„Welche Arten von Korpora gibt es?“ – Korpustypologie

1. Korpustypologie

„*Korpus ist nicht gleich Korpus.*“ (Scherer 2014: 16)

Wie wir bereits gelernt haben, ist Repräsentativität ein zentrales Merkmal linguistischer Korpora. Ein Korpus soll nämlich eine bestimmte sprachliche Grundgesamtheit abbilden: entweder eine Sprache insgesamt (z. B. Deutsch der Gegenwart) oder eine sog. Varietät, d. h. eine bestimmte Ausprägung einer Sprache (z. B. politische Reden im Bundestag, Frühneuhochdeutsch). Daraus geht hervor, dass Korpora nicht alle gleich sind, sondern sich zumindest darin unterscheiden, welche Grundgesamtheit sie widerspiegeln sollen. Es gibt aber andere Kriterien, hinsichtlich deren Korpora voneinander abweichen können. Diese zu kennen ist nicht unerheblich, und zwar aus mindestens zwei Gründen. Erstens muss man, sobald eine Fragestellung vorliegt, Korpora darauf testen, ob sie genügend Sprachdaten für diese Fragestellung liefern können; die „Korpuslandschaft“ ist allerdings sehr breit und man braucht Kriterien, an denen man sich orientieren kann, um Korpora herauszufiltern. Zweitens ist es in einer wissenschaftlichen Arbeit vorgesehen, dass man das verwendete Korpus kurz beschreibt: Dem Leser muss nämlich durch eine knappe Beschreibung des Korpus klar werden, warum sich dieses Korpus für die Fragestellung der Arbeit eignet.
Eine Übersicht über mögliche Klassifikationskriterien für linguistische Korpora bietet bspw. Scherer (2014: 16–32), wobei sie zwischen formalen und inhaltlichen Kriterien unterscheidet.[1] **Formale Kriterien** hängen mit der Zusammenstellung des Korpus zusammen, z. B.: Besteht das Korpus aus vollständigen Texten oder aus Textauszügen? **Inhaltliche Kriterien** (Scherer 2014: 23–32) betreffen hingegen die Natur der enthaltenen Texte, etwa: Geht es um geschriebene oder gesprochene Sprache?

1.1 Formale Kriterien

Zu den formalen Kriterien zählen **Speichermedium**, **Vollständigkeit der Texte**, **Abgeschlossenheit des Korpus** und **Aufbereitung des Korpus** (Scherer 2014: 17–22). Was diese Kriterien genau bedeuten, sehen wir gleich.

a) Speichermedium

In welcher Form liegt das Korpus vor? Die Korpora, mit denen man meist zu tun hat, sind sog. **maschinen-/computerlesbare Korpora**, d. h. Korpora in elektronischer Form. Die (wenn auch unpraktische) Alternative stellen **Papierkorpora** dar, die, wie der Name schon sagt, in Papierform vorliegen. Künftig werden wir uns ausschließlich mit elektronischen Korpora auseinandersetzen. Es ist jedoch sinnvoll, kurz darüber nachzudenken, dass die Computerlesbarkeit nicht Teil der Definition von linguistischen Korpora sein muss: Es sind andere Eigenschaften, die ein linguistisches Korpus ausmachen (z. B. Repräsentativität).

[1] Ähnlich wie Scherer (2014) gehen Lemnitzer & Zinsmeister (2015: 137–142) vor.

b) Vollständigkeit der Texte

Sind die Texte, die das Korpus bilden, vollständige Texte oder Textauszüge? Demnach spricht man entweder von **Volltextkorpora** oder von **Probenkorpora**. Beide Formen haben Vor- und Nachteile.
Volltextkorpora sind von Vorteil besonders bei textlinguistischen[2] Fragestellungen: Nehmen wir an, Sie interessieren sich für die Textsorte der Kochrezepte und möchten untersuchen, ob und wie der Aufbau von Kochrezepten zwischen zwei Sprachen (z. B. Deutsch vs. Niederländisch) variiert. Dafür eignet sich am besten ein Volltextkorpus, bspw. ein Korpus mit Blogeinträgen, wo hoffentlich auch Kochrezepte in vollständiger Form vorkommen. Ein Nachteil von Volltextkorpora ist hingegen, dass die Länge der Texte nicht so konstant bleibt wie bei Probenkorpora: Angenommen, dass hinter jedem Text jeweils nur ein Sprachproduzent „steckt“, sind Sprachproduzenten nicht in gleichem Maße vertreten, insofern Texte aufgrund ihrer unterschiedlichen Länge mit einem unterschiedlichen Gewicht zum Korpus beitragen. Diese Tatsache kann eine gewisse Relevanz haben, wenn Sie Anomalien (sowohl im positiven als auch im negativen Sinne) in Ihren Korpusdaten beobachten: Sind diese Anomalien auf einen gleichen langen Text bzw. einen gleichen Sprachbenutzer zurückzuführen oder tauchen sie in mehreren Texten bzw. bei mehreren Sprachbenutzern auf?
Das, was bei Volltextkorpora von Vor- und Nachteil ist, ist von Nach- und Vorteil bei Probenkorpora. Da hier nur Textproben vorkommen, lässt sich der Textaufbau bei einer bestimmten Textsorte schwieriger untersuchen; die Textlänge bleibt allerdings konstant, sodass die Sprecher hinter den Texten im gleichen Maße vertreten sind.[3]

c) Abgeschlossenheit des Korpus[4]

Ist das Korpus abgeschlossen oder wird es über die Zeit mit neuen Texten erweitert? Abgeschlossene Korpora werden auch als **statische Korpora** bezeichnet, unabgeschlossene als **Monitorkorpora**. Typische Monitorkorpora sind Zeitungskorpora, die fortlaufend mit neuen Zeitungsausgaben ergänz werden. Besonders beim Gebrauch von Monitorkorpora ist es Usus, im Methodenteil einer wissenschaftlichen Arbeit das Datum anzugeben, an dem die Korpusdaten, auf denen die Arbeit basiert, erhoben wurden.

d) Aufbereitung des Korpus

Ist das Korpus **annotiert** oder **nicht annotiert**? Was bedeutet aber der Begriff ***annotieren*** bzw. ***Annotation*** in diesem Zusammenhang? Stellen wir uns vor, Ihr Untersuchungsgegenstand ist der Gebrauch von *furchtbar* als Gradpartikel[5] vor Adjektiven (*furchtbar langweilig*, *furchtbar kompliziert* usw.). In einem solchen Fall würde leider eine einfache Korpussuche nach dem Wort *furchtbar* nicht reichen, denn sie würde zu viele Fehltreffer ergeben, sprich Korpusbelege für andere Verwendungen von *furchtbar*, die möglicherweise interessant sind, aber nicht Ihren Untersuchungsgegenstand darstellen. Abb. 1

[2] Die Textlinguistik ist ein Zweig der Linguistik, die sich mit dem Aufbau von Texten und mit Textsorten befasst.
[3] Ein zusätzlicher Aufwand bei Probenkorpora besteht allerdings für die Personen, die sie zusammenstellen müssen: „Einerseits müssen Texte, die die Normlänge überschreiten, gekürzt werden. Dies wirft die Frage auf, an welcher Stelle eines Textes die Probe entnommen werden soll, da Texte nach bestimmten Prinzipien strukturiert sind, die sich auch in der verwendeten Sprache niederschlagen. Insbesondere wissenschaftliche Texte folgen zum Teil strengen Gliederungsprinzipien. So kann sich eine Probe, die der Einleitung eines Fachaufsatzes entnommen wird, sprachlich komplett anders gestalten als ein Textausschnitt aus der Mitte oder dem Schluss desselben Aufsatzes. Handelt es sich hingegen um kurze Texte, so müssen möglicherweise mehrere Texte zusammengefasst werden, um die vorgegebene Normlänge zu erreichen.“ (Scherer 2014: 19f.)
[4] Man spricht auch von Persistenz (Lemnitzer & Zinsmeister 2015: 140f.).
[5] Gradpartikeln geben den Intensitätsgrad einer Eigenschaft an (z. B. *sehr salzig*, *zu salzig*).

veranschaulicht das Problem: Würden Sie nach dem Wort *furchtbar* im DWDS-Korpus *Blog* suchen, würde eine zufällige Stichprobe aus 10 Korpusbelegen nur einen passenden Treffer enthalten (*Das wird ja alles ganz furchtbar teuer*, Zeile 4) – was nur 10% (!) der Stichprobe entspricht. Möchten Sie etwa 100 passende Korpusbelege für Ihre Arbeit haben, müssten Sie eine Stichprobe aus ca. 1000 Korpusbelegen erheben und alle Fehltreffer dort manuell entfernen. Dann denken Sie bestimmt: „Was habe ich mir mit der Auswahl dieses Themas angetan…“

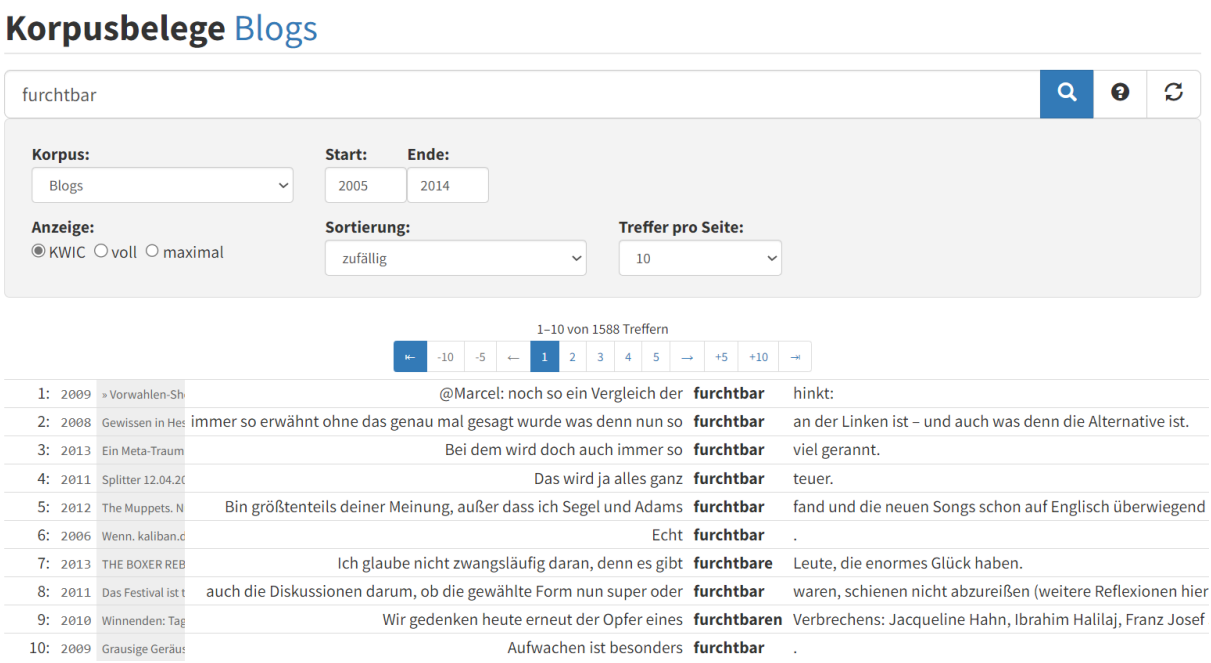

Abb. 1. Suche nach dem Wort *furchtbar* im DWDS-Korpus *Blog*

Vielleicht denken Sie aber: „Wie schön wäre es, wenn ein Korpus zusätzliche Mittel bieten würde, um die Menge der Fehltreffer zu reduzieren.“ Wie schön wäre es, wenn man z. B. nur nach Korpusbelegen mit der Wortform *furchtbar* (d. h. nicht *furchtbare*, *furchtbaren* etc. wie in den Zeilen 7 und 9) suchen könnte, und zwar nur nach denjenigen, die nach der Wortform *furchtbar* ein Adjektiv aufweisen (z. B. *teuer* wie in Zeile 4). Diese zusätzlichen Informationen haben bei näherem Hinsehen mit Sprachwissenschaft zu tun: Die oben genannte Information ‚Adjektiv‘ ist bspw. allgemein betrachtet eine Information über die Wortart und auch mit Wortarten (Nomen, Verben usw.) befasst sich ja die Sprachwissenschaft. Informationen solcherart stehen zwar nicht in allen Korpora zur Verfügung, aber in vielen, und sie haben auch einen Namen, nämlich **(linguistische) Annotationen** (lat. *annotatio* ‚Anmerkung‘). Das heißt, ein **annotiertes Korpus** ist ein Korpus, das die Möglichkeit bietet, die Suche nach einer Zeichenkette (etwa *furchtbar*) mit zusätzlichen Bedingungen – Annotationen – zu ergänzen, z. B.: „Suche nach der Zeichenkette *furchtbar*, wobei die Zeichenkette danach ein Adjektiv sein muss“ (s. Abb. 2).

Welche Arten von linguistischen Annotationen existieren und wie man sie einsetzen kann, soll als separates Thema behandelt werden. Wichtig ist an dieser Stelle, sich darüber klar zu sein, dass es überhaupt Korpora gibt, die mit linguistischen Annotationen versehen sind, die genauere Korpussuchen ermöglichen, wenn die Menge der Fehltreffer ansonsten nicht überschaubar wäre.

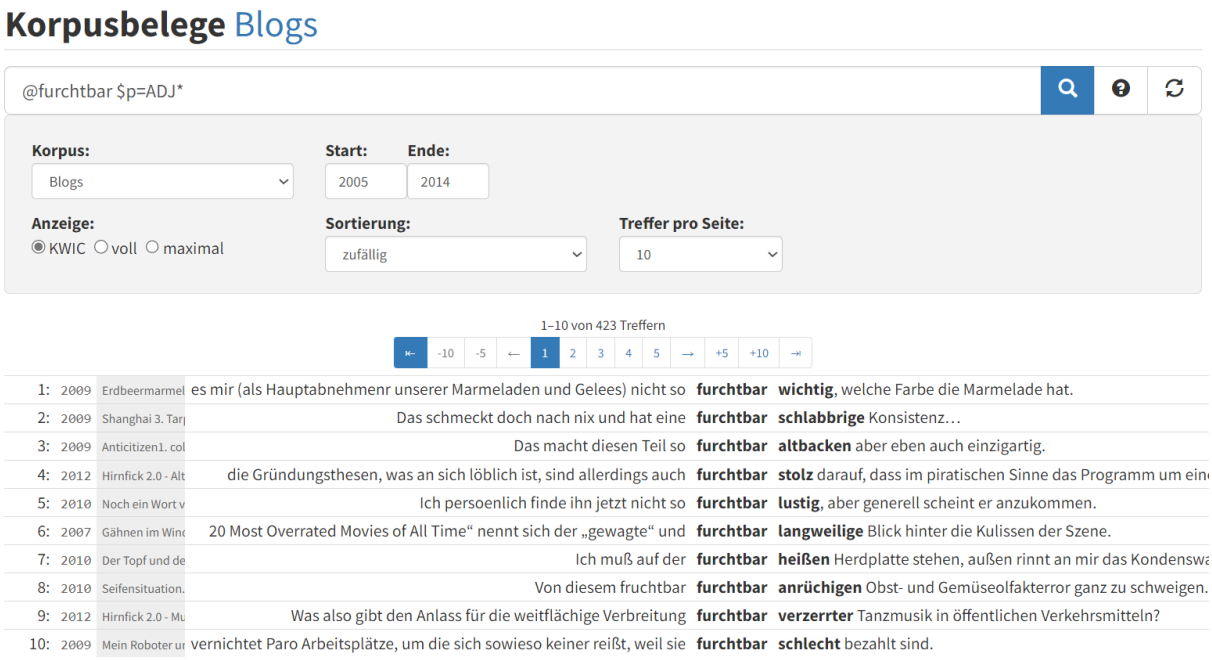

Abb. 2. Einsatz von Annotationen bei der Suche nach der Wortform *furchtbar* im DWDS-Korpus *Blog*

1.2 Inhaltliche Kriterien

Zu den inhaltlichen Kriterien zählen hingegen der **Geltungsbereich des Korpus**, der **zeitliche Bezug**, das **Sprachmedium** und die **Anzahl der Sprachen** (Scherer 2014: 23–32).

a) Geltungsbereich des Korpus

Soll das Korpus eine Sprache insgesamt widerspiegeln oder nur eine bestimmte Varietät? Sog. **Referenzkorpora** haben den Anspruch, eine Sprache in ihrer Gesamtheit abzubilden, z. B. das *COCA* (*Corpus of Contemporary American English*). Für das Deutsche werden in der Regel die folgenden zwei Referenzkorpora erwähnt: das **Deutsche Referenzkorpus (DeReKo)** des Instituts für Deutsche Sprache (IDS) und das **DWDS-Kernkorpus (1900-1999)**. **Spezialkorpora** möchten hingegen nur eine bestimmte Varietät, d. h. eine bestimmte Ausprägung einer Sprache, wiedergeben: Ein **Zeitungskorpus** repräsentiert z. B. nur Zeitungssprache (d. h. die Sprache eines bestimmten Mediums); ein sog. **Lernerkorpus** repräsentiert die Sprache von Nicht-Muttersprachlern (d. h. die Sprache einer bestimmten Sprechergruppe), ein **Webkorpus** die Online-Kommunikation (z. B. das Webkorpus *Ballsportarten*) usw.

b) Zeitlicher Bezug

Umfasst das Korpus Gegenwartssprache oder Sprache aus einer älteren Sprachstufe (etwa Frühneuhochdeutsch)? Demnach spricht man entweder von **Korpora der Gegenwartssprache** oder von **historischen Korpora**. Ein Korpus kann allerdings je nach Fragestellung als historisch betrachtet werden oder nicht: „[M]anche sprachlichen Veränderungen wie etwa Neuerungen im Wortschatz [finden] innerhalb relativ kurzer Zeit statt[], wohingegen sich andere Veränderungen, etwa in der Satzstruktur, deutlich langsamer vollziehen“ (Scherer 2014: 25f.).

c) Sprachmedium

Geht es im Korpus um Äußerungen der gesprochenen oder der geschriebenen Sprache? Die meisten Korpora sind **Korpora der geschriebenen Sprache**. Das ist nicht überraschend, denn bei **Korpora der**

gesprochenen Sprache ist der Korpusaufbau wesentlich zeitintensiver (man muss Gespräche aufnehmen, diese transkribieren usw.), weshalb sie einen kleinen Anteil der „Korpuslandschaft“ darstellen. Gelegentlich lassen sich aber auch Mischformen, sog. **multimodale Korpora**, finden, nämlich Korpora, die sowohl geschriebene als auch gesprochene Sprache umfassen, etwa das ICE (*International Corpus of English*).

d) Anzahl der Sprachen

Enthält das Korpus Texte nur aus einer Sprache oder aus mehreren Sprachen? Hier wird grob zwischen **einsprachigen/monolingualen Korpora** und **mehrsprachigen/multilingualen Korpora** unterschieden. Ein besonderer Fall von mehrsprachigen Korpora sind sog. **Parallelkorpora**. Parallelkorpora umfassen Originaltexte einer Sprache und deren Übersetzung in eine oder mehrere Sprachen: Ein Musterbeispiel ist das Korpus *Europarl*, das Protokolle des europäischen Parlaments in über 20 Sprachen enthält. Solche Korpora sind offensichtlich für sprachvergleichende Arbeiten gedacht.

2. Praxis

2.1 Korpusbeschreibungen dekonstruieren

Jetzt, wo wir mehrere Klassifikationsmerkmale für Korpora kennengelernt haben, sollten wir auch in der Lage sein, diese Merkmale wiederzuerkennen, wenn wir empirische Studien lesen, in denen mit Korpora gearbeitet wurde und Korpusbeschreibungen vorkommen. Nehmen wir etwa die Studie von Stathi (2012), wo der Gebrauch von *selbst* vs. *persönlich* anhand eines Korpus untersucht wurde. Abb. 3 zeigt, wie die Autorin im Methodenteil das Korpus beschrieben hat.

> Für diese Korpusstudie wurde das Korpus des Instituts für deutsche Sprache in Mannheim *Cosmas II* verwendet.[10] Genauer gesagt ist die Datenbasis der Korpusstudie der Teil „W-öffentlich“ des Mannheimer Korpus, der alle öffentlichen Korpora des Archivs W (50%) enthält. Das Archiv W umfasst 162 Korpora, die eine Größe von etwa zehn Millionen Texten und ungefähr drei Milliarden Textwörtern (sogenannte Tokens) haben. Diese Texte stammen aus dem Zeitraum 1946 bis 2008. In Bezug auf die Textsorte handelt es sich vor allem um Zeitungstexte.

Abb. 3. Beispiel für eine Korpusbeschreibung (Stathi 2012: 62)

Zunächst wird erwähnt, wie das Korpus heißt und wo dies zu finden ist (der Link wird in der Fußnote 10 angegeben). Die Korpusgröße wird genannt. Daraufhin geht die Autorin auf die Natur der Korpustexte ein. Sie nennt den Zeitraum, den die Texte abdecken (Kriterium „zeitlicher Bezug“) und sagt uns, dass eine bestimmte Textsorte (Zeitungstexte) deutlich überwiegt, was implizit dafür spricht, dass das Korpus eher Richtung Spezialkorpora geht als Richtung Referenzkorpora (Kriterium „Geltungsbereich des Korpus“). Da es von Zeitungstexten die Rede ist, lässt sich entnehmen, dass es sich um ein Korpus der geschriebenen Sprache handelt (Kriterium „Sprachmedium“). Es versteht sich von selbst, dass das Korpus monolingual ist (Kriterium „Anzahl der Sprachen“). Auf die restlichen Kriterien wird nicht eingegangen; das bedeutet, dass sie keine Relevanz für den Untersuchungsgegenstand haben. Ein Musterbeispiel ist das Kriterium „Aufbereitung des Korpus“: Hätte die Autorin sich linguistischer Annotationen bedient, hätte sie in der Korpusbeschreibung auf jeden Fall erwähnt, dass das Korpus annotiert ist (und wie). Das ist aber eben nicht der Fall ist und diese Tatsache wird dann bestätigt, wenn sie beschreibt, wie sie Korpusbelege herausgefiltert hat, die ihren Untersuchungsgegenstand enthalten

(s. Abb. 4): Sie hat einfach nach den Wortformen *selbst* und *persönlich* gesucht, ohne also linguistische Annotationen gebrauchen zu müssen.

> Die Textbasis wurde nach den Wortformen *selbst* und *persönlich* durchsucht.[11] Die Suche ergab insgesamt 562.799 Treffer (d.h. Textbeispiele) für *selbst*[12] und 46.075 Treffer für *persönlich*. Aus dieser Gesamtmenge wurden jeweils 1000 Treffer für beide Intensifikatoren zufällig ausgewählt und in die Analyse einbezogen. Alle Beispiele wurden manuell sortiert und ihrer Bedeutung und Funktion nach (siehe Abschitt 3) klassifiziert.

Abb. 4. Beschreibung einer Korpussuche (Stathi 2012: 62)

Was können wir nun anhand dieses Beispiels lernen? Erstens reicht ein Paragraf von wenigen Zeilen aus, um ein Korpus adäquat zu beschreiben. Zweitens müssen Sie in Ihrer Korpusbeschreibung nicht auf alle möglichen Kriterien eingehen, anhand deren Korpora klassifiziert werden, sondern nur auf diejenigen, die jeweils relevant sind: Wenn für Ihre Fragestellung unerheblich ist, ob Ihr Korpus etwa aus vollständigen Texten oder aus Textauszügen besteht (s. Kriterium „Vollständigkeit der Texte“), brauchen Sie nicht, diese Information herauszufinden und in Ihrer Arbeit darüber zu berichten. Daneben geht der Leser tendenziell davon aus, dass Ihr Korpus monolingual ist (sprich ein Korpus des Deutschen) und dass es um ein Korpus der geschriebenen Sprache geht (weil es sich bei den meisten Korpora so verhält).
Die Fragen, die Sie hingegen immer – unabhängig von Ihrer Fragestellung – angehen sollen, sind: Wie heißt Ihr Korpus? Wo ist dies zu finden? Ist das ein Referenz- oder ein Spezialkorpus? (Wenn eine Textsorte stark überwiegt bzw. andere Textsorten komplett ausgeblendet sind, spricht dies in der Regel für eine Klassifikation als Spezialkorpus.) Was für Texte enthält das Korpus (z. B. Zeitungstexte, Texte aus einer älteren Sprachstufe)? Gibt es etwa einen thematischen Schwerpunkt (z. B. Sport, s. das Webkorpus *Ballsportarten*)? Welche Zeitspanne deckt das Korpus ab? Haben Sie linguistische Annotationen verwendet, müssen Sie auf jeden Fall erwähnen, dass es um ein annotiertes Korpus geht.

2.2 Ein Korpus auswählen

Man sagt, das Leben sei kein Ponyhof: Dasselbe gilt auch für die Korpuslandschaft. Das Korpus zu finden, das hundertprozentig zu der eigenen Fragestellung passt, ist einfach eine Utopie. Die Korpuslinguistik ist nämlich ein sehr junges Fach, was bedeutet, dass die Vielfalt der bisher erstellten Korpora noch begrenzt ist. Sie müssen Kompromisse eingehen: Nicht alle oben beschriebenen Korpusmerkmale haben in der Praxis das gleiche Gewicht; man priorisiert stattdessen das Merkmal oder die Merkmale, die für die Fragestellung entscheidend sind, z. B.:

- Kriterium „zeitlicher Bezug“ – Möchten Sie etwa Aussagen über den Sprachgebrauch in einer bestimmten Sprachstufe treffen, muss das Korpus zuerst diese Sprachstufe abdecken.
- Kriterium „Aufbereitung des Korpus“ – Brauchen Sie für die Operationalisierung Ihrer Fragestellung linguistische Annotationen, um die hohe Anzahl der Fehltreffer zu senken (s. nochmal das Beispiel von *furchtbar* als Gradpartikel in Abb. 1), muss das Korpus auf jeden Fall annotiert sein.
- Kriterium „Geltungsbereich des Korpus“ – Möchten Sie sich einer bestimmten Textsorte (z. B. Zeitungstexten) widmen, muss das Korpus diese Textsorte abbilden.

Dulcis in fundo ist aber ein Faktor unabhängig von der Fragestellung der entscheidendste: die Datenverfügbarkeit. Sie brauchen als Allererstes ein Korpus, das Daten über Ihren Untersuchungsgegenstand liefern kann; denn ohne (genügend) Daten können Sie keine empirische Antwort auf Ihre Fragestellung geben. Diesbezüglich spielt vor allem eine Rolle, wie häufig das Sprachphänomen, das man untersuchen möchte, generell im Sprachgebrauch vorkommt. Untersuchen Sie ein seltenes Sprachphänomen (z. B. das *gehören*-Passiv wie in *Die Täter gehören bestraft*), müssen Sie notwendigerweise zu einem sehr großen Korpus greifen, um eine adäquate Anzahl an Korpusbelegen für Ihre Analyse zu haben. Hinzu kommt, dass Korpora „noisy“ sind:

> “One thing to bear in mind when dealing with web data is that it can be rather ‘noisy’. For instance, it may lack punctuation or whitespace and there may be a higher proportion of non-standard spellings than in conventional texts.” (Kehoe 2020: 339)

Korpora umfassen authentische Sprache und somit auch Äußerungen, wo die Sprachproduzenten bspw. mehr oder weniger von der Rechtschreibung abweichen, was dazu führen kann, dass diese Korpusbelege für uns Sprachwissenschaftler wenig verständlich sind. Ist ein Korpusbeleg nicht verständlich genug (um etwa die Lesart eines Suchwortes X zu erfassen), ist dieser unbrauchbar. Nun finden sich in Korpora sehr viele Belege solcherart. Das bedeutet, Ihr Korpus muss mehr Daten liefern können, als Sie tatsächlich brauchen. Möchten Sie etwa 200 gute Korpusbelege, um Aussagen über Ihr Sprachphänomen zu treffen, darf dieses Korpus nicht nur 200 Belege zu diesem Sprachphänomen enthalten, denn viele davon werden mit Sicherheit unbrauchbar sein: Entweder einige Korpusbelege sind nicht klar genug, um ausgewertet zu werden, oder es handelt sich um Fehltreffer (d. h., es sind Korpusbelege, die etwa Ihr Suchwort enthalten, aber nicht Ihren Untersuchungsgegenstand erfassen; s. nochmal das Beispiel von *furchtbar* als Gradpartikel in Abb. 1).

3. Zusammenfassung

Um zusammenzufassen, sollten wir jetzt in der Lage sein, Korpora nach mehreren Kriterien zu beschreiben sowie diese Kriterien in Korpusbeschreibungen, die in empirischen korpusbasierten Studien vorkommen, wiederzuerkennen. Informieren Sie den Leser über die Eigenschaften Ihres Korpus kurz und bündig, indem Sie nur auf die für Ihre Fragestellung relevanten Korpuseigenschaften eingehen: Wer liest, muss sozusagen ein Gefühl für Ihr Korpus bekommen als auch den Eindruck haben, dass Sie nicht nach Zufall Ihr Korpus ausgewählt haben. Schließlich kennen wir jetzt auch einige wichtigen Überlegungen, was den Prozess der Auswahl eines Korpus anbetrifft (z. B. die allgemeine Auftretenshäufigkeit des zu untersuchenden Sprachphänomens im Sprachgebrauch).
Möchten Sie mehr über die deutsche Korpuslandschaft erfahren (d. h. welche Korpora es für das Deutsche überhaupt gibt), bietet sich bspw. die Lektüre von Kap. 7 aus der Einführung in die Korpuslinguistik von Lemnitzer & Zinsmeister (2015: 136–156) an: Dort finden Sie u. a. eine Übersicht über zahlreiche Portale und Archive. Sehr benutzerfreundlich sind v. a. die DWDS-Korpora (https://www.dwds.de/r), die mehrfach genau aus diesem Grund als Beispiele herangezogen wurden.

4. Literatur

Kehoe, Andrew. 2020. Web Corpora. In: Paquot, Magali & Stefan Th. Gries. (eds.), *A practical handbook of corpus linguistics*, 329–351. Berlin & New York: Springer.

Lemnitzer, Lothar & Heike Zinsmeister. 2015. *Korpuslinguistik: Eine Einführung*. 3. Auflage. Tübingen: Narr Francke Attempto.

Scherer, Carmen. 2014. *Korpuslinguistik*. 2., aktualisierte Auflage. Heidelberg: Winter.

Stathi, Katerina. 2012. *Selbst* vs. *persönlich* im deutschen Sprachgebrauch: eine Korpusanalyse. In: König, Ekkehard, Gunter Gebauer & Jörg Volbers (Hrsg.), *Selbstreflexionen: Performative Perspektiven*. München: Fink, 59–72.

5. Aufgaben

a) Warum ist es wichtig, sich mit den Kriterien der Korpustypologie vertraut zu machen?
b) Was ist der Unterschied zwischen Monitorkorpora und statischen Korpora?
c) Was ist ein Referenzkorpus? Was ist ein Spezialkorpus?
d) Welche Informationen über das Korpus sollte eine Korpusbeschreibung unabhängig von der Fragestellung enthalten?
e) Welcher Faktor ist der entscheidendste bei der Auswahl eines Korpus?
f) Nehmen wir an, eine empirische Studie hat Substantive mit dem Erstglied *Corona-* (z. B. *Corona-Krise*, *Corona-Zeit*) untersucht, und zwar anhand von Korpusdaten. Es folgt hier eine Beschreibung des Korpus. Welche Informationsarten bzw. Kriterien von Scherer (2014) kommen im Text vor?

   Als Datenquelle diente das Corona-Korpus (https://www.dwds.de/d/korpora/corona): ein Spezialkorpus aus Texten deutscher Webseiten, die sich thematisch mit der COVID-19-Pandemie beschäftigen (Zeitraum: 2018/01 bis 2020/12). Das Korpus ist morphosyntaktisch mit dem Stuttgart-Tubingen-Tagset annotiert. Zum Zeitpunkt der Datenerhebung (Datum: 06.06.2023) umfasste das Korpus ca. fünfzig Millionen Tokens (50 017 373) aus über siebzig Tausend Dokumenten (71 282). Diese stammten insgesamt aus 215 Quellen (z. B. Tages- und Wochenzeitungen, Magazinen, Fachpublikationen, Blogs).