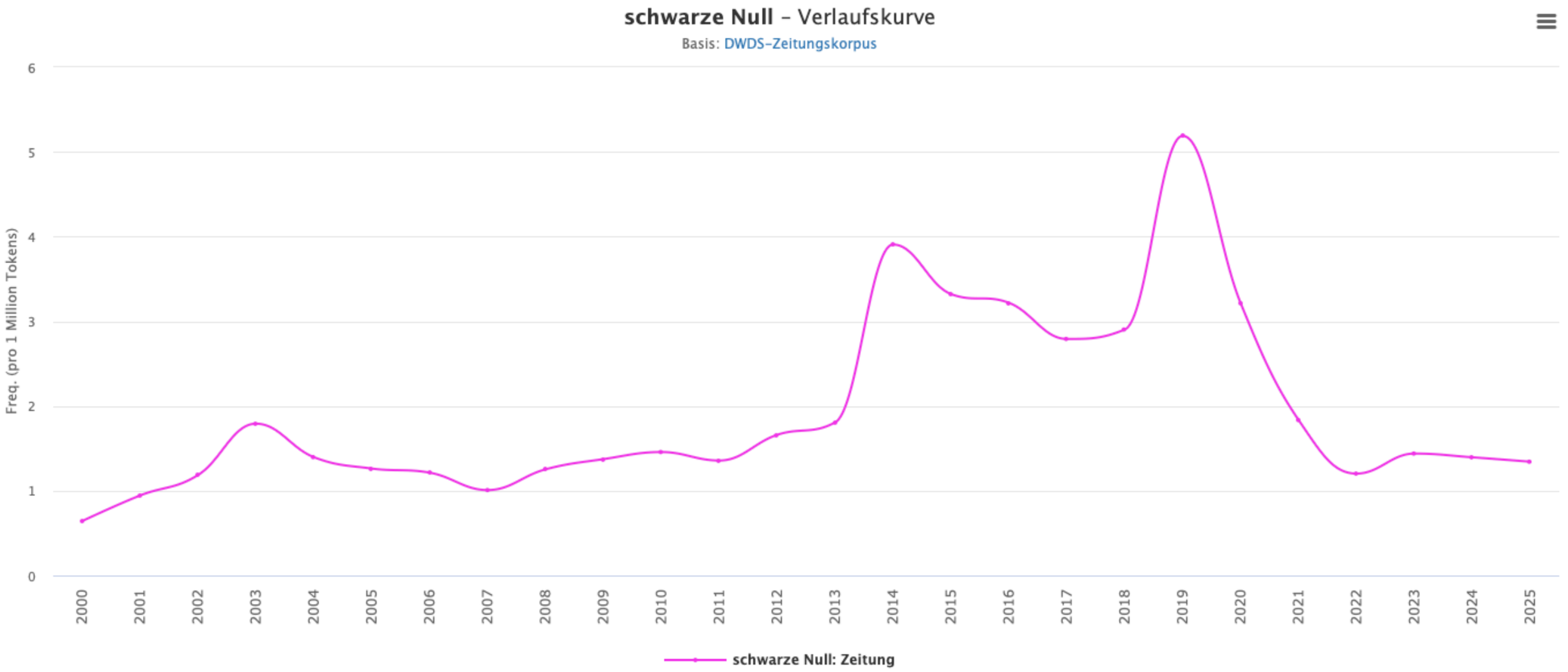
Wortverlaufskurven



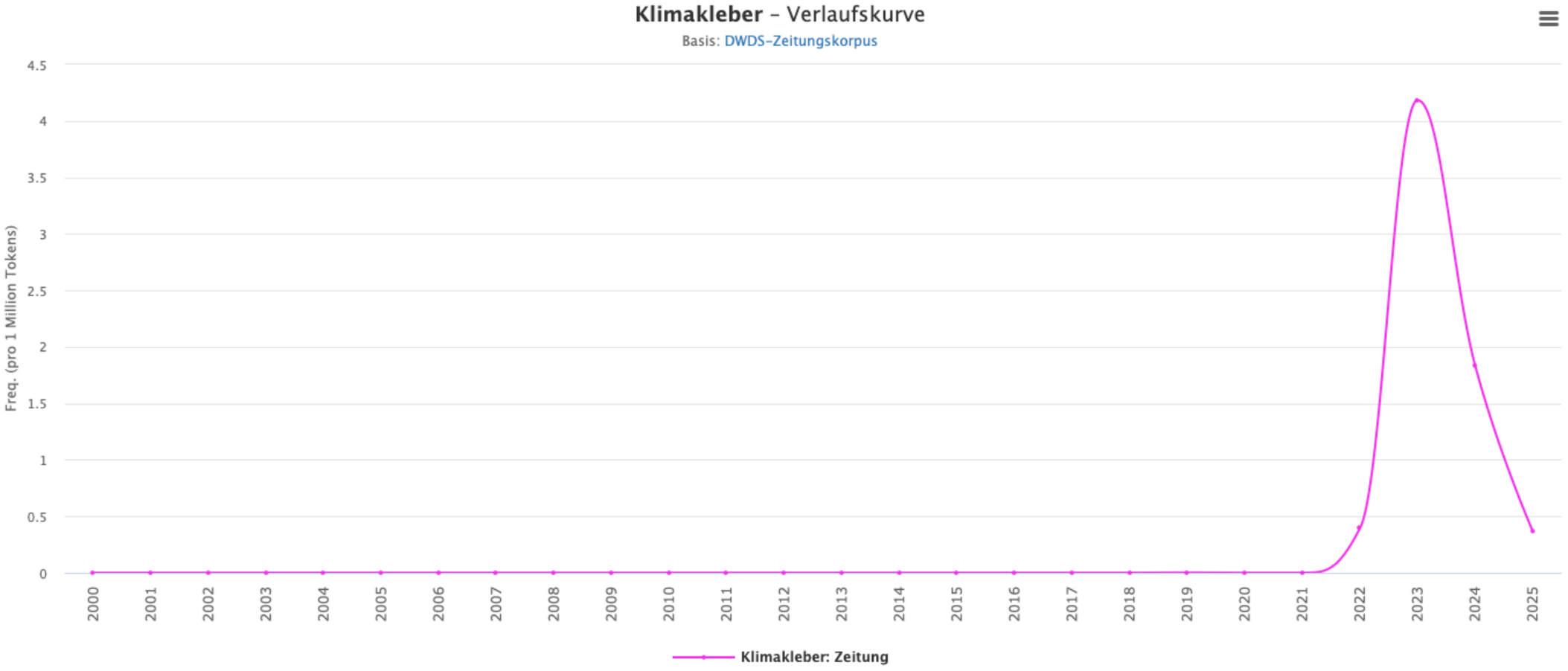
Wortverlaufskurven



Wortverlaufskurven



Wortverlaufskurven



Kollokationen

In **Kollokationsanalysen** wird betrachtet, welche Wörter besonders häufig miteinander vorkommen.

Hierbei werden die Nachbarn eines Zielwortes gezählt, wobei ein **Kontextfenster** einer festen Größe zugrundegelegt wird.

Berechnet wird, ob ein anderes Wort **übermäßig häufig** im Kontext des Zielwortes vorkommt (also Kookkurrenzhäufigkeit, normalisiert durch generelle Worthäufigkeit).

(https://www.tiktok.com/@fussballinguist/video/7555470919637126403)

Aufgabe

Erstelle dir einen Probe-Account auf SketchEngine unter https://sketchengine.eu.

Verschaffe dir einen Überblick über die verfügbaren Korpora und Analysen.

Suche dir ein Thema heraus, das du stichprobenartig analysierst. Fokussiere dich dabei auf einen speziellen Korpus (z.B. Bundestagsreden); behalte den Referenzkorpus (deTenTen23) im Hinterkopf.

Schreibe eine kleine Analyse (max. eine halbe Seite) zu deinen Beobachtungen und lade sie im Ordner “Studienleistungen” auf StudIP hoch.