„Was ist ein Korpus?“ – Das linguistische Korpus

1. (*Das*) *Korpus* im weiten und engen Sinne

Suchen wir nach dem Lexem *Korpus* (n.)[1] in einem Wörterbuch – etwa im Digitalen Wörterbuch der deutschen Sprache (DWDS) –, finden wir heraus, dass bei diesem Begriff eine weite und eine enge Lesart vorliegen. Im weiten Sinne bedeutet *Korpus*, wie das folgende Beispiel aus dem DWDS zeigt, einfach ‚**Sammlung**‘:

> *„[…] er erzählt von einem gewaltigen Korpus abstrakter Gemälde, das der Öffentlichkeit auf Anweisung der Künstlerin verborgen blieb, bis es dann in den Achtzigerjahren geborgen wurde wie ein Schatz […].“* [Süddeutsche Zeitung, 27.06.2020]

Im engen Sinne hingegen heißt *Korpus* ‚**Sammlung von Texten**‘:

> *„Große literarische Korpora, die anders gar nicht mehr angemessen bearbeitet werden können, werden mit den Methoden des »Distant Reading« makroanalysiert, um diachrone Entwicklungen besser beschreiben zu können.“* [Die Welt, 14.03.2020]

Genau diese Lesart liegt nun der Begriff *Korpus* zugrunde, wenn er in der Sprachwissenschaft verwendet wird. Ist aber jede Textsammlung gleichzeitig auch ein **linguistisches Korpus** – d. h. eine Textsammlung, das Sprachwissenschaftler nutzen können, um Aussagen über den Sprachgebrauch zu treffen? Kann ich anhand einer beliebigen Textsammlung linguistische Fragen beantworten wie: Hatte das Lexem *Urlaub* im Mhd. dieselbe Bedeutung wie heute? Welche Wörter mit dem Letztglied *-heini* bilden Sprecher in der Online-Kommunikation (s. Abb. 1)? Anders als eine beliebige Textsammlung zeichnet sich ein linguistisches Korpus durch eine Reihe von Eigenschaften aus, die uns ermöglichen, dies als adäquate Quelle sprachlicher Daten zu nutzen, um solche Fragestellungen anzugehen.

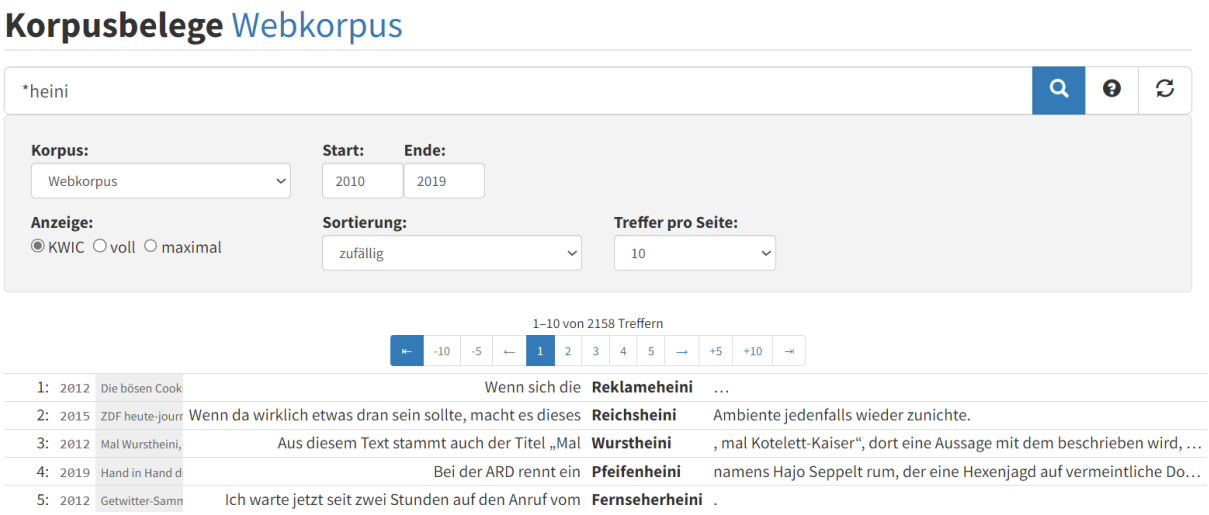

Abb. 1. Suche nach Wörtern mit dem Letztglied *-heini* im DWDS-*Webkorpus*

[1] Es heißt *das Korpus* im Singular und *die Korpora* im Plural.

2. Das linguistische Korpus

Vor allem drei Eigenschaften machen ein linguistisches Korpus aus: **Repräsentativität**, **Authentizität** und **(Korpus-)Größe** (Scherer 2014: 5–10; Stefanowitsch 2020: 22–38).

2.1 Repräsentativität

Wie oben angedeutet, ist das Ziel eines Sprachwissenschaftlers, Aussagen über den Sprachgebrauch zu treffen. Es ist aber naheliegend, dass wir eine Sprache nicht in ihrer Gesamtheit untersuchen können (sprich alle Millionen von Äußerungen/Texten, die auf Deutsch in einem bestimmten Zeitraum produziert wurden); wir müssen uns stattdessen mit einer Stichprobe – einer Auswahl an Texten – zufriedengeben: einem Korpus. Diese Auswahl an Texten darf aber nicht beliebig sein, da wir anhand dieser Auswahl, wie eben gesagt, Aussagen über den Sprachgebrauch treffen möchten. Sie muss die Grundgesamtheit, die uns interessiert (z. B. Deutsch der Gegenwart), abbilden: Sie muss **repräsentativ** sein. Der erste Unterschied zwischen linguistischen Korpora und beliebigen Textsammlungen ist also, dass linguistische Korpora in ihrer Zusammensetzung immer **zweckgebunden** sind: Sie wurden so zusammengestellt bzw. die Texte wurden so ausgewählt, dass das Korpus entweder eine Sprache insgesamt (Deutsch, Englisch usw.) oder eine sog. Varietät, d. h. einen „Ausschnitt“ einer Sprache (z. B. Zeitungssprache) repräsentieren soll. Glücklicherweise müssen wir für eine wissenschaftliche Arbeit kein Korpus selbst bilden, sondern fertige Korpora stehen uns zur Verfügung, bei deren Zusammenstellung andere Sprachwissenschaftler sich bemüht haben, die Repräsentativität zu gewährleisten. Repräsentativität stellt allerdings zugleich ein Ideal dar (Stefanowitsch 2020: 28–36). Letztendlich ist es sehr schwierig, zu ermitteln, in welchem Umfang etwa die unterschiedlichen Genres (Belletristik, Zeitungen etc.) im Sprachgebrauch vorkommen. Sicher ist aber, dass ein linguistisches Korpus den Sprachgebrauch getreuer als eine beliebige Textsammlung widerspiegelt, da diese ein reines Zufallsprodukt ist.

2.2 Authentizität

Man sagt außerdem, dass die Texte, die ein linguistisches Korpus bilden, **authentisch** sind. Was bedeutet aber Authentizität in diesem Zusammenhang? Die Texte eines Korpus sind authentisch, insofern sie für natürliche kommunikative Zwecke produziert wurden:

> „[*A*]*uthentic* language is language produced for the purpose of communication, not for linguistic analysis or even with the knowledge that it might be used for such a purpose.” (Stefanowitsch 2020: 23)

> „The texts were written or spoken […] for some authentic communicative purpose as opposed to, let’s collect those things for a corpus. […] For instance, journalese, newspaper language in corpora, would obviously meet this criterion, because journalists write articles to communicate something to the newspapers, and not because they know that that kind of stuff will later end up in a corpus, obviously.” (Gries 2020: 5)

Nehmen wir etwa das Korpus *Briefe von Jean Paul (1780–1825)*: Es ist offensichtlich, dass Jean Paul seine Briefe nicht geschrieben hat, damit zwei Jahrhunderte später ein Korpus mit seinem Namen zusammengestellt werden konnte und wir seinen Sprachgebrauch untersuchen. Ähnlich verhält es sich bei den allermeisten Korpora. Eine Ausnahme dazu können Korpora der gesprochenen Sprache darstellen, insofern die Sprachproduzenten in der Regel darüber informiert sind, dass sie

aufgenommen werden, spricht dass ihre Äußerungen Gegenstand von linguistischen Untersuchungen werden können. In diesem Fall würde ein sog. „**observer's paradox**“ vorliegen: “[W]e want to observe speakers interacting linguistically as they would if no linguist was in sight“ (Stefanowitsch 2020: 25).

2.3 Korpusgröße

Die Anzahl der Texte und Wörter, die ein Korpus bilden, kann von Korpus zu Korpus stark variieren. Es gibt sowohl Korpora, die „nur“ hunderte Millionen von Wörtern umfassen, als auch solche, die aus mehreren Milliarden bestehen. Allgemein gesprochen gilt: Je größer das Korpus, desto besser. Denn ein Korpus stellt, wie gesagt, eine Stichprobe von Texten dar und es ist naheliegend, dass die Repräsentativität des Korpus sich umso mehr verbessert, je mehr Texte darin enthalten sind. Nehmen wir z. B. das Korpus *Politische Reden (1982–2020)*: Je mehr politische Reden unterschiedlicher Politiker berücksichtigt wurden, umso genauer können die Rückschlüsse auf das Genre der politischen Reden sein. Wie groß muss aber ein Korpus sein?

> „There is no principled answer to the question “How large must a linguistic corpus be?”, except, perhaps, an honest “It is impossible to say” […]. The more modest answer is that it must be large enough to contain a sample of instances of the phenomenon under investigation that is large enough for analysis.” (Stefanowitsch 2020: 38)

Ein Korpus ist also groß genug, wenn es genug Sprachdaten für eine bestimmte Fragestellung liefern kann. Das heißt, je seltener ein Sprachphänomen ist, desto größere Korpora muss man in Erwägung ziehen. Ein Musterbeispiel stellt das anfangs erwähnte Letztglied *-heini* dar, das relativ selten im Sprachgebrauch vorkommt. Das Korpus *Blogs* umfasst über 200 000 Texte: Ist diese Textanzahl adäquat oder zu gering? Allgemein betrachtet ist diese Frage schwierig zu beantworten, aber im Bezug auf den Untersuchungsgegenstand *-heini* ist die genannte Korpusgröße definitiv zu klein (s. Abb. 2): Für die Zeitspanne 2000-2014 sind nämlich nur 31 Treffer hier vorhanden – was eine inadäquate Datenbasis ist, damit wir Aussagen über die Verwendung von *-heini* treffen können.

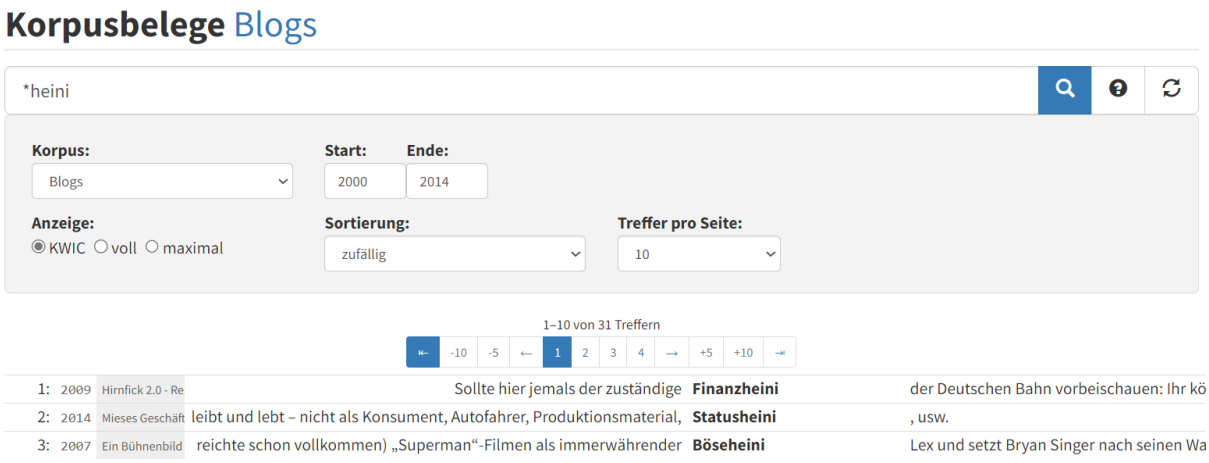

Abb. 2. Suche nach Wörtern mit dem Letztglied *-heini* im DWDS-Korpus *Blog*

3. Zusammenfassung

Fassen wir das bisher Gesagte zusammen, sollten wir jetzt in der Lage sein, eine vorläufige Definition für den Begriff *Korpus* im sprachwissenschaftlichen Bereich zu geben, z. B.: Ein Korpus ist eine Sammlung authentischer Texte, die eine Sprache insgesamt (z. B. Deutsch der Gegenwart) oder eine bestimmte Varietät (z. B. die deutsche Zeitungssprache) abbilden soll und im besten Fall genügend Texte enthält, um Aussagen über den Sprachgebrauch zu ermöglichen. Profis der Korpuslinguistik würden zustimmend nicken:

> „Ein Korpus ist eine Sammlung von Texten oder Textteilen, die bewusst nach bestimmten sprachwissenschaftlichen Kriterien ausgewählt und geordnet wurden.“ (Scherer 2014: 3)

> „In corpus linguistics, the term is used differently – it refers to a collection of samples of language use with the following properties: […] the instances of language use contained in it are authentic; […] the collection is representative of the language or language variety under investigation; […] the collection is large.” (Stefanowitsch 2020: 22f.)

4. Literatur

Gries, Stefan T. 2020. *Ten lectures on corpus linguistics with R: Applications for usage-based and psycholinguistic research*. Leiden/Boston: Brill. [Online verfügbar]

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

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press. [Online verfügbar]

5. Wiederholungsfragen

a) Welche Bedeutungen kann der Begriff *Korpus* haben?
b) Hinsichtlich welcher Eigenschaft unterscheidet sich ein linguistisches Korpus von einer beliebigen Textsammlung?
c) Inwiefern sind die Texte, die ein linguistisches Korpus bilden, authentisch?
d) Wann ist ein Korpus groß genug?
e) Wie lässt sich ein linguistisches Korpus definieren?
f) Kann man das Web als linguistisches Korpus betrachten?