Medienlinguistische Methodik

Korpuslinguistik

Arne Rubehn
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau

11.11.2025

Überblick

In dieser Sitzung widmen wir uns den Grundlagen der **Korpuslinguistik.**

Was ist ein Korpus?

Wie werden Korpora generiert und aufgebaut?

Wie können Korpora in der Sprachforschung verwendet werden?

Empirismus und Rationalismus

Zwei gegenüberstehende **Perspektiven** der Wissenschaft.

Rationalismus. Erkenntnisse werden geistig durch Begriffe und Urteile gewonnen.
Untersuchung des Sprachsystems und der Sprachkompetenz. (*I-Sprache, langue)*

Empirismus. Erkenntnisse werden durch Erfahrung gewonnen.
Untersuchung des Sprachgebrauchs und der Sprachperformanz. (*E-Sprache, parole)*

Wer hat Recht? Was hat das mit Korpuslinguistik zu tun?

Empirismus und Rationalismus

Ferdinand de Saussure unterscheidet zwischen ***langue*** und ***parole***.

langue. abstraktes **Sprachsystem**, einheitliche Regeln, unabhängig von einzelnen Sprecher:innen

parole. konkreter **Sprachgebrauch**, variabel, kontrolliert durch einzelne Sprecher:innen

Noam Chomskys ***I-Sprache*** und ***E-Sprache*** entsprechen im Kern ungefähr ***langue*** und ***parole.***

Empirismus und Rationalismus

Beide Ansätze sind in ihrer “Reinform” unfruchtbar.

Zur Herausbildung abstrakter, **rationalistischer Konzepte** sind zumindest einmal grundsätzliche Beobachtungen notwendig.

Empiristische Beobachtungen alleine generieren ohne Abstraktion kein Wissen über zugrundeliegende Objekte und Prozesse.

Beide Ansätze sind nicht exklusiv, sondern komplementär zu verstehen.

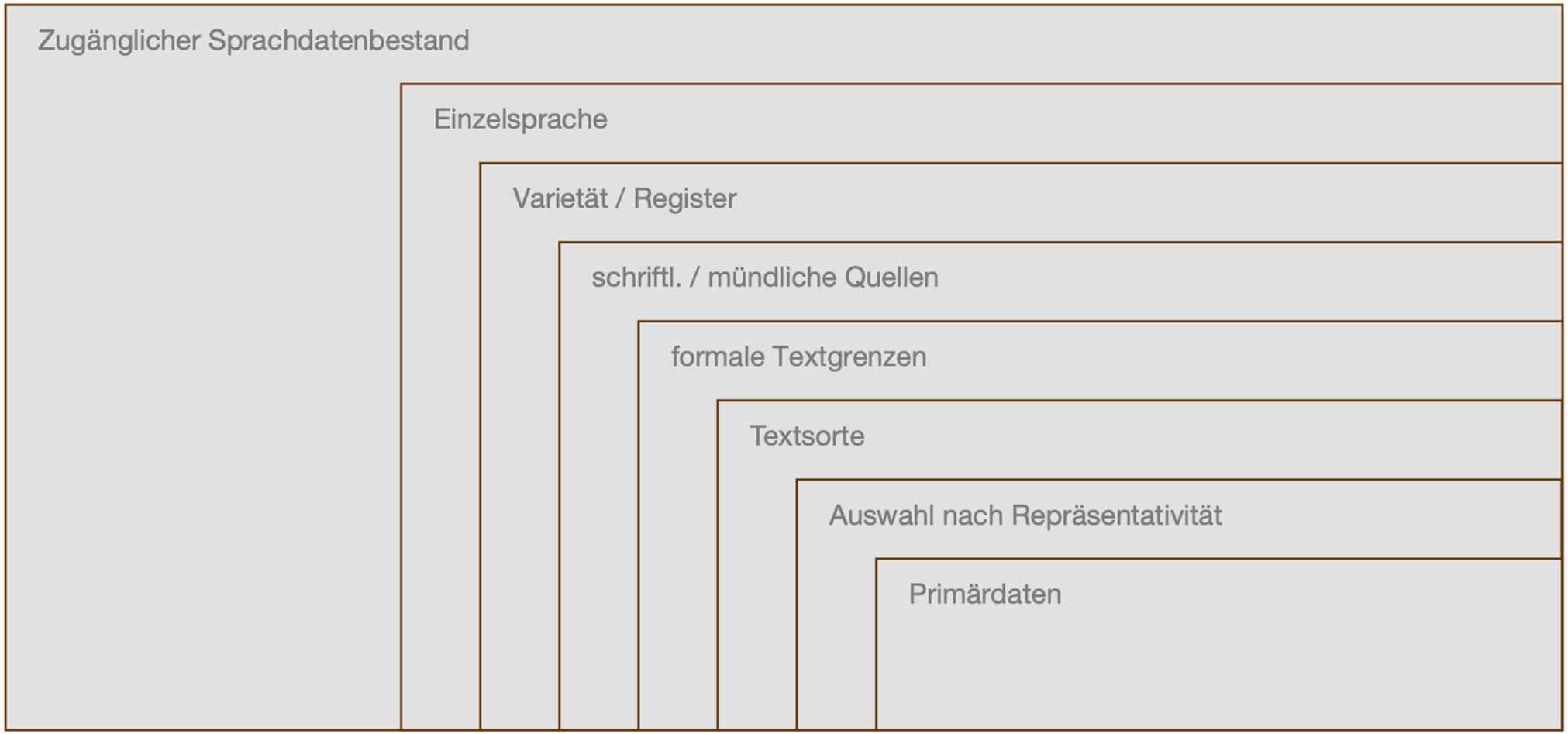
Korpus

Ein **Korpus** (von lat. *corpus* ‘Körper’, Pl. *Korpora*) bezeichnet in der Linguistik eine Sammlung an Texten, die zu einem gewissen Zweck zusammengetragen wurde.

Der Begriff leitet sich metaphorisch vom Körper als **Hauptbestandteil** ab und wurde in der Klassik für ein mehrteiliges Werk (Cicero), später auch für die wesentlichen Teile einer größeren Sammlung (Vitruvius) verwendet.

Korpora haben den Anspruch, möglichst **repräsentativ** und **authentisch** zu sein.

Korpus


Zugänglicher Sprachdatenbestand
Einzelsprache
Varietät / Register
schriftl. / mündliche Quellen
formale Textgrenzen
Textsorte
Auswahl nach Repräsentativität
Primärdaten

Anforderungen an Korpora

Digitalität. Korpora sollten in maschinenlesbarer Form vorliegen.

Authentizität. Die Daten behalten den originalen Kontext, was die Grundlage jeder qualitativen Auswertung ist. Einfache Wortlisten sind also keine Korpora!

Zweck. Ein Korpus ist kriterienbasiert und für einen bestimmten Forschungszweck zusammengestellt.

Konstanz. Korpora bleiben i.d.R. konstant, ändern sich also nicht über die Zeit.

Anforderungen an Korpora

Synchronie. Korpora sind i.d.R. synchron (allerdings existieren auch dezidiert diachrone Korpora)

Repräsentativität. Der Korpus bildet den Gesamtdatenbestand gut ab.

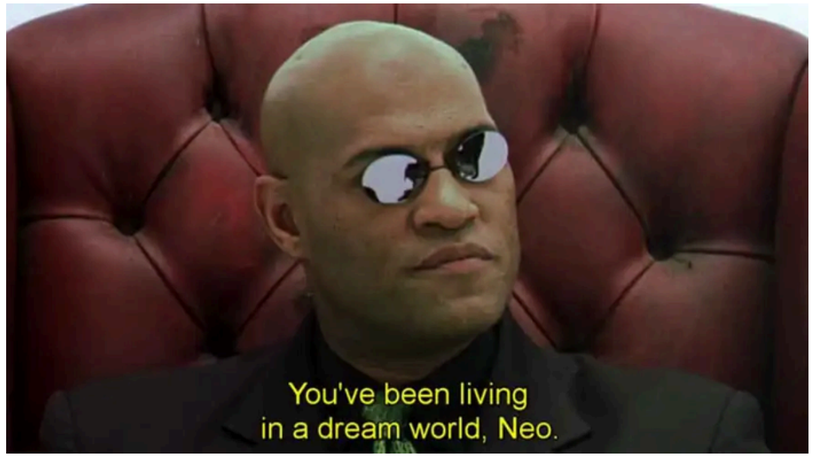
Referentialität. Eine Sonderform von Korpora ist der **Referenzkorpus,** der den *gesamten* Bestand der Sprache abdecken soll. (durchaus umstritten!)

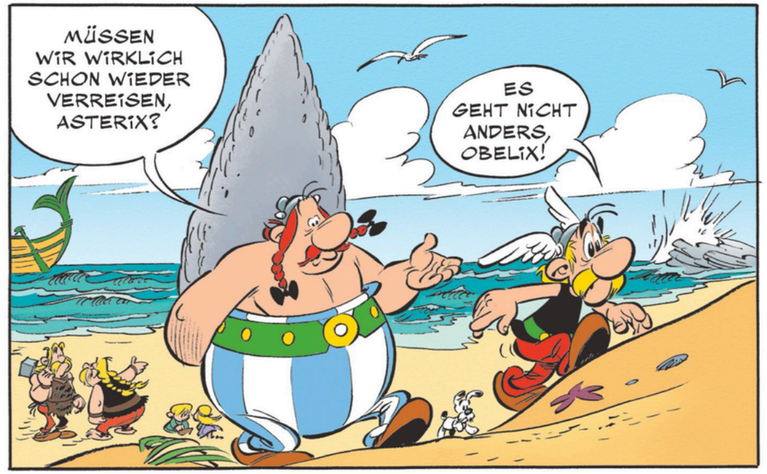
Mindestanforderungen: Authentizität und Repräsentativität!

Multimodale Korpora

Insbesondere im Bereich der Medienlinguistik spielen **multimodale Korpora** eine Rolle, die üblicherweise Text- und Bilddaten kombinieren.

Diese Multimodalität findet sich auch bei Videoaufnahmen aus **Beobachtungen** oder bei **Gebärdensprachen** wieder.





Daten in Korpora

Wenn wir in Korpora “Wörter zählen”, unterscheiden wir zwischen **Token, Types** und **Lemmata.**

Token. Kleinste Einheit des Korpus, jede Instanz zählt. Satzzeichen werden meistens als eigenständige Tokens mitgezählt.

Type. Die Menge *unterschiedlicher* Tokens (also, jedes Token wird nur einmal gezählt).

Lemma. Die lexikalische Grundform (“Wörterbuchform”) eines Tokens.

Daten in Korpora

Wenn Robben hinter Robben robben, robbten Robben hinter Robben.

Wie viele Tokens?

Wie viele Types?

Wie viele Lemmata?

Daten in Korpora

1 2 3 4 5 6 7 8 9 10 11

Wenn Robben hinter Robben robben , robbten Robben hinter Robben .

Wie viele Tokens? 11

Wie viele Types?

Wie viele Lemmata?

Daten in Korpora

1 2 3 2 4 5 6 2 3 2 7

Wenn Robben hinter Robben robben , robbten Robben hinter Robben .

Wie viele Tokens? 11

Wie viele Types? 7

Wie viele Lemmata?

Daten in Korpora

1 2 3 2 2 4 5 2 3 2 6

Wenn Robben hinter Robben robben , robbten Robben hinter Robben .

Wie viele Tokens? 11

Wie viele Types? 7 (6 wenn wir Groß-/Kleinschreibung ignorieren)

Wie viele Lemmata?

Daten in Korpora

1	2	3	2	4	-	4	2	3	2	-
Wenn	*Robben*	*hinter*	*Robben*	*robben*	*,*	*robbten*	*Robben*	*hinter*	*Robben*	*.*

Wie viele Tokens? 11

Wie viele Types? 7 (6 wenn wir Groß-/Kleinschreibung ignorieren)

Wie viele Lemmata? 4

Daten in Korpora

Jede dieser Einheiten kommt mit ihren eigenen, kleinen **Herausforderungen.**

Tokenisierung. *New York* vs. *new shoe; I will* vs. *I'll; hand luggage* vs. *handbook*

Typisierung. Homonyme und Polyseme fallen unter einen Type – unterschiedliche Formen eines Wortes allerdings nicht.

Lemmatisierung. Keine einfache automatische Lösung – erfordert entweder manuelle Annotation oder komplexe (nicht komplett verlässliche) Computermodelle.

Korpusgenerierung

Korpora werden üblicherweise aus bereits **bestehenden Quellen** “zusammengefügt”.

Je nach Zweck des Korpus sollten unterschiedliche Quellen sorgfältig **gewichtet** werden (*Repräsentativität)*, wodurch u.U. **Kürzungen** oder **Auszüge** der Primärdaten erforderlich sind.

Eine “Sonderform” stellen hierbei Daten aus dem **Internet**, insbesondere aus Social Media dar. Auch Sprachdaten, die im Zuge von **Befragungen** oder **Experimenten** erhoben wurden, können als Korpora dienen.

Normalisierung

Um sinnvoll mit Korpora zu arbeiten, ist häufig eine **Normalisierung** notwendig.

Dies beinhaltet typischerweise die Zusammenführung von **Varianten** (Lemmatisierung, variierende Normen) und die **Korrektur** von Fehlschreibungen.

Auch in der Anwendung gibt es typische Normalisierungsschritte, wie das Ignorieren von **Groß- und Kleinschreibung** oder von häufigen **Funktionswörtern.**

Eine korrekt ausgeführte Normalisierung ändert die Primärdaten nicht, sondern ist eine Annotation!

Normalisierung: Problemfälle

~~»Lieber Magnus,~~

~~ich habe zusammengezählt, wie oft wir uns im Leben gesehen habe: [redacted]. Aber das bedeutet nichts, wie alle Welt sehr gut weiß. Petrarca hat Laura wie oft gesehen? Nun bin ich …«~~

»Wir haben Frl. Natalie von Wilde als eine sehr sympathische, sehr gebildete Dame kennen gelernt, sie hat ein prächtiges Organ, und man vergißt manchmal, auf den Sinn ihrer Worte zu hören, weil man schöner Musik zu lauschen vermeint.«

Und ich stamme aus einer ~~Künstler- und Intellektuellen-~~ Bauern- und Polizistenfamilie aus der DDR.

Damit er sah, dass ich auch nicht niemand war.

Ich hab ihn gefragt, ob er keine Angst hat, mit wem ich mich herumtreibe, wenn ich ohne ihn in Berlin bin. Er hat gesagt, *weg'n wos sullt'* er Angst haben, seiner Erfahrung nach bin ich *net so ane*, aber sollte ich doch so eine sein, *wenn's dees is' wos t wüllst*, dann sollte ich lieber gleich dableiben, *für wos sullnma t Fülzleis* und des Drama durchs *holwarte Eyropa schloaffa*. Was nicht von »schlafen«, sondern von »schleppen« kommt, wie man Leuten aus dem *Dieringischn* erklären musste, aber was für ein schöne, passende Fehlleistung das doch war!

Dinge hineinsteigern.

!?!? Ich sollte was nicht?

Er sagte nicht, dass ich die Einzige wä

Anwendung

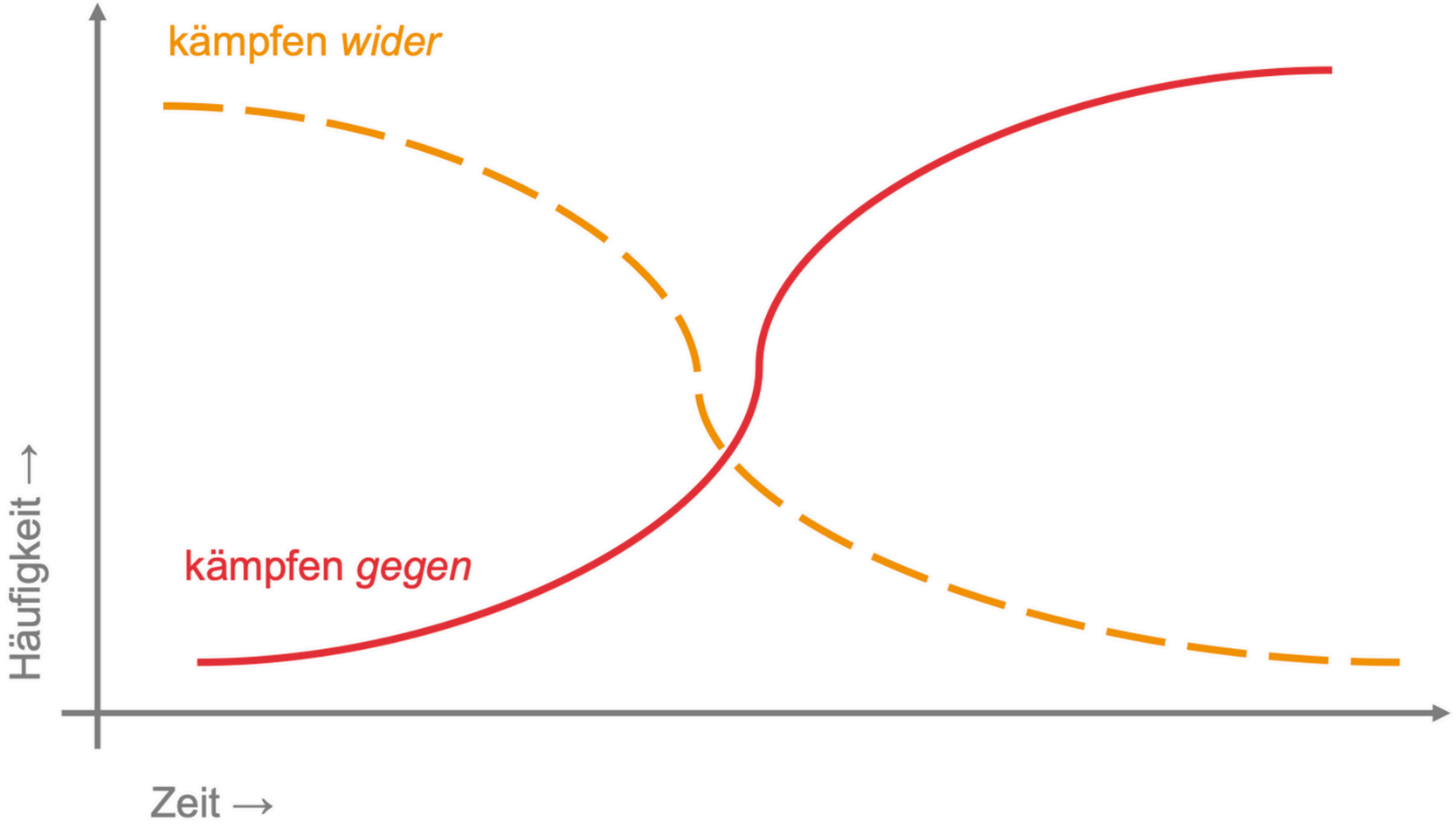
Korpora können im weitesten Sinne sowohl zur **Theoriebildung** als auch zur **Theorieprüfung** dienen.

Hierbei kann eine weite Reihe an Phänomenen untersucht werden: Argumentationsstrukturen, Meinungen oder Framing zu bestimmten Themen, Grammatikalität von Satzstrukturen, Bedeutungswandel, lexikographische Analysen, …

Korpora eignen sich also für die Untersuchung von **Sprachsystem** und **Sprachgebrauch.**

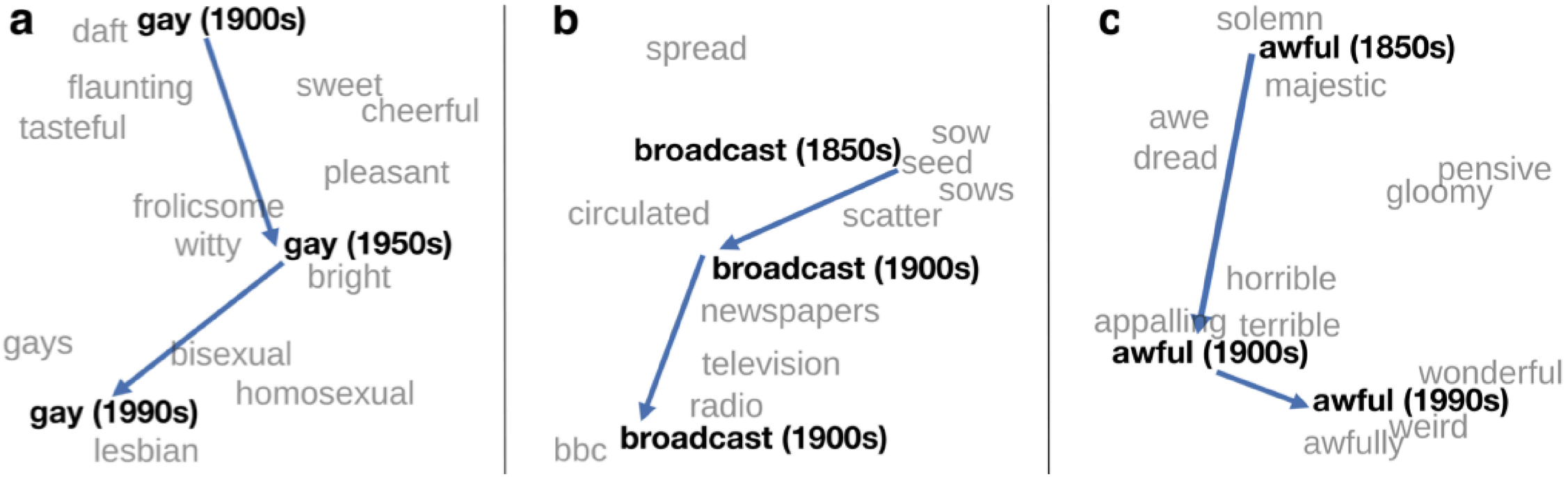
Anwendung

Lexikographie, Sprachwandel



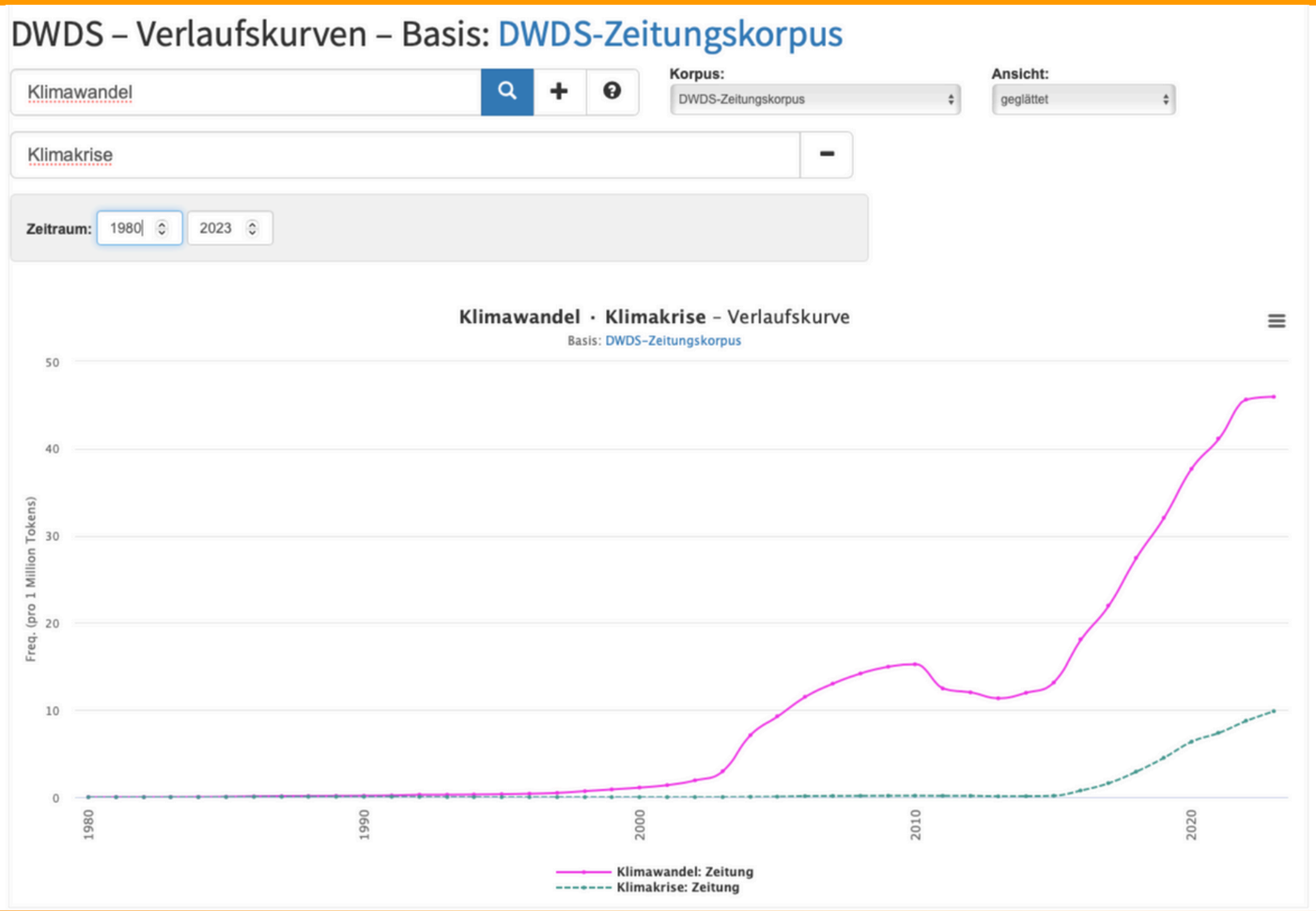
Anwendung

Sprachwandel, Wortfelder



Anwendung

Lexikographie, Diskursanalyse



Korpora: Übersicht

Eine Sammlung einiger relevanten Korpora findet sich unter:
https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/links/korpora_links

Verschaffe dir einen Überblick.

Welche Korpora sind für welche Fragestellungen geeignet?

Welche Probleme gibt es eventuell?