Inhalt

Vorwort

Diese Einführung in die Korpuslinguistik richtet sich an Studierende der Linguistik und alle anderen Interessierten. Wir setzen kein Fachwissen voraus, Kenntnis linguistischer Grundbegriffe ist jedoch hilfreich.

Kapitel 1 umfasst eine Einführung in korpuslinguistische Grundbegriffe, Kapitel 2 eine praktische Einführung in die Suche in Korpora. Annotationen und ihre Abfrage sind Gegenstand von Kapitel 3 und 4. Kapitel 5 und 6 befassen sich mit der Erstellung und Auswertung eigener Korpora. Kapitel 7 gibt abschließende Hinweise zur Präsentation korpuslinguistischer Studien in wissenschaftlichen Texten. Für eine Vertiefung der Inhalte empfehlen wir das Narr Studienbuch *Korpuslinguistik* von Lothar Lemnitzer und Heike Zinsmeister (2015, 3. Auflage).

Begleitend zu diesem Buch finden Sie unter doi.org/10.5281/zenodo.1410445 ein Korpus zur Textsorte Foodblog zum Download. Viele Beispiele in diesem Buch können Sie anhand dieser Daten selbst reproduzieren. Wir danken dem Kurs „Korpuslinguistik“ (Sommersemester 2018) an der Universität Hamburg für die Hilfe bei der Erstellung des Korpus.

Für die Unterstützung bei der Gestaltung dieses Buches danken wir unserer Kollegin Sarah Jablotschkin sowie Julia Schumacher und Tillmann Bub vom Narr Francke Attempto Verlag.

Hamburg, im Dezember 2018
Melanie Andresen und Heike Zinsmeister

1. Korpora kennenlernen

Wie lösen wir dieses schwere/schwierige Problem?[1] Sind beide Varianten möglich oder gibt es hier einen Unterschied? Wie gehen Sie vor, wenn Sie beim Schreiben eines Textes Zweifel an einer Formulierung haben? Sie können jemand anderen fragen, vielleicht sogar eine Expertin oder einen Experten in einer Schreibberatung. Oder Sie schlagen in einem Lexikon oder einer Grammatik nach. Wahrscheinlich werden Sie in vielen Fällen einfach googeln. Eventuell finden Sie Webseiten, die Ihren Problemfall unmittelbar thematisieren, damit sind Sie quasi wieder bei der Expertenmeinung (auch wenn das Expertenniveau im Internet nicht immer nachvollziehbar ist). Spannender ist es, wenn Sie keine spezielle Seite finden, die Ihr Problem erklärt, sondern wenn Sie anhand der Trefferliste erkunden, welche Varianten andere Autoren und Autorinnen gewählt haben und vielleicht sogar eine Verwendungsregel ableiten können. Sie haben dann einen **empirischen** Weg gewählt, indem Sie nicht nur Ihre eigene Sprachkompetenz befragt oder etabliertes Wissen übernommen haben, sondern Ihre Erkenntnis aus Beobachtungen gewonnen haben. Dies ist sinnvoll, weil sowohl das eigene Sprachgefühl als auch überliefertes Wissen falsch sein können. Wenn man bewertet, ob ein Sprachbeispiel akzep-

1 Zur Notation: Sprachbeispiele werden *kursiv* gesetzt, Fachbegriffe **fett** und Ausdrücke, die in Suchfelder eingegeben werden müssen, werden in `Courier` dargestellt. Für Personenbezeichnungen verwenden wir in diesem Buch abwechselnd die weibliche und männliche Form sowie manchmal beide Formen.

tabel ist oder nicht – was auch als **Introspektion** (‚Hineinsehen') bezeichnet wird –, kann das eigene Urteil durch Normregeln getrübt sein. Das kann dazu führen, dass man manchmal Strukturen ausschließt, die man im tatsächlichen Sprachgebrauch durchaus produziert („Man sagt nicht *Da hat er nichts von*, sondern *Davon hat er nichts.*").

In der Korpuslinguistik wird das empirische Verfahren systematisch betrieben, nicht durch einfache Internetrecherchen, sondern auf der Grundlage informierter Abfragen von **linguistischen Korpora**. Diese Korpora, die der Korpuslinguistik ihren Namen geben, sind Textsammlungen, die speziell für empirische linguistische Untersuchungen zusammengestellt wurden. Sie haben gegenüber dem Internet den entscheidenden Vorteil, dass ihr Inhalt nachvollziehbar und damit überprüfbar ist. Bei Internetrecherchen bleibt die genaue Datengrundlage unbekannt, sodass man als Nutzerin nicht genau weiß, welche Webseiten von der Suchmaschine tatsächlich durchsucht werden. Das Internet ist zudem ständig im Wandel: Täglich kommen neue Webseiten hinzu, während andere geändert oder gelöscht werden.[2] Die Datengrundlage ist daher nicht stabil. Bei linguistischen Korpora besteht diese Unsicherheit nicht.[3] Man kann zumindest nachvollziehen, aus welchem Sprachmaterial ein Korpus zum Zeitpunkt der Abfrage besteht und die genaue Wortanzahl bestimmen, die der Untersuchung zugrunde liegt.

2 Über „Internet Archive" (web.archive.org) können viele geänderte oder gelöschte Webinhalte rekonstruiert werden, die in normalen Suchanfragen nicht mehr auftauchen. Diese und alle anderen URLs im Buch wurden im Oktober 2018 zuletzt besucht.

3 Ausnahmen sind hier sogenannte Monitorkorpora, die permanent wachsen und zum Beispiel von Verlagen genutzt werden.

Dies ist wichtig, wenn man verschiedene Häufigkeiten vergleichen möchte.

1.1 Annotation: Interpretationen offenlegen

Eine weitere wichtige Eigenschaft von Korpora ist, dass sie häufig linguistisch aufbereitet sind. Das bedeutet, dass die **Primärdaten**, also die reinen Texte, mit zusätzlichen Informationen angereichert wurden. Dadurch kann man nicht nur nach den Wörtern bzw. genauer, den konkreten Wortformen suchen, aus denen die Texte bestehen, sondern allgemeinere Anfragen stellen und vor allem auch umfassendere Ergebnisse erhalten. Beispielsweise kann es sinnvoll sein, wenn man neben der Wortform *schwierig* auch Treffer erhält, in denen andere Formen von *schwierig* wie *schwierige, schwieriger* und (*am*) *schwierigsten* auftreten. In Korpora wird dies erreicht, indem für jedes Textwort jeweils das **Lemma** hinterlegt wird. Ein Suchprogramm kann dann zusätzlich zur Textoberfläche auch noch auf der Lemmaebene suchen und findet dadurch alle Belege eines Wortes ungeachtet der tatsächlichen Wortform im Text.[4] Man kann sich vorstellen, wie hilfreich dieses Vorgehen ist, wenn man beispielsweise an die Anzahl der Wortformen des Verbs *sein* denkt: *bin, bist, ist, sind, seid, sei, war, wäre, werde, würde, gewesen* usw.[5]

ex\erc\ise

Gehen Sie auf die Webseite www.dwds.de. Das ist die Startseite des Projekts Digitales Wörterbuch der Deutschen Sprache. Geben Sie

4 Suchmaschinen wie Google leisten diese Art der Analyse inzwischen automatisch direkt bei der Abfrage.

5 Für eine vollständige Liste der Wortformen von *sein* siehe www.canoo.net/inflection/sein:V:sein oder www.duden.de/rechtschreibung/sein_Hilfsverb.

> in das Suchfeld folgenden Suchausdruck genau so ein, wie er hier gezeigt wird (einschließlich der Anführungsstriche): `"schweres Problem"`.
> Die Anführungsstriche signalisieren dem System, dass die beiden Wörter unmittelbar aufeinander folgen sollen. Klicken Sie dann auf das Such-Icon und sichten Sie die Treffer im DWDS-Kernkorpus (1900 – 1999). Wie viele Treffer werden Ihnen angezeigt? Wie viele gibt es insgesamt im Korpus? Welche Wortformen wurden gefunden?
> Wiederholen Sie die Anfrage mit `"schwieriges Problem"` und vergleichen Sie die Ergebnisse. Welche Unterschiede stellen Sie in den Häufigkeiten und Verwendungsweisen der beiden Kombinationen fest? Welche Variante würden Sie für den Beispielsatz am Kapitelanfang wählen?

Die Textanreicherung mit zusätzlichen Informationen heißt **Annotation** – sowohl der Prozess des Hinzufügens als auch die hinzugefügte Information selbst. Im Beispiel in der Aufgabe wurde der Text auf Wortebene mit lexikalischen Basisformen oder **Lemmata** annotiert. Eine weitere sehr gängige Annotation ist die Angabe der **Wortart** wie Verb oder Substantiv. Liegt diese vor, kann man zum Beispiel ermitteln, mit welchen anderen Substantiven *schwierig* und *schwer* typischerweise auftreten. Eine Korpussuche nach „*schwierig* + Substantiv“[6] findet Vorkommnisse wie *schwieriges Umfeld* oder *schwierige Lage*. Spontan können wir beurteilen, dass im Vergleich dazu *schweres Umfeld* seltsam klingt. *Schwere Lage* hingegen wäre durchaus möglich, wirkt aber nicht ganz so eingängig wie *schwierige Lage*. Interessant wäre jetzt eine vergleichende Anfrage für „*schwer* + Substantiv“, um zu

6 Im DWDS-Suchfenster lautet die entsprechende Anfrage: `"schwierig $p=NN"`, siehe Kap. 3.2.

ermitteln, mit welchen Substantiven *schwer* typischerweise auftritt und ob es sich größtenteils um dieselben handelt wie bei *schwierig*. Aufschlussreich sind hier besonders Kombinationen wie die mit *Umfeld*, die nur mit einem der beiden Wörter auftreten. Sie weisen auf Bedeutungsunterschiede zwischen *schwierig* und *schwer* hin. Es kann sich dabei auch um mehr oder weniger feste Wendungen handeln, die in Wörterbüchern aufgeführt werden sollten, damit Deutschlernende darauf aufmerksam gemacht werden können.

Neben Lemma und Wortart können je nach Forschungsinteresse der Korpusersteller beliebige andere Informationen im Text annotiert sein. Eigentlich gibt es hier keine Grenzen. Sie sollten sich allerdings klarmachen, dass Sie immer nur solche Informationen abfragen können, die über die reinen Wortformen vermittelt oder in Annotationen hinterlegt sind. Zunächst wollen wir uns noch genauer mit der Korpuszusammenstellung befassen und sehen, welche Möglichkeiten bzw. Grenzen der Nutzung von Korpora damit einhergehen.

1.2 Metadaten: Datenumstände erfassen

Anders als viele Beispielsätze, die man in Grammatiken findet, haben Korpusbelege den Charme, dass sie „aus dem Leben gegriffen" sind, was bedeutet, dass es sich um **authentische** und nicht um erfundene Sprachbeispiele handelt. Authentische Beispiele eröffnen uns einen Zugriff auf die Vielfalt der Sprache und können uns gleichzeitig Hinweise auf mögliche Verwendungsbeschränkungen geben und damit auf das zugrundeliegende Sprachsystem oder relevante kommunikative Regeln.

Um diese Vielfalt der Sprache systematisch ausschöpfen zu können, benötigen wir ähnlich zu den oben eingeführten

linguistischen Annotationen zusätzliche Informationen zu den Texten. Wir würden gerne wissen, wer gesprochen hat bzw. wer der Autor war, wann der Text entstanden ist, in welcher Situation gesprochen wurde bzw. in welchem Kontext ein geschriebener Text veröffentlicht wurde usw. Solche zusätzlichen Informationen werden normalerweise nicht als Annotationen, sondern als **Metadaten** bezeichnet. Es sind „Daten über die (Text-)Daten". Je nach Korpustyp können die Metadaten aber auch direkt im Text annotiert sein, beispielsweise die Angabe der Sprecher und Sprecherinnen in einem Gespräch mit mehreren Teilnehmenden.

Bei geschriebenen Texten wird häufig die **Textsorte** als Metadatum angegeben: Je nach (situativem) Kontext und kommunikativer Funktion unterscheiden sich Texte systematisch. Es macht einen Unterschied, ob man privat eine WhatsApp-Nachricht an die WG-Mitbewohnerin schreibt oder ob man im Rahmen eines Bachelorstudiums eine linguistische Hausarbeit verfasst, die anschließend bewertet wird. Ebenso beeinflusst die kommunikative Funktion des Textes Form und Inhalt: Mit Texten wollen wir informieren, beeinflussen, versprechen, persönliche Kontakte knüpfen und pflegen und manchmal sogar neue Fakten schaffen und damit unmittelbar die Welt verändern, beispielsweise wenn wir kündigen oder jemandem eine Vollmacht erteilen (vgl. Brinker et al. 2014, 139 f.).

Das DWDS-Kernkorpus (1900 – 1999), das Sie in der ersten Aufgabe kennengelernt haben, sieht nur eine sehr grobe Textsorten-Klassifikation für geschriebene Sprache vor. Es unterteilt Texte in Belletristik (d. h. unterhaltende, schöngeistige Literatur), Zeitungs-, Wissenschafts- und Gebrauchstexte. In anderen Online-Korpora können Sie auf feinere Klassifizierungen für Textsorten zugreifen.

Zum Beispiel unterscheiden die Metadaten des Deutschen Referenzkorpus (DeReKo)[7] insgesamt 75 Textsorten wie Leserbrief, Literaturhinweis oder Lokales. Diese beruhen allerdings nicht auf einer linguistisch motivierten Analyse, sondern wurden automatisch aus den Zeitungstexten ermittelt. Daher liegt die Angabe auch nur bei einem Teil der Texte im DeReKo vor.

Ein anschauliches Beispiel für den Einsatz von Metadaten finden Sie in einer korpuslinguistischen Studie zu den Adjektiven *ewig* und *unendlich*. Meißner (2008) untersucht für die Lesart ‚zeitlich ohne Ende/Grenze', ob die Textsorte einen Einfluss auf die Wortwahl hat. Auf der Basis von insgesamt 207 Belegen von *ewig/unendlich lang(e)* aus dem DeReKo stellte sie fest, dass

> „[*ewig lang*(*e*)] eher in den Rubriken Lokales, Sport und Vermischtes vorkam. Auch trat die Verstärkung mit *ewig* eher in der direkten Rede sowie im Kontext von dialektal gefärbten oder umgangssprachlichen Ausdrücken auf. Die Verstärkung mit *unendlich* fand sich hingegen eher in den Rubriken (sachlicher) Bericht oder Kommentar und wurde kaum in dialektal gefärbten oder umgangssprachlichen Kontexten verwendet." (Meißner 2008, 12)

Meißner schließt aus diesen Beobachtungen, dass *ewig* eher in informellen Bereichen verwendet wird und *unendlich* eher „in (Zeitungs-)Texten höherer Stilebene" (Meißner 2008, 13).

7 Das Deutsche Referenzkorpus DeReKo, www.ids-mannheim.de/kl/projekte/korpora/, am Institut für Deutsche Sprache, Mannheim.

1.3 Grundlegende Definitionen

Wir haben den Begriff „das Korpus" – mit seiner Pluralform „Korpora" – bereits mehrfach verwendet und hoffen, dass Sie inzwischen eine gewisse Vorstellung davon haben, was sich dahinter verbirgt. Die folgende Definition aus dem Narr Studienbuch zur Korpuslinguistik von Lemnitzer und Zinsmeister (2015) fasst die wichtigsten Aspekte zusammen:

> Ein **Korpus** ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. (Lemnitzer/Zinsmeister 2015, 13)

In der Alltagssprache stellen wir uns unter „Äußerungen" meist etwas Gesprochenes vor, unter „Text" etwas Geschriebenes. In der Korpuslinguistik werden beide Begriffe für gesprochene und geschriebene Sprache verwendet. Damit Tonaufnahmen im korpuslinguistischen Sinn maschinenlesbar werden, müssen sie zuerst **transkribiert** werden, sodass sie in Textform vorliegen. Die Transkription kann phonetisch sein und möglichst viele Details der Lautsprache abbilden (z. B. durch das Internationale Phonetische Alphabet). Für viele linguistische Fragestellungen ist eine mehr oder weniger orthographische Transkription ausreichend, die neben dem Gesprächsinhalt auch Phänomene der Mündlichkeit wie Abbrüche oder besondere Pausen dokumentiert, z. B. nach dem Transkriptionssystem HIAT (Rehbein et al. 2004).

Korpora geben Zugriff auf authentische Sprachbeispiele und Belege für konkrete linguistische Phänomene. Sie dienen dabei als Mittel, **empirische Evidenz** zur Beantwortung linguistischer Fragestellungen zu finden. Korpusuntersuchungen haben oft das Ziel, diese Evidenz statistisch auszuwerten, um zu überprüfen, ob die Beobachtungen allgemeingültig sind. Statistische Auswertungen sind Ihnen sicher aus der Sozialforschung bekannt. Im bundesweiten Mikrozensus 2017[8] war zum ersten Mal eine sprachbezogene Frage enthalten, die empirisch ermitteln sollte, welche Sprache in deutschen Haushalten vorwiegend gesprochen wird. Hierfür muss zuerst festgelegt werden, was „Haushalte" bedeutet, z. B. auf der Basis von Registern des Einwohnermeldeamts. Aus dieser Gesamtheit wird eine repräsentative, d. h. zufällige Stichprobe gezogen und befragt. Die Ergebnisse der Stichprobe lassen dann Rückschlüsse auf die Gesamtheit der Haushalte zu. Schön wäre es, wenn Korpora ebenfalls eine zufällige Stichprobe der Sprache darstellen würden. Leider gibt es hier ein schwerwiegendes Problem: Die Grundgesamtheit des linguistischen Forschungsobjekts, beispielweise des Deutschen/der Jugendsprache/der Arzt-Patient-Kommunikation/der Rechtssprache usw. ist nicht bekannt. Es ist unmöglich festzulegen, welche geschriebenen oder gesprochenen Äußerungen alle dazugehören. Ganz abgesehen von den vielen möglichen Sätzen, die zufällig noch niemand ausgesprochen oder geschrieben hat. Korpora

8 Siehe z. B. www.datenschutz-hamburg.de/ihrrechtaufdatenschutz/mikrozensus. Warum die Fragestellung aus linguistischer Sicht problematisch ist, erklärt Adler (2018).

sind also keine zufälligen Stichproben der Sprache im statistischen Sinn. Trotzdem liefern sie Linguistinnen und Linguisten empirische Evidenz dafür, wie Sprache tatsächlich verwendet wird, und sind damit objektiver als die Betrachtung von beliebigen Einzelfällen wie der Rückgriff auf die eigene Sprachintuition. Bei der Einzelbetrachtung ist die Gefahr besonders groß, gerade auf die große Ausnahme, den statistischen Ausreißer, gestoßen zu sein und den eigentlichen Trend zu verpassen.

Da die Grundgesamtheit der Sprache nicht bekannt ist, können Korpora auch nicht im statistischen Sinn repräsentativ dafür sein. In der Korpuslinguistik versucht man stattdessen, Korpora so zusammenzustellen, dass sie nach bestimmten Kriterien, von denen man annimmt, dass sie die Sprache beeinflussen, **ausgewogen** sind. Das DWDS-Kernkorpus (1900 – 1999) ist zum Beispiel in Bezug auf die Kriterien Entstehungszeit und Textsorte ausgewogen zusammengestellt und enthält ungefähr gleich viel Textmaterial für jede Textsorte in jeder Dekade des Jahrhunderts. Bevor wir Ihnen zum Abschluss des ersten Kapitels ein paar für uns relevante Korpora des Deutschen vorstellen, wollen wir zuerst noch die Disziplin der Korpuslinguistik definieren, in der Korpora als empirische Evidenz für linguistische Untersuchungen eingesetzt werden:

de\?finition

Man bezeichnet als **Korpuslinguistik** die Beschreibung von Äußerungen natürlicher Sprache, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind. Korpuslinguistik ist eine wissenschaftliche Disziplin, d. h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Korpusbasierte Sprachbeschreibung kann

verschiedenen Zwecken dienen, zum Beispiel dem Fremdsprachenunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen Sprachverarbeitung bzw. Computerlinguistik. (Lemnitzer/Zinsmeister 2015, 14 f.)

1.4 Korpora des Deutschen

Mit dem DWDS-Kernkorpus und dem DeReKo haben Sie schon zwei häufig verwendete Korpora für das Deutsche kennengelernt. Tabelle 1 fasst Metadaten für diese und eine Auswahl weiterer wichtiger Korpora zusammen, auf die wir in dieser Einführung Bezug nehmen. Als Namen verwenden wir Kurzformen. Alle Korpora stehen kostenlos zur Verfügung, wobei Sie sich teilweise zuerst anmelden müssen. Die Größe der Korpora ist entweder in Wörtern oder in sog. Token angegeben. Letztere zählen Satzzeichen als eigenständige Textbausteine mit.

Tab. 1: Korpora des Deutschen

Name	Informationen zum Korpus
DECOW	German web corpus by COW – Corpora from the Web (Bildhauer/Schäfer 2012, Schäfer 2015); automatisch aus dem Internet heruntergeladen; 2011 und 2014; DECOW16A: mehr als 15 Milliarden Wörter in mehr als 17 Millionen Texten. Online: www.webcorpora.org[9]

9 Kurz vor Drucklegung dieses Buches wurde der Zugang zu DECOW für Studierende leider eingeschränkt.