Korpuslinguistik

Medienlinguistische Methodik

Arne Rubehn
arne.rubehn@uni-passau.de

11.11.2025

Zusammenfassung

Diese Sitzung bietet eine Einführung in die Korpuslinguistik. Hierbei soll zunächst einmal ergründet werden, was wir unter einem Korpus verstehen. Dann verschaffen wir uns einen Überblick darüber, wie Korpora generiert und aufgebaut werden, und zuletzt beleuchten wir einige Anwendungsfälle für Korpora in der Sprachforschung.

1 Einführung

In vielen Bereichen der Wissenschaft stehen sich Vertreter des **Empirismus** und des **Rationalismus** gegenüber. Obwohl heute empirische Forschung einen unbestreitbaren Vorrang hat, bedeutet dies nicht, dass nur der Empirismus eine wissenschaftlich vertretbare Haltung ist. Was genau steht hinter den Begriffen?

- **Rationalismus:** Erkenntnisse werden geistig durch Begriffe und Urteile gewonnen.
- **Empirismus:** Erkenntnisse werden durch Erfahrung gewonnen.

Wenn es hier ein Problem gibt, dann liegt es vermutlich im Anspruch auf ausschließliche Geltung. Dass es mit der Entscheidung für eine Seite nicht getan sein kann, liegt auf der Hand, denn es ist nicht möglich, empirische Forschung zu betreiben, die von Theorien unabhängig ist. Empirisches Vorgehen bestätigt oder widerlegt Hypothesen, die zuvor von Theorien abgeleitet werden müssen. Nun ist es aber nicht möglich, Theorien selbst als direktes Ergebnis aus Erfahrungen zu erhalten. Es handelt sich bei ihnen vielmehr um Erklärungsmodelle, die uns helfen, Erfahrungen in einen kohärenten Zusammenhang einzuordnen und im Idealfall auch Voraussagen zu treffen.

Obwohl letztlich also beide Ansätze gebraucht werden, führen sie doch zu verschiedenen Vorge- hensweisen in der Linguistik. Rationalistische Ansätze argumentieren **kompetenzbasiert** und beschreiben das **Sprachsystem**, also ein abstraktes System, das unabhängig von einzelnen Sprecher:innen einheitlichen Regeln folgt (vgl. Saussures *langue* oder Chomskys *I-Sprache*). Empiristische Ansätze hingegen argumentieren **performanzbasiert** und untersuchen den konkreten **Sprachgebrauch**, der viel variabler ist und durch einzelne Sprecher:innen kontrolliert wird (vgl. Saussures *parole* oder Chomskys *E-Sprache*).

Der Beginn der Sprachwissenschaft am Anfang des 20. Jhdts. war in diesem Sinne weitgehend rationalistisch, und empirische Überprüfung beschränkte sich im Wesentlichen auf selbstgewählte und meist konstruierte Beispiele. Dieses Vorgehen ist bis heute noch in manchen Bereichen Standard und zieht zu Recht Kritik auf sich, weil es rein subjektiv und wenig nachprüfbar erscheint. Die Stützung auf authentische Daten aus Textkorpora stellt

in diesem Punkt sicher eine Verbesserung dar. Andererseits sind rationalistische Ansätze und die damit verbundenen Abstraktionen auch nicht ohne Weiteres von der Hand zu weisen: Ob nun regelbasiert oder nicht, eine Sprache, die so variantenreich ist, dass jeder mittelkomplexe Satz wahrscheinlich nur ein Mal in der gesamten Menschheitsgeschichte formuliert wurde, kann nicht allein anhand von Performanzdaten beschrieben werden. Selbst ein extrem großer Korpus würde die sprachlichen Möglichkeiten niemals auch nur annähernd ausschöpfen.

Beide Ansätze sind also in ihrer „Reinform“ unfruchtbar. Zur Herausbildung abstrakter, rationalistischer Konzepte sind zumindest einmal grundsätzliche Beobachtungen notwendig; empiristische Beobachtungen alleine generieren allerdings ohne Abstraktion kein Wissen über zugrundeliegende Objekte und Prozesse. Beide Ansätze sind also nicht exklusiv, sondern komplementär zu verstehen.

2 Korpora

2.1 Was ist ein Korpus?

Der Terminus Korpus kommt von lat. *corpus* (Neutrum, Gen. *corporis*, Nom. Pl. *corpora*), verbreitet vor allem durch die juristische Fachsprache: das Corpus Delicti. Daher also die Pluralform Korpora, allerdings wird teilweise auch von Korpussen gesprochen. Die Bedeutung ist wörtlich Körper, seit Cicero auch mehrbändiges Textwerk, bei Vitruvius erweitert zu eine Sammlung von Texten, die das Wesentliche eines Gegenstandes darstellt. Das ist in etwa auch, was wir in der Linguistik darunter verstehen: ein Kompendium von Texten, das einen bestimmten sprachlichen Bereich (z.B. Ostbairisch oder Jugendsprache) einigermaßen vollständig und (falls möglich) repräsentativ, abbildet.

Linguistische Korpora, die nicht aus verschriftlichten Texten bestehen, sind ebenfalls möglich, z.B. aus Aufzeichnungen gesprochener Sprache, oder auch aus ganz anderen Objekten, wie z.B. Bild- Texten (Memes etwa oder Nachrichten mit Emojis).

Wie schon erwähnt, ist ein Korpus immer ziemlich weit vom Gesamtbestand verfügbarer Sprachda- ten entfernt, denn eine sinnvolle Reduktion der Daten ist ja meistens die Absicht. Mögliche Stufen der Auswahl sind z.B. Einzelsprache → Varietät / Register → visuelle / auditive Daten → formale Textgrenzen → Textsorte → Auswahl nach Repräsentativität → Primärdaten. Das ist nur ein Bei- spiel, es sind durchaus noch mehr und andere Stufen der Reduktion möglich. Hier geht es nur darum aufzuzeigen, wie weit sich ein Korpus von „der Sprache“ entfernt.

Eine gewisse Ausnahme stellen vielleicht die Trainingsdaten von *Large Language Models* dar, da hier das Streben gerade in Richtung Maximierung der Datenmenge geht. Aller-

dings sind auch solche Korpora auf verfügbare (verschriftlichte, rechtlich zugängliche[1], kommerziell verwertbare) Daten be- schränkt und bislang ist immer noch die größte Menge erzeugter Sprachdaten volatil, weil sie aus gesprochener Sprache besteht und nirgends aufgezeichnet wird. Fast alle wissenschaftliche Arbeit wird daher von geschriebener Sprache dominiert, obwohl diese sich z.T. erheblich von gesprochener Sprache unterscheidet.

2.2 Anforderungen an Korpora

Es gibt praktische und theoretische Anforderungen an Korpora. Praktische beziehen sich auf die **Verarbeitbarkeit** und in erster Linie werden **digitale Korpora** für maschinengestützte Auswertung erzeugt. Nichtdigitale Korpora in diesem Sinne, z.B. aus als digitale Bilder vorliegenden Archivalien von Handschriften, sind aber durchaus in Gebrauch, denn oft erscheint der Aufwand für eine maschinenlesbare Digitalisierung zu hoch. Prinzipiell lassen sich die gleichen Verfahren auch auf solche oder komplett analoge Korpora anwenden, aber der Anteil manueller Arbeit ist weitaus höher. Die wichtigste theoretische Anforderung ist **Authentiziät**. Was bedeutet das? Korpora enthalten nicht nur Daten, sondern auch Relationen zwischen diesen Daten (z.B. syntaktische Informationen, Kollokationen etc.). Von Authentizität wird gesprochen, wenn der ursprüngliche Textzusammenhang bzw. Kontext im Korpus erhalten bleibt (Hirschmann, 2019, 3). Vorsicht ist also bei Standardisierung (Kürzungen) und Normalisierung (Schreibvereinheitlichung) geboten. Hieraus folgt auch, dass reine Wortlisten keine Korpora in diesem Sinne sein können, denn hier ist der Textzusammenhang zerstört.

Daneben gibt es weitere Anforderungen:

Zweck: Korpora werden meist für bestimmte Forschungszwecke gezielt zusammengestellt (Scherer, 2006, 5), die Daten müssen sich also für das eignen, was man erforschen möchte.

Konstanz: Korpora ändern sich gewöhnlich nicht nach ihrer Erzeugung (Scherer, 2006, 6), es sei denn es handelt sich um Monitorkorpora (wie z.B. der Bestand an Zeitungsartikeln, der dem *Digitalen Wörterbuch der Deutschen Sprache* zugrundeliegt und der sich permanent erweitert) (Scherer, 2006, 20).

Synchronie: Korpora sind gewöhnlicherweise für synchrone Untersuchungen angelegt, man spricht auch von *sampled / balanced corpora* (Scherer, 2006, 11). Volatile Phänomene (Jugendsprache, Metaphern) sind daher schwer zu erfassen. Monitorkorpora (s.o.) können als Ganzes natür lich nicht synchron sein, aber ggf. trotzdem synchron ausgewertet werden.

[1] Auch dieser Punkt ist streitbar – OpenAI verwendete z.B. auch Daten, die urheberrechtlich geschützt sind, zum Training von ChatGPT. Ob das Urheberrecht diesen Fall abdeckt, ist derzeit noch nicht geklärt. Jedoch wurde genau heute (11.11.2025) das erste deutsche Gerichtsurteil zu diesem Thema gefällt, demzufolge die Wiedergabe von Liedtexten durch Chatbots eine rechtswidrige Vervielfältigung von urheberrechtlich geschütztem Material darstellt – ein potenziell richtungsweisendes Urteil (Schröder-Ringe, 2025).

Synchronie ist allerdings nicht notwendigerweise ein Kriterium für Korpora, da es auch dezidiert **diachrone Korpora** gibt.

Repräsentativität: Für die quantitative Untersuchung ist es wichtig, dass im Korpus festgestellte Phänomene in derselben relativen Häufigkeit auftreten, wie dies im Gesamtdatenbestand, aus dem das Korpus erstellt wurde, der Fall wäre (externe Validität). Obwohl spezielle Maßnahmen ergriffen werden, um dieses Kriterium möglichst gut zu erfüllen (s.u. bei Korpusgenerierung), bleibt hier immer eine gewisse Unsicherheit bestehen. Bei Umfragen wird versucht, die Representativität sicherzustellen, indem bekannte Eigenschaften der Befragten zwischen Stichprobe und Grundgesamtheit ausgeglichen werden, z.B. die Verteilung auf verschiedene Alterskohorten. Für Texte gibt es solche allgemein anerkannten oder überhaupt nur offensichtlichen Kriterien nicht, daher muss man sich darauf verlassen, dass gleichartige Verfahren bei der Generierung, z.B. gleiche Textlänge, hoffentlich auch größtmögliche Repräsentativität erzeugen.

Referentialität: Eine Sonderform von Korpora sind **Referenzkorpora**, die den gesamten Bestand einer Sprache abdecken sollen. Das ist schon aus den o.a. Gründen schwierig, insbesondere bestehen aber Einwände hinsichtlich der Repräsentativitiät (Scherer, 2006, 24/28).

Wie schon angedeutet, können theoretisch alle Arten von Daten (oder Artefakten) in Korpora zusammengefasst werden. Ernsthafte Anwendungen gibt es z.B. für Filme, Bild-Text-Medien (z.B. Social Media), Videoaufzeichnungen für Gesprächsanalyse oder auch Gebärdensprache (Scherer, 2006, 25).

2.3 Daten in Korpora

Wenn wir in Korpora „Wörter zählen“, unterscheiden wir zwischen drei verschiedenen Repräsentationen:

- **Token:** Kleinste Einheit des Korpus; meistens ein Wort. Jedes Vorkommen und jedes Satzzeichen wird gezählt, Leerzeichen aber nicht.
- **Type:** Die Menge *unterschiedlicher* Tokens – identische Tokens zählen nur einmal.
- **Lemma:** Lexikalische Grundform („Wörterbuchform“) eines Tokens (z.B. *gehen* → *gegangen*).

Sehen wir uns als Beispiel den folgenden Satz an:

> *Wenn Robben hinter Robben robben, robbten Robben hinter Robben.*

- **11 Tokens**[2]**:** `["Wenn", "Robben", "hinter", "Robben", "robben", ",", "robbten", "Robben", "hinter", "Robben", "."]`

- **7 Types:** `{"Wenn", "Robben", "hinter", "robben", ",", "robbten", "."}`

- **4 Lemmata**[3]**:** `{"Wenn", "Robbe", "hinter", "robben"}`

Jede dieser Einheiten kommt mit ihren eigenen, kleinen Herausforderungen. Die Tokenisierung von Dokumenten ist nicht immer eindeutig – bislang haben wir sehr simpel anhand von Leerzeichen tokenisiert, allerdings gibt es Beispiele, bei denen das nicht genügt. Der Ortsname *New York* besteht beispielsweise aus zwei Wortteilen, sollte aber – genau wie *London* – als ein Token aufgefasst werden. Bei regulären Konstruktionen wie *new shoe* sollte allerdings unbedingt die Trennung in zwei Tokens beibehalten werden! Wenn Wörter als kleinste Einheit, also als Tokens dienen sollen, stellt sich generell das Problem, dass Leerzeichen und Wortgrenzen nicht immer identisch sind. Vergleiche die englische Komposita *hand luggage* und *handbook*: Hier liegt offensichtlich die gleiche Konstruktion vor, allerdings wird das eine Kompositum zusammen geschrieben und das andere auseinander. Eine gute Tokenisierung sollte idealerweise beide Komposita – trotz ihrer unterschiedlichen Schreibweisen – identisch behandeln. Ein ähnliches Beispiel sind Kontraktionen wie *I will* zu *I'll* – auch hier möchte man vermeiden, dass die gleiche Konstruktion im ersten Fall als zwei Tokens und im zweiten Fall als ein Token interpretiert wird.

Auch bei der Typisierung und der Lemmatisierung gibt es kleinere Probleme. Typen sind zwar sehr einfach zu inferieren, da wir lediglich prüfen müssen, ob der gleiche String vorliegt oder nicht. Das führt allerdings dazu, dass Homonyme und Polyseme zusammengeführt werden (z.B. *Bank* als Sitzgelegenheit oder als Finanzinstitution), während unterschiedliche Formen des selben Lemmas getrennt bleiben. Letzteres kann insbesondere bei morphologisch komplexen Sprachen (z.B. Finnisch oder Türkisch) problematisch sein, die aus einem Stamm teilweise hunderte von Flektionsformen bilden können. Eine gute Lemmatisierung löst dieses Problem, allerdings bedarf sie entweder aufwändiger, manueller Annotation oder komplexe Computermodelle, die keineswegs komplett verlässlich sind. Neben Lemmatisierungen können über Annotationen auch viele weitere Metainformationen hinzugefügt werden, wie zum Beispiel syntaktische Zusammenhänge, morphologische Prozesse oder Eigennamen.

[2]Wenn wir Groß- und Kleinschreibung ignorieren – was auch häufig getan wird – fallen „Robben“ und „robben“ zusammen; in diesem Fälle würden wir 6 Types zählen.

[3]Nur tatsächliche Wörter können lemmatisiert werden; Satzzeichen fallen hier also heraus.

2.4 Generierung von Korpora

Korpora können nach den o.a. Kriterien aus vorhandenen (digitalen oder analogen) Quellen selbst zusammengestellt werden. Sofern die Daten aus unterschiedlichen Quellen stammen, sollte die Datenmenge aus allen Quellen gleich gehalten werden, da sonst Beeinträchtigungen der Repräsentativität zu erwarten sind. Das erreicht man durch Trunkierung der Texte auf gleiche Länge. Allerdings sollte bei Quellen mit starker innerer Strukturierung auch der Ausschnitt passen, damit man nicht z.B. aus Zeitungsquellen von einer Quelle hauptsächlich den Wirtschaftsteil in den Korpus aufnimmt und von einer anderen den Sportteil.

Alternativ können auch Daten, die durch Befragungen, Beobachtungen oder Experimente erhoben wurden, als Korpora fungieren. Eine weitere Sonderform sind Daten aus dem Internet, insbesondere aus Social Media, die wir uns in einer späteren Sitzung gesondert anschauen.

Beim Zusammenstellen und Nutzen von Korpora sind häufig diverse Normalisierungsschritte notwendig. Unter Normalisierung versteht man die Zusammenführung verschiedener Schreibweisen – das geht typischerweise mit einer Lemmatisierung einher. Allerdings werden auch variierende Normen (z.B. <ß> in Deutschland und Österreich vs. <ss> in der Schweiz) und eventuelle Fehlschreibungen vereinheitlicht. Bei stärker automatisierten Ansätzen zur maschinellen Sprachverarbeitung *Natural Language Processing*, kurz *NLP*) kommen häufig einige Normalisierungsschritte zum Einsatz, die die Verarbeitung der Daten vereinfachen. Häufig wird Groß- und Kleinschreibung ignoriert, indem alles in Kleinbuchstaben gesetzt wird. Des weiteren können häufige Funktionswörter (wie *und, oder, der, die*) ignoriert werden, die eine vergleichsweise schwache Semantik haben. In gleicher Weise werden häufig *Hapax Legomena*, also Tokens, die nur einmal in einem Korpus vorkommen, ignorieren, um die Komplexität statistischer Berechnungen zu senken.

In allen Fällen gilt jedoch: Eine korrekt augeführte Normalisierung ändert die Primärdaten nicht, sondern ist eine Annotation. Die Primärdaten sollten niemals durch Normalisierungsprozesse überschrieben werden; viel eher ist sicherzustellen, dass diese Normalisierungsprozesse transparent und reproduzierbar sind.

3 Anwendungsfälle

Korpuslinguistische Methoden können auf einfache Weise zur Theorieprüfung verwendet werden. So kann man mit Hilfe des Monitorkorpus des DWDS herausfinden, ob kausale Relativsätze mit ‚weil' und Verbzweitstellung (statt Verbendstellung) mittlerweile zunehmend in der Schriftsprache auftreten, was zu erwarten ist, wenn man behauptet, dass ihre Akzeptabilität allgemein zunimmt. In gleicher Weise kann man natürlich auch Futter für die

Abbildung 1: Semantische Nachbarn der Wörter *gay, broadcast* und *awful* im Laufe der Zeit (Hamilton et al., 2016, 1490).

Bildung neuer Theorien suchen. Ein reales Beispiel ist der quantitative Vergleich von Fehlern bei L1 und L2-Sprecherinnen+ mit Hilfe des Falko-Korpus der Humboldt-Universität Berlin. Hieraus könnte man z.B. schließen, welchen Fehlern der Fremsprachenunterricht besondere Beachtung schenken sollte (Lüdeling, 2007, 33).

Weitere Beispiele für die Verwendung diachroner Korpora sollen im Folgenden kurz beleuchtet werden. Eine unveröffentlichte Korpusanalyse[4] zeigt, dass der Präpositionsgebrauch nach dem Verb ‚kämpfen' relativ zügig (<100 Jahre) von ‚wider' zu ‚gegen' wechselte. Abb. 1 visualisiert anhand von semantischen Nachbarn die Bedeutungswandel der Wörter *gay, broadcast* und *awful*, wobei diese Wortfelder automatisch in Form von Word Embeddings aus einem diachronen Korpus berechnet wurden (Hamilton et al., 2016). Abb. 2 illustriert wiederum eine Veränderung eines Diskurses, die sich dadurch zeigt, dass neben dem Begriff ‚Klimawandel' in den vergangenen Jahren auch zunehmend von einer ‚Klimakrise' gesprochen wird.

[4] Dagobert Höllein, persönliche Kommunikation

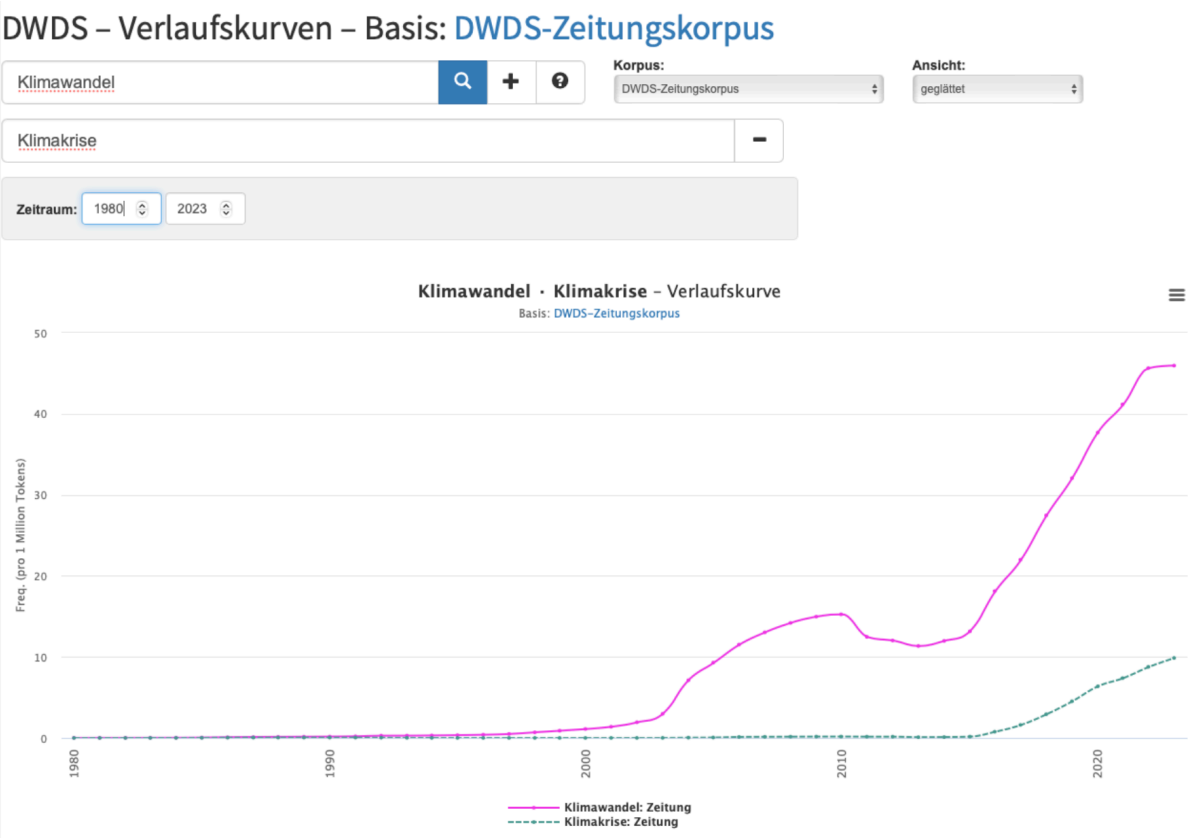

Abbildung 2: Vorkommen der Wörter ‚Klimawandel' und ‚Klimakrise' in deutschen Zeitungsartikeln im Laufe der Zeit ("DWDS-Zeitungskorpus", 2025).

Literatur

DWDS-Zeitungskorpus. (2025). In Berlin-Brandenburgische Akademie der Wissenschaften (Hrsg.), *DWDS – Das Wörterbuch der Deutschen Sprache*.

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)* (S. 1489-1501). Berlin: ACL. Zugriff auf https://www.aclweb.org/anthology/P16-1141 doi: 10.18653/v1/P16-1141

Hirschmann, H. (2019). *Korpuslinguistik: Eine Einführung*. J.B. Metzler.

Lüdeling, A. (2007). Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In W. Kallmeyer & G. Zifonun (Hrsg.), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*. De Gruyter.

Scherer, C. (2006). *Korpuslinguistik*. Winter.

Schröder-Ringe, P. (2025). *GEMA gewinnt gegen OpenAI: Stärkung des Urheberrechts im KI-Zeitalter*. Berlin: Härting Rechtsanwälte PartGmbB. Zugriff auf https://haerting.de/wissen/gema-gewinnt-gegen-openai-staerkung-des-urheberrechts-im-ki-zeitalter/