Originalveröffentlichung in: Diskursmonitor.
https://diskursmonitor.de/glossar/korpus/
(Stand: 01.08.2023, 12:15 Uhr)

31. Januar 2021

Jan Oliver Rüdiger

Korpus

Kurzzusammenfassung

In den Sprach- als auch Literaturwissenschaften versteht man unter *Korpora* (Plur. Korpora, die / Sing. Korpus, das) ganz allgemein Textsammlungen. Nach Lemnitzer und Zinsmeister (2010, S. 40) ist ein *Korpus*: „[…] eine Sammlung [authentischer] schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen“. Die Zusammenstellung erfolgt nach verschiedenen wissenschaftlichen Kriterien, die sich am zu untersuchenden Gegenstand orientieren (Bsp. 1: Soll strategische Kommunikation in politischen Reden analysiert werden, so wird ein *Korpus* aus ‚Politischen Reden‘ zusammengestellt, die strategisch/kommunikative Praktiken enthalten – Bsp. 2: Für die Analyse von Modalpartikeln im Fremdsprachenerwerb wird ein *Korpus* aus transkribierten Redebeiträgen verschiedener Erwerbsstufen benötigt). Prinzipiell kann ein *Korpus* auch analog (gedruckt) vorliegen und manuell ausgewertet werden – In der empirischen Linguistik ist ein *Korpus* aber i. d. R. immer ein digitales (maschinenlesbares) *Korpus*, das automatisiert (mittels Software) ausgewertet wird.

Erweiterte Begriffsklärung

Korpora bestehen nicht nur aus den Texten (den so genannten Primärdaten), sie umfassen auch eine ganze Reihe weiterer Sekundärdaten. Im Wesentlichen handelt es sich dabei um Metadaten und Annotationen (vgl. hierzu Perkuhn et al. (2012)). Metadaten sind Zusatzinformationen zu einzelnen Texten (z. B. der Titel, Autor*in, Datum, Textsorte etc.). Diese Metadaten können in der Korpusanalyse genutzt werden, um etwa Akteursgruppen, Zeitfenster oder Textsorten miteinander zu vergleichen bzw. sie zueinander in Beziehung zu setzen (z.B. Vergleich von Sprachgebrauchsmuster bestimmter Autor*innen / Unterschiedlicher Sprachgebrauch in zwei oder mehr definierten Zeitfenstern).
Annotationen sind Sekundärdaten, die direkt mit dem Text verknüpft sind. Annotationen können sowohl manuell erstellt oder automatisch erzeugt werden. Elektronische Korpora werden i. d. R. mehrstufig automatisch annotiert. Zusätzliche manuelle Annotationen oder Nachkorrekturen der automatischen Annotation sind je nach Forschungsinteresse notwendig.

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-120368

Folgende automatische Prozessschritte sind weit verbreitet: Zerteilung der Texte in einzelne Sätze, Zerteilung der Sätze in einzelne Token (Unter den Begriff ‚Token' fallen sowohl Wortformen (Berg, Berge, Berges etc.) als auch Satzzeichen), automatische Lemmatisierung der Token (Token: Häuser > Lemma: Haus), automatische Zuordnung der Wortart (Token: Berge > Wortart: Nomen), Annotation von Phrasen (Token: Das wundersame Fest > Phrase: Nominalphrase). Ein so aufbereitetes Korpus erlaubt sehr komplexe Analyse- und Abfragemöglichkeiten (z. B.: Suche alle Sätze mit Nominalphrasen, die das Lemma ‚Krise' enthalten).

Wie Mukherjee (2009) anmerkt, arbeitet die Sprachwissenschaft bereits in der ‚Vor-Computer-Zeit' mit Textsammlungen, also Korpora. So wird u. a. das Beispiel der Konkordanz-Analyse der *King James Bibel* von Alexander Curden aus dem Jahre 1736 als eine händische Korpusanalyse angeführt. Bei dieser Analyse werden Konkordanzen/Belegsätze geordnet und ggf. gefiltert. Durch diese Art der Darstellung lässt sich der Kontext einfacher erkennen und auswerten. Während man in der ‚Vor-Computer-Zeit' auf händische Arbeit angewiesen war (Belege abschreiben/abtippen, ausschneiden, auf Karteikarten aufkleben etc.), kann eine solche Konkordanz-Analyse heutzutage mittels entsprechender Software in Sekundenschnelle auf riesige Korpora angewendet werden. Folgendes Beispiel zeigt eine Konkordanz-Analyse zum Stichwort ‚Europa' in Plenarprotokollen des Europäischen Parlaments:

Dennoch muß die französische Delegation der Fraktion der Union für das	Europa	der Nationen einige der darin enthaltenen Behauptungen mißbilligen.
Denn was uns in	Europa	verbindet, sind an erster Stelle doch die Werte und Grundsätze der Freiheit, der Demokratie und der Achtung der Menschenrechte.
Denn	Europa	ist und bleibt die stärkste Garantie gegen Fremdenhaß und Demagogie oder gar den Rückfall in die Barbarei.
Dem können wir nicht zustimmen, denn ein solches Abgleiten in ein	Europa	der Richter wäre dem ordnungsgemäßen Funktionieren der Demokratie in unseren Staaten abträglich.
Das wurde in Südostasien nie gemacht, das wurde in	Europa	nie gemacht, und daß der Kollege Karas aus Österreich nicht mehr da ist, das kann ich auch verstehen, bei deren Finanzgebaren.
Das Weiterbestehen der Dossiers müßte einem	Europa	unangemessen sein, das sich alle sechs Monate verändern würde, doch es ist ein Mechanismus, der sich durch neue Herangehensweisen erneuert und u...
Das weckt in diesen Menschen natürlich das Gefühl, daß	Europa	den Anstrengungen kulturellen Schöpfertums fremd, abweisend, ja nachgerade feindlich gegenübersteht, für das sich die Europäische Union doch eigent...
Das war lediglich eine Erweiterung der politischen Ziele der Vollendung des Binnenmarktes in	Europa	.
Das soziale	Europa	, das die EU benutzt, um sich in ein besseres Licht zu rücken, ist völlig von der Tagesordnung verschwunden.
Das sind die öffentlichen Hilfen in	Europa	, die nach diesem Kodex vergeben wurden und die die Kommission geprüft hat.
Das schlägt meine Parlamentsfraktion, die Fraktion für das	Europa	der Demokratien und der Unterschiede, ebenso vor wie unsere gemeinsame Gruppe von Gesinnungsgenossen, die sich SOS Demokratie nennt.
Das Parlament hat traditionell in hohem Maße zu einer starken Umweltgesetzgebung in	Europa	beigetragen, und es würde mich verwundern und betrüben, wenn dies nicht auch heute der Fall wäre.

Grafik 1: Konkordanzen als KWIC (Keyword in Context) dargestellt.

Die ersten computergestützten *Korpora* entstanden Mitte des 20. Jahrhunderts. Das Brown University Corpus of Present-Day American English (vgl. Francis und Kučera (1964) war das erste rein computergestützte *Korpus* und umfasste bereits eine Millionen Wörter. Es setzte sich, wie der Name schon andeutet, aus unterschiedlichsten schriftsprachlichen Genres der amerikanischen Gegenwartssprache

zusammen. Ebenfalls 1964 entstand mit dem Mannheimer Korpus I (MK_I) am *Leibniz-Institut für Deutsche Sprache* unter der Leitung von Paul Grebe und Ulrich Engel ein vergleichbares *Korpus* des Deutschen mit sogar bereits 2,2 Mio. Token. Durch die Verbreitung und schnell steigenden Computerkapazitäten wuchsen auch die Möglichkeiten der Korpuslinguistik. Bereits in den 1990ern erreichte das British National Corpus (BNC) einen Umfang von über 100 Millionen Token. Im Jahr 2000 überschritt das Deutsche Referenz Korpus (DeReKo) – basierend auf dem Projekt *Mannheimer Korpus* (siehe oben) – die Schwelle von einer Milliarde Token. Mit Stand 02.02.2021 umfasst *DeReKo* 50 Milliarden Token (vgl. zur Entwicklung auch Kupietz et al. (2018)). Viele, auch kleinere, spezifischere Korpusprojekte entstanden in den letzten Jahrzehnten. Eine gute Ausgangsbasis für eigene Recherchen bietet die Plattform CLARIN – hier sind viele freie Korpus-Ressourcen gelistet. Für das Teilprojekt Barometer im *DiskursMonitor* wurden verschiedene freie Korpora als Referenzkorpora aggregiert (eine Beschreibung finden Sie hier). Außerdem wurde eine Infrastruktur entwickelt, die eine LIVE-Analyse ermöglicht. Die *Korpora* stehen intern für Lehrstuhlprojekte aber auch externen Forschenden zur Verfügung. Besucher*innen der Webseite können auf aggregierte Korpusdaten einfache Analysen durchführen (zu den Analysen).

Beispiele

Korpora dienen nicht nur zur Suche nach Belegen. Auf ein *Korpus* lassen sich verschiedene (statistische) Analysen anwenden (sowohl auf die primären Textdaten als auch auf die sekundären Metadaten/Annotation – und auch in Kombination [Text <> Metadaten/Annotation]). Einen guten Überblick und eine Einbettung/Anknüpfung in die linguistischen Grundlagen bietet Bubenhofer (2009). Im folgenden haben wir eine Liste mit beispielhaften Analysen zusammengestellt, die auf den *Barometer-Korpora* (siehe oben) basieren und die Sie selbst explorativ testen können:

- **Frequenzanalysen** stellen eine einfache und effektive Möglichkeit dar, große Textsammlungen (Korpora) zu untersuchen. Mit der Frequenzanalyse lassen sich verschiedene Fragen beantworten (für detaillierter Informationen siehe Frequenzanalyse):
  → Was sind die häufigsten Token (z. B. Lemmata) in einem Korpus?
  → Wie oft kommt ein bestimmtes Token im Korpus vor?
  → Wie oft kommt ein bestimmtes Token zu einem bestimmten Zeitpunkt vor?
  Dadurch wird die oben angesprochene Kombination aus Text und

Metadaten (Zeit) zum Analysegegenstand. Das Resultat sind z. B. folgende Zeitverlaufsanalysen:

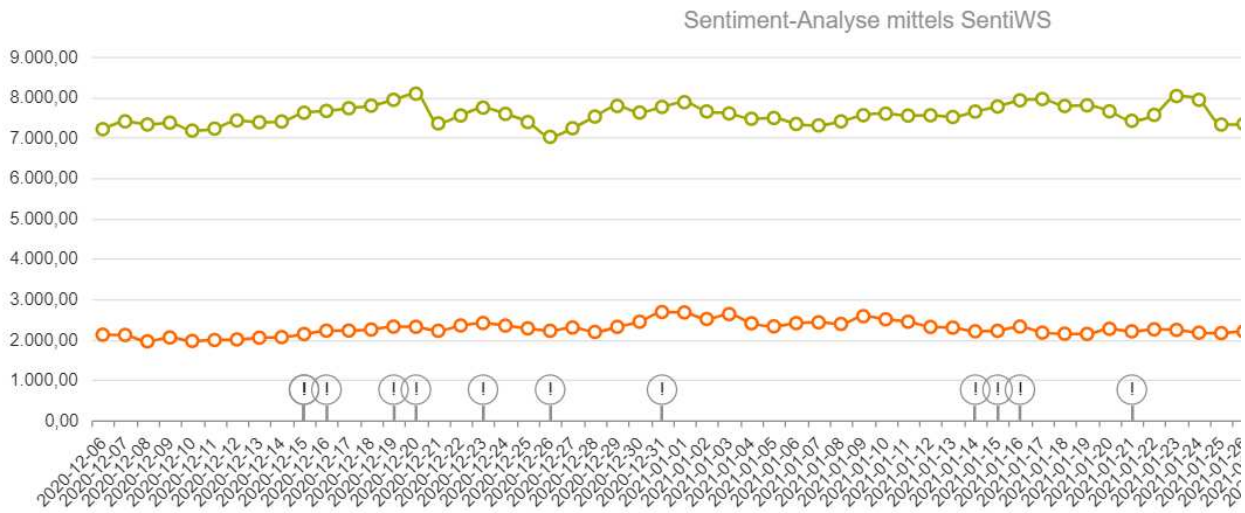
- JSON-Rohdaten
- PNG-Grafik
- TSV-Tabelle
- EXCEL-Tabelle
- PDF-Tabelle

Relative Frequenzen: ☐ Granulierung:

- Tag
- Woche
- Monat

Grafik 2: Beispiel Frequenzverlauf zu ‚SARS-CoV-2'

- Eine mögliche Anwendung der oben genannten Frequenzanalyse ist die **Sentiment-Analysen** (Sentiment von engl. Gefühl/Empfindung). Im Wesentlichen handelt es sich um eine Positiv/Negativ-Frequenzanalyse. Dieses Verfahren hat methodische Grenzen und ist daher nur unter Vorbehalt einsetzbar: Zum einen fixiert das Verfahren die Textoberfläche und erkennt Tiefenstrukturen, Kontexte, Ironie und ähnliche sprachliche Phänomene nicht; zum anderen ist die Erstellung der Ausgangslisten und damit Bewertung von Einzeläußerungen oftmals von der subjektiven Einordnung des/der Bearbeiter*in abhängig. Für weiterführende Informationen zur Sentiment-Analyse siehe hier.

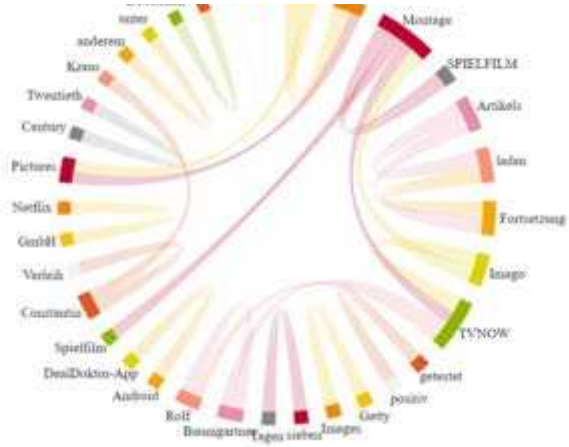


Grafik 3: Sentiment-Detection im diskursmonitor LIVE-Korpus

- **Schlagwörter** (Keywords) geben Auskunft darüber, welche Ausdrücke zu einem Zeitpunkt (verglichen mit einem anderen Zeitpunkt) besonders auffällig häufig sind. Schlagwörter werden mithilfe eines statistischen Signifikanztests ermittelt (hier: Poisson-Verteilung). Die folgende Auswertung vergleicht unterschiedlich große Zeitabschnitte miteinander und ermöglicht damit, kurzfristige von langfristigen Diskurstrends zu differenzieren. Eine tagesaktuelle Auswertung finden Sie hier.

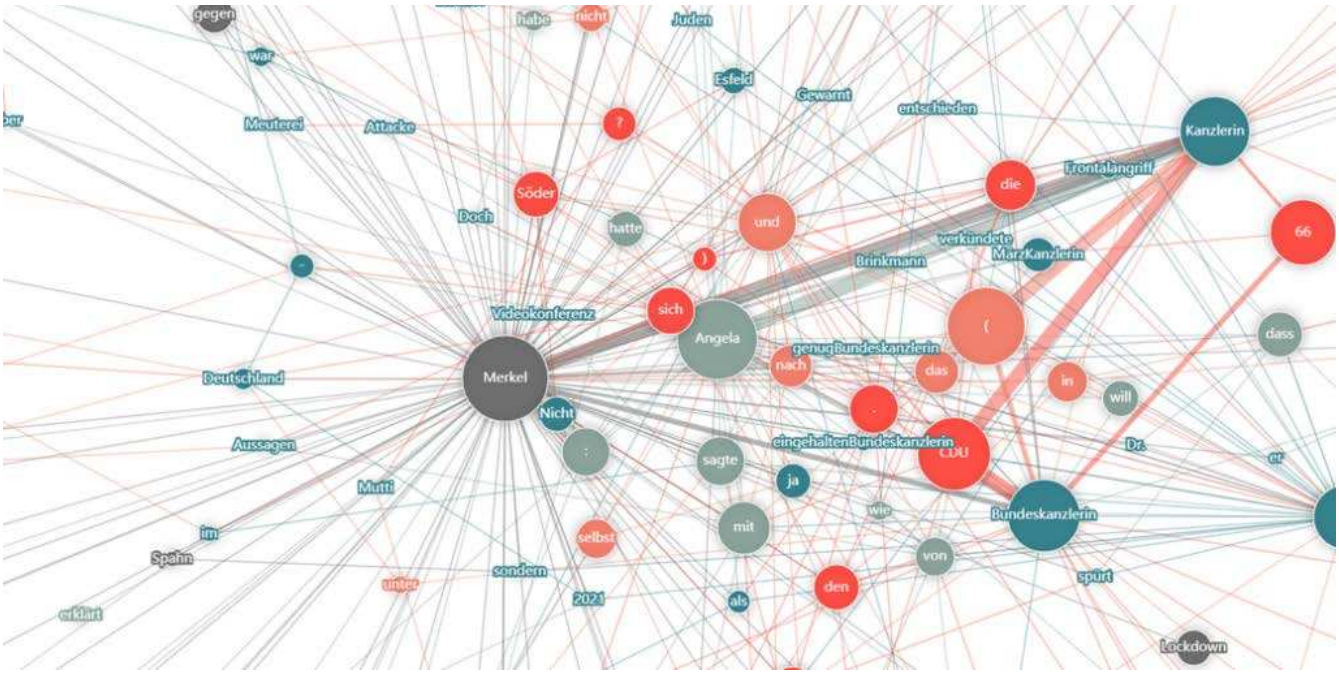
Grafik 4: Beispiel-Schlagworte

- **Kookkurrenz-Analysen – Kookkurrenzen** (engl. co-occurrence – gemeinsames Vorkommen) sind zwei Token (hier: Wörter), die innerhalb einer Textmenge besonders häufig zusammen (hier: innerhalb eines Satzes) und selten in anderen Kombinationen vorkommen (z. B. mit anderen Token). Eine interaktive Analyse finden Sie hier.



Grafik 5: Beispiel-Kookkurrenzen

- **N-Gramme** sind hochfrequente Wortfolgen und gehören zur Analysekategorie der Mehrworteinheiten. Das N steht dabei für eine natürliche Zahl größer 0. Durch N-Gramme lassen sich Mehrwort-Verbindungen und Sprachgebrauchsmuster in Texten identifizieren. Solche Muster können z. B. Floskeln (z. B. „Guten Tag“, „Meine Damen und Herren“ etc.), Phrasen (z. B. „den Tag nicht vor dem Abend loben“ etc.) oder syntaktische Abfolgen realisieren (z. B. „ich gehe ins“ etc.). Weitere Details zu N-Grammen finden Sie hier.



Grafik 6: N-Gramme zu ‚Merkel‘

- **Konkordanz-Analyse / KWIC-Analyse**
  Wie im ersten Nutzungsbeispiel (siehe oben – Konkordanzen) beschrieben, kann eine Konkordanz-Analyse Einblicke in die Realisationsformen eines

Token geben. Wenn Sie das einmal selbst ausprobieren möchten, finden Sie hier die Möglichkeit und weiterführende Informationen.

Grafik 7: KWIC-Belege zu Merkel

Literatur

Zitierte Literatur

- Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Zugl.: Zürich, Univ., Diss., 2008. Berlin: de Gruyter.
  Calzolari, Nicoletta et al. (Hrsg.) (2018): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).
- Francis, W. Nelson; Kučera, Henry (1964): Manual of Information to Accompany, A Standard Sample of Present-Day Edited American English, for Use with Digital Computers*. Brown University, Providence. Department of Linguistics.
- Kupietz, Marc et al. (2018): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, Nicoletta et al. (Hrsg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).
- Lemnitzer, Lothar; Zinsmeister, Heike (2010): Korpuslinguistik. Eine Einführung. Tübingen: Narr Verlag.
- Mukherjee, Joybrato (2009): Anglistische Korpuslinguistik. Eine Einführung. Berlin: Schmidt.

- Perkuhn, Rainer; Keibel, Holger; Kupietz, Marc (2012): Korpuslinguistik. Paderborn: Fink.
- Scherer, Carmen (2014): Korpuslinguistik. Heidelberg: Winter.
- Stede, Manfred (2007): Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik. Tübingen: Narr.

Abbildungsverzeichnis