Erschienen in: Bergmann, Rolf/Pauly, Peter/Stricker, Stefanie: Einführung in die deutsche Sprachwissenschaft. - Heidelberg: Universitätsverlag C. Winter, 2010, Fünfte, überarbeitete und erheblich erweiterte Auflage von Rolf Bergmann und Stefanie Stricker. S. 385-400.

Kapitel 28: Sprachdaten als Grundlage für die Sprachwissenschaft

28.1. Einstieg: Wozu braucht man Sprachdaten?

Gegenstand sprachwissenschaftlicher Untersuchung ist die Sprache, wie sie in Redeakten, Äußerungen, Zeitungsartikeln, Fachschriften und anderen Formen mündlicher oder schriftlicher Form verwendet wird. Sprachwissenschaftliche Untersuchungen sind also immer empirisch ausgerichtet: Sie stützen sich auf eine (mehr oder weniger) große Anzahl von Redeakten oder Texten. Je nach Erkenntnisinteresse beschreiben sie beispielsweise die darin vorliegende Sprachnorm und das zugrunde liegende System, wie anhand der folgende Aussage zum Online-Angebot 'grammis – Grammatik in Fragen und Antworten' (Institut für Deutsche Sprache, Mannheim) deutlich wird:

> Um das gewünschte Maß an Übereinstimmung mit dem Sprachhandeln kompetenter Sprachteilhaber zu erreichen, müssen sie [d.h. grammatische Regelformulierungen] auf einer empirischen Erforschung des tatsächlichen Sprachverhaltens der Mitglieder der Sprachgemeinschaft beruhen. In diesem Sinn stützen wir unsere Antworten, wo immer und wann immer dies möglich scheint, auf eine Auswertung einschlägiger Daten aus den Textkorpora des Instituts für Deutsche Sprache sowie weiterer maschinenlesbarer Textsammlungen. Jede Regel, die wir formulieren, muss mit den gefundenen Daten kompatibel sein oder, wenn die Datenlage nicht eindeutig ist, zumindest mit einer substantiellen Teilmenge der Daten.

Derartigen Untersuchungen gehen komplexe Arbeitsschritte voraus. Dazu gehören zunächst Beobachtungen an der Sprache, sodann die Entwicklung einer Fragestellung sowie Hypothesenbildung, die genaue und angemessene Begriffsbildung sowie eine konsequent anzuwendende Untersuchungsmethode.

Schließlich geht es vor allem auch darum, für die Recherche geeignete Sprachdaten zu sammeln bzw. zusammenzustellen. Solch eine Zusammenstellung kann für jede sprachwissenschaftliche Untersuchung eigenständig und neu erstellt werden. Da inzwischen aber viele Texte von verschiedenen Institutionen für einen breiteren Benutzerkreis in digitalisierter Form und für wissenschaftliche Untersuchungen zugänglich gemacht sind, wird häufig auf dieses Sprachmaterial zugegriffen. Zunehmend werden auch die vielen Texte im Internet als Recherchegrundlage verwendet. Eine aus solchen oder anderen Sprachdaten gezielt ausgewählte und strukturierte Zusammenstellung von (gesprochensprachlichen oder schriftsprachlichen) Texten wird als **Corpus** bezeichnet.

28.2. Corpora, Textarchive und Belegsammlungen

Nicht jede Zusammenstellung von Sprachmaterial ist jedoch ein Corpus. Als Grundlage für sprachwissenschaftliche Untersuchungen dienen auch Sammlungen von **Belegen** (d.h. von authentischen sprachlichen Äußerungen in Form kürzerer Textauszüge) und Archive digitalisierter Texte für nicht-sprachwissenschaftliche Zwecke (z.B. Texte eines Autors, einer literarischen Epoche). Im Folgenden sollen diese Sammlungen von Sprachdaten als Grundlage für sprachwissenschaftliche Untersuchungen vorgestellt werden.

28.2.1. Corpora

Spachwissenschaftliche Corpora sind meist sehr umfangreiche Sammlungen (vollständiger) Texte einer (manchmal auch mehrerer) Sprachen. Neben den Texten selbst sind im Corpus auch **Metadaten**, d.h. Informationen zu diesen Texten (z.B. Entstehungszeit, Publikationsort, Einordnung in ein Sachgebiet) enthalten. Das Corpus kann auch linguistisch annotiert sein, d.h. Informationen zur Wortart der enthaltenen Wörter und ihren Grundformen, Bestimmung von Konstituenten in Nominalphrasen, Markierung von Eigennamen oder Fehlern usw. anbieten.

Ein Corpus wird aber vor allem mit der Grundidee zusammengestellt, dass die in ihm enthaltenen Sprachausschnitte repräsentativ für die Sprache sind. Da die Grundgesamtheit (die Sprache) aber nicht vollständig und genau als Untersuchungsgegenstand zu bestimmen ist (sie verändert sich täglich), können Aussagen über die Stichproben (die Corpustexte) nur mit Vorsicht verallgemeinert werden. Je größer, aber auch je ausgewogener ein Corpus hinsichtlich der enthaltenen Textsorten ist, desto verlässlicher werden quantitative Aussagen, aber auch qualitative Aussagen über sprachliche Phänomene.

28.2.1.1. Schriftsprachliche Corpora

Digitalisierte Textcorpora, die für eigene sprachwissenschaftliche Untersuchungen genutzt werden können, werden von verschiedenen Institutionen angeboten. Ein für sprachwissenschaftliche Fragestellungen besonders geeignetes Corpus ist das am 'Institut für Deutsche Sprache' in Mannheim erarbeitete 'Deutsche Referenzkorpus DeReKo', die „weltweit größte Sammlung deutschsprachiger Korpora als empirische Basis für die linguistische Forschung". Es ist mithilfe des Corpusanalyse- und -recherchetools COSMAS II (= Corpus Search, Management, and Analysis System) zu untersuchen und besteht aus mehreren Einzelcorpora, die zeitlich und regional sowie hinsichtlich der Textsorten differenziert

sind. Derzeit umfasst das Corpus 3,9 Milliarden Textwörter; das entspricht über 9 Millionen Buchseiten, wenn durchschnittlich 400 Wörter pro Seite zugrundegelegt werden. Über Art und Zusammensetzung des Corpus und die reichen Suchmöglichkeiten informiert der Programmbereich Korpuslinguistik des Instituts. Die Mannheimer Korpora sind nach Registrierung als Nutzer über eine Webschnittstelle[1] oder durch kostenlosen Download einer Applikation für Windows-Oberflächen[2] zu benutzen. Eine umfangreiche Online-Hilfe mit zahlreichen Beispielen für Suchanfragen führt in die Benutzung ein.

Das 'Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts' (DWDS) verfügt über ein Kerncorpus mit 100 Millionen Textwörtern in knapp 80.000 Dokumenten aus dem gesamten 20. Jahrhundert sowie über verschiedene Zeitungs- und Sondercorpora[3]. Hiervon ist besonders das (allerdings nur intern nutzbare) DWDS-Ergänzungscorpus mit ca. 1 Milliarde Textwörtern aus Zeitungstexten zwischen 1990 und 2000 zu erwähnen. Das DWDS-Kerncorpus umfasst Texte aus den Bereichen Zeitung, Belletristik, Wissenschaft, Gebrauchsliteratur und gesprochene Sprache, die nach Textsorten geordnet chronologisch beschrieben werden. Die DWDS-Korpora sind vollständig mit Metadaten und linguistischer Annotation aufbereitet, so dass in ihnen z.B. mit einfachen Suchen, Lemmasuchen oder Suchen nach Wortarten recherchiert werden kann.

Für den weiter unten (Kapitel 29.1. und 29.2.3.) behandelten Wandel des Verbs *winken* von der schwachen Flexion *gewinkt* zur starken Flexion *gewunken* kann beispielhaft das Textcorpus des DWDS nach Auftreten, Häufigkeit und zeitlicher Streuung der jeweiligen Verbformen befragt werden. Dabei zeigt sich, dass die starke Partizipialform *gewunken* 14mal belegt ist, wovon allerdings nur 11 Treffer aus nutzungsrechlichen Gründen angezeigt werden, und zwar in Texten aus den Jahren 1907 (6 Belege), 1917, 1927, 1935, 1953 und 1994 (man vergleiche die Treffer-Anzeige zu *gewunken* in Abbildung 1).

---

[1] http://www.ids-mannheim.de/cosmas2/web-app/

[2] http://www.ids-mannheim.de/cosmas2/win-app/

[3] http://www.dwds.de

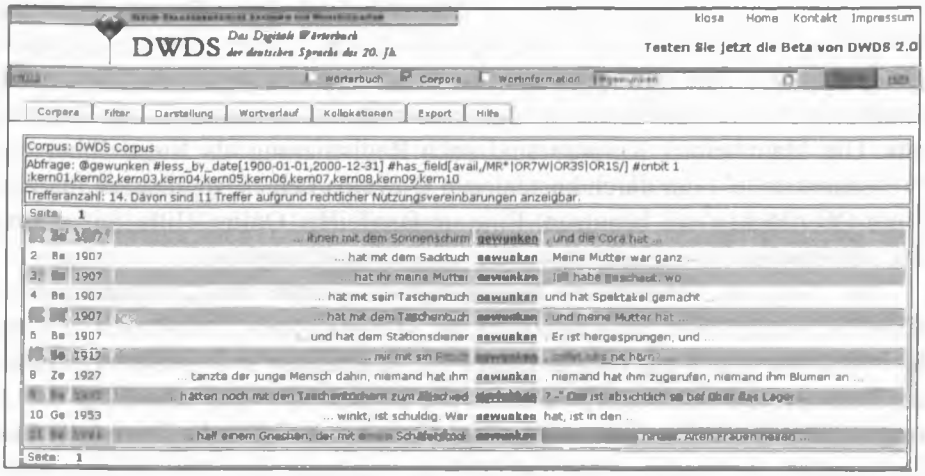

Abb. 1: 'Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts' (DWDS): Treffer-Anzeige zu *gewunken* im Kerncorpus

Die schwache Flexionsform *gewinkt* ist 51mal belegt, wovon allerdings nur 30 Treffer aus nutzungsrechlichen Gründen angezeigt werden, und zwar in Texten von 1902, 1905, 1907, 1910 (2 Belege), 1912, 1915, 1916, 1917, 1918, 1920 (2 Belege), 1922, 1925, 1926, 1927, 1928, 1929, 1931, 1932, 1940, 1946, 1951, 1957, 1963, 1970 (2 Belege), 1986, 1996 und 1998. Für einen schnelleren Überblick über die zeitliche Verteilung der Wortform in Texten des 20. Jahrhunderts bietet sich ein Blick in so genannte 'Verlaufsstatistiken im DWDS-Kerncorpus' an. Die Recherche bezeugt einerseits eine gleichmäßige Verwendung der schwachen Form über das ganze 20. Jahrhundert, andererseits ist die starke, hochsprachlich als nicht korrekt geltende Form ebenfalls im ganzen 20. Jahrhundert gleichmäßig belegt. Die in Kapitel 29.2.3. zu diesem Verb dargestellten Angaben aus grammatischen und lexikografischen Werken decken sich also mit dem Befund in diesem Textcorpus.

28.2.1.2. Gesprochensprachliche Corpora

Zwar liegt der Schwerpunkt bei der Erstellung von Corpora zum Deutschen generell auf geschriebener Sprache, doch gibt es auch einige Ressourcen zur gesprochenen Sprache. So ist im 'Digitalen Wörterbuch der deutschen Sprachen des 20. Jahrhunderts' (DWDS) beispielsweise ein 'Corpus Gesprochene Sprache' enthalten, dass Transkripte (also Verschriftlichungen der gesprochenen Äußerungen) von Reden, Rundfunkansprachen, Interviews, Talkshows usw. aus

dem gesamten 20. Jahrhundert im Umfang von ca. 2,5 Millionen Wortformen enthält.

In der 'Datenbank gesprochenes Deutsch' (DGD, Teil des 'Archivs für gesprochenes Deutsch') des 'Instituts für Deutsche Sprache' in Mannheim können zu den Treffern einer Suchanfrage in den Transkripten verschiedener gesprochensprachlicher Teilcorpora Tondateien mit den gesprochenen Äußerungen geöffnet werden. Dies ist z.B. dann interessant, wenn man untersuchen möchte, ob eine Partikel wie *halt* betont oder unbetont verwendet wird. Aus den Transkripten (wie in Abbildung 2 gezeigt) lässt sich die Bedeutung dieser Partikel erkennen (sie wird unter anderem verwendet, um zu betonen, dass an einer Tatsache nichts geändert werden kann, vgl. beispielsweise Ausschnitte 1 und 12). Durch Abspielen der Tondateien wird deutlich, dass diese Partikel immer unbetont verwendet wird.

Die Suche nach "halt" [KOSMAS-Recherche: STR('halt')]
führte zu 15 Fundstellen in Transkripten ausgewählter Korpora

Liste der Treffer Nr. 1 - 15

Die **Kennung** führt zur Dokumentation und T zum Transkript der Interaktion. W gibt den Ausschnitt der Tonaufnahme im WMA- und M im MP3-Format wieder

1 FR023 T W/M ..... in dieser Arbeit nich wahr es erfordert ja **halt** nun mal den den den Gesamteinsatz einer .....

2 FR023 T W/M ... allgemeine Temperamentsfrage was ich bewege mich **halt** immer sehr viel daß man einfach ein bißchen ...

3 FR023 T W/M .... haben sie beantwortet muß man sich sehr bemühen **halt** sich also dauernd nun zu ordnen weil man .....

4 FR030 T W/M .... solchen Augenblick nun alles aufzugeben und das **halt** ich für zu gefährlich
S3: und da meine ...

5 OS041 T W/M ..... das Feld hatten wir sehr weit. Und da bin ich **halt** mit dem bissel Essen, was ich so konnte .....

6 OS041 T W/M ..... hören, die hat mir's vorgesagt und da habe ich **halt** so meine Schularbeit meist auf dem Felde .....

7 OS041 T W/M ..... sonnabends, da war er daheim. Und da ging's **halt** bis in die Nacht, da mußte ich schon das .....

8 OS041 T W/M ..... waren dazumal noch alle klein, und da habe ich **halt** auch müssen so machen. Mein Mann hat sich ....

9 OS041 T W/M .... habe auch immer dreimal gemolken, denn ich habe **halt** gedacht, wenn das Kü wenn die Kühe auf dem ...

10 PF123 T W/M .... ganz nett und ganz interessant.Und dann geh ich **halt** schrecklich gern auf den Wochenmarkt.Und .....

11 PF123 T W/M ..... ungeputzt gegessen hab Und dann bekommt man da **halt** manchmal die Ware nicht so schön wie bei .....

12 PF123 T W/M ..... hab ich noch gleich 'n Essen.Und dann geh ich **halt** auch gern in die Geschäfte, nat wahr?Also ....

13 PF123 T W/M ..... gekauft hab, wieder zu kriegen.Dann bin ich **halt** raus und hab sie woanders gekauft.Woanders ...

14 PF123 T W/M ..... was sieht man da noch?Ja, und dann guckt man **halt** zu, wie die Elektrokarren aneinanderfahren, ...

15 PF180 T W/M ..... da kennt einer 'n andern scho nat.Und das ist **halt** da drauß ist ganz fürs ander.Da, wenn ganz ....

Abb. 2: Trefferanzeige zur Suche nach der Partikel *halt* in der 'Datenbank gesprochenes Deutsch' (DGD)

28.2.1.3. Zusammenfassung

Einen Überblick zu deutschsprachigen Corpora (geschrieben- und gesprochensprachlich) bieten L. Lemnitzer und H. Zinsmeister im Kapitel 'Deutsche Korpuslandschaft'[4]. In die Arbeit mit dem 'Deutschen Referenzkorpus' (DeReKo) und dem 'Digitalen Wörterbuch der deutschen Sprache des 20. Jahrhunderts' (DWDS) führt C. Scherer im Kapitel 'Arbeiten mit bestehenden Korpora'[5] ein. In beiden Bänden wird auch die Frage diskutiert, ob das World Wide Web ein Corpus ist. Tatsächlich bietet das World Wide Web eine riesige Menge an authentischen Texten, die prinzipiell auch für sprachwissenschaftliche Untersuchungen herangezogen werden können – allerdings nur mit Vorbehalt. So ist es z.B. nicht einfach, nur die deutschsprachigen Texte zu finden, häufig fehlen Metadaten zu den Texten (wer hat sie wann und wo geschrieben?) und Texte wiederholen sich. Sowohl für verlässliche qualitative wie quantitative Aussagen ist das World Wide Web daher nur bedingt geeignet. Das World Wide Web ist deshalb, aber auch, weil es nicht gezielt zum Zweck sprachwissenschaftlicher Untersuchungen zusammengestellt wurde, kein richtiges Corpus.

28.2.2. Textarchive

Sowohl im Internet wie auf elektronischen Datenträgern (z.B. CD-ROM) stehen Sammlungen digitalisierter, meist literarischer Texte zur Verfügung, die vornehmlich zum Zweck der Archivierung und besseren Verfügbarkeit angelegt wurden. Solche **Textarchive** auf CD-ROM umfassen große Einzelwerke der Weltliteratur wie die Bibel oder vollständige Textausgaben eines Autors. So gibt es beispielsweise die Werke Johann Wolfgang von Goethes oder Friedrich Schillers auf CD-ROM. Erwähnenswert ist insbesondere das Projekt 'Digitale Bibliothek'[6], das Anthologien zu zahlreichen literarischen Epochen und Gattungen, Werkausgaben zahlreicher deutsch- und anderssprachiger Autoren, allgemeine Nachschlagewerke und Anthologien sowie Texte zu verschiedenen Fachgebieten auf CD-ROM anbietet. Mit dem Projekt 'Gutenberg'[7] steht im Internet eine umfangreiche Sammlung digitalisierter Texte der Weltliteratur zur Verfügung.

Solche Textarchive können für verschiedene sprachwissenschaftliche Fragestellungen als Untersuchungsgrundlage dienen, z.B. wenn es um den Wortschatz eines Autors oder grammatische Entwicklungen in einem Sprachstadium

---

[4] in: Korpuslinguistik. Eine Einführung, S. 107-126

[5] in: Korpuslinguistik, S. 74-91

[6] http://www.digitale-bibliothek.de/

[7] http://www.gutenberg.org bzw. für das Deutsche http://gutenberg.spiegel.de/

geht. Hierzu sind die Textarchive, ähnlich wie Corpora, meist mit einem Recherchesystem ausgestattet und für dieses auch entsprechend aufbereitet. Diese Systeme bieten beispielsweise alphabetische **Lemmatisierungen** der Textwörter, Angaben zur Frequenz der Wortformen, Nennungen der Quelle und genaue Stelle, Wort- und Syntagmensuche mit Einsatz von Stellvertretern für bestimmte Wortbestandteile (z.B. *-ung*) oder variable Buchstabenfolgen (z.B. *r-f-* für *rief*, *riefen*, *rufte* etc.) sowie eine individuelle Bestimmung der Beleglänge. Die ermittelten Daten sind, wie bei Corpora, speicherbar und in Textverarbeitungsprogramme kopierbar, so dass ein zeitaufwändiges und fehleranfälliges Abschreiben der Texte entfällt.

Möglichkeiten der Recherche können erneut an einem Beispiel aus dem Zusammenhang der Verbflexion veranschaulicht werden. Als Beispiel soll das Verb *rufen* dienen, das bis in neuhochdeutsche Zeit hinein stark und schwach flektiert worden ist. Es bestand also ein Nebeneinander von *rief*, *gerufen* und *rufte*, *geruft*. Im 19. Jahrhundert ist die schwache Flexion dann aufgegeben worden. Das bei anderen Verben gegenwartssprachlich existierende Normproblem, das aus dem Nebeneinander unterschiedlicher Flexionsformen resultiert (man vergleiche Kapitel 23 und 29.2.3.), ist bei *rufen* heute nicht mehr gegeben, da nur noch die starken Flexionsformen gebräuchlich sind. Der relevante Zeitraum, in dem sich der Wandel in der Flexion von *rufen* vollzogen hat, ist das 18./19. Jahrhundert. Will man die Verhältnisse bei J.W. von Goethe aufdecken, so erweist sich die CD-ROM-Version der Weimarer Goethe-Ausgabe als gutes Hilfsmittel. Das Verb *rufen* ist dort mit über 1.600 Belegen erfasst und kann damit als hochfrequent bezeichnet werden. Tabelle 1 veranschaulicht die Belegfrequenz der stark flektierten Verbformen, die zusammen 1.381 Belege ausmachen:

Wortform	**Frequenz**
rief	1.248
rief's	5
riefe	8
riefen	104
riefest	1
riefst	9
rieft	6

Tabelle 1: Belegfrequenz stark flektierter Verbformen von *rufen* in 'Goethes Werke auf CD-ROM'. Weimarer Ausgabe. Chadwyck / Healey, Cambridge 1996

Es fragt sich, ob und gegebenenfalls in welchem Umfang in den Goethe-Texten daneben schwache Flexionsformen auftreten. Die Recherche führt zur Erhebung von 14 schwach flektierten Verbformen (*rufte* 13 Belege, *ruften* 1 Beleg). Tabelle 2 zeigt das Ergebnis der Recherche des flektierten Wortes *rufte* in der Weimarer Ausgabe. Aufgelistet sind Abteilung und Band der Ausgabe, der Titel des Werke, in dem die Wortform steht, und die jeweilige Trefferzahl.

Abtlg.,Bd.	**Titel**	**Treffer**
I,13ii	Prolog zu dem Schauspiel Der Krieg, von Goldoni. Gesprochen von Madame Becker, geb. Neumann. Den 15. October 1793. [Lesarten]	1
I,19	Die Leiden des jungen Werther. [Apparat]	7
I,51	Wilhelm Meisters theatralische Sendung [Buch 1 – Buch 3]	2
I,51	Wilhelm Meisters theatralische Sendung [Buch 1 – Buch 3], [Lesarten]	2
V,4	Gespräch mit Dutitre, Frau: [undatiert]	1

Tabelle 2: Nachweis der schwach flektierten Verbformen von *rufen* in 'Goethes Werke auf CD-ROM'. Weimarer Ausgabe. Chadwyck / Healey, Cambridge 1996

Tabelle 3 zeigt die beiden *rufte*-Belege aus 'Wilhelm Meisters theatralische Sendung' in Form eines kurzen Kontextes, der etwa eine Zeile umfasst. Zudem wird die Seite genannt, auf der sich der Beleg findet. Oft reicht ein kurzer Kontext aus, um festzustellen, ob der Beleg für die jeweilige Fragestellung relevant ist und – eventuell mit einem längeren Kontext – aufgenommen werden sollte oder ob er auszuschließen ist.

Abtlg./Bd.	**Titel/Kontext**
I,51	Wilhelm Meisters theatralische Sendung [Buch 1 – Buch 3]. Buch 1, Capitel 5, S. 16 *ihn seine Mutter manchmal herein* ***rufte****, um ihr etwas heraus tragen*
I,51	Wilhelm Meisters theatralische Sendung [Buch 1 – Buch 3]. Buch 3, Capitel 14, S. 278 *war, faßte sich zusammen und* ***rufte*** *die ersten Versen seiner Rolle*

Tabelle 3: Kontexte der *rufte*-Belege in 'Goethes Werke auf CD-ROM'. Weimarer Ausgabe. Chadwyck / Healey, Cambridge 1996

Zu dem zweiten Beleg lautet der vollständige Satz: *Die Symphonie des Stückes ging an, und sein Geist, der aus einer Leidenschaft in die andere geworfen war, faßte sich zusammen und* ***rufte*** *die ersten Verse seiner Rolle dem Gedächtnisse hervor.* (Wilhelm Meisters theatralische Sendung [Buch 1 – Buch 3]. Buch 3. Capitel 14, S. 278, Z. 24-28).

Für die Frage der Ablösung der schwachen Flexion durch die starke ist auch der Beleg für *ruften* bemerkenswert: *bald wanndt' ich mich hierher zu meiner Mutter, und lebte still, biß sie die Götter* ***ruften*** *(übergeschrieben riefen, Herder) bey ihr.* (WA. I, 11, S. 382, 14-16: Elpenor. Paralipomena). J.G. Herder hat bei seiner kritischen Textdurchsicht die schwache Verbform *ruften* durch die starke Form *riefen* ersetzt.

Um ein Phänomen wie die Schwankung zwischen starker und schwacher Verbflexion zu untersuchen, sind alle flektierten Formen eines Verbs zu erfassen. Bei dem Beispiel *rufen* zeigt sich, wie gering der Anteil schwach flektierter Formen im Vergleich zu den starken Formen in den Goethe-Texten ist. Erst bei einer vollständigen Erhebung ist ein solcher Befund aussagekräftig und erlaubt Feststellungen über die genaue Frequenz einer Erscheinung und damit über den Stand der Entwicklung bei dem jeweiligen Autor oder in dem ausgewählten Zeitraum.

28.2.3. Belegsammlungen

Insbesondere zu lexikografischen Zwecken (man vergleiche Kapitel 27), aber auch für sprachwissenschaftliche Untersuchungen allgemein werden **Belegsammlungen** angelegt. Solche Sammlungen können aus einem für eine Untersuchung verwendeten Corpus extrahiert werden, sie können aber auch aus anderen Quellen (z.B. allen Werken eines Autors) ermittelt werden.

Eine umfangreiche Belegsammlung zum historischen deutschen Wortschatz ab 1450 besitzen beispielsweise die Arbeitsstellen für die Neubearbeitung des 'Deutschen Wörterbuchs' von Jacob und Wilhelm Grimm in Berlin und Göttingen. Für die Alphabetteile *A* bis *F* liegen dort circa 5 Millionen Belegzettel[8] vor (man vergleiche die Abbildung und Erläuterung eines solchen Belegzettels in Kapitel 27.8.).

Auch andere Wörterbuchunternehmen haben große Belegsammlungen, die Benutzern zugänglich gemacht werden. Beispielsweise verfügt das Goethe-Wörterbuch (mit Arbeitsstellen in Berlin, Hamburg und Tübingen) über ein Belegzettelarchiv, das die gut 90.000 Wörter der Sprache Goethes in über 3,3 Mil-

[8] http://grimm.adw-goettingen.gwdg.de/

lionen Belegen repräsentiert. Informationen über deutschsprachige Wörterbücher, ihre Belegsammlungen und Zugriffsmöglichkeiten enthält eine von M. Schlaefer herausgegebene Broschüre (Deutschsprachige Wörterbücher. Projekte an Akademien, Universitäten, Instituten).

Darüber hinaus gibt es an etlichen Forschungsstellen und Instituten verschiedener Universitäten spezielle Belegsammlungen, die aus Forschungsunternehmen hervorgegangen sind und die Interessierte in der Regel auf Anfrage einsehen können. Für Untersuchungen zu neueren Entwicklungen des Lexikons kann z.B. die Sammlung 'Wortwarte' von L. Lemnitzer[9] zugrundegelegt werden, zur Untersuchung syntaktischer Phänomene beispielsweise die von M. Volk aufgebaute 'Grammatiktestumgebung'[10].

Einen guten Überblick über Belegsammlungen zur Sprachgeschichte des Deutschen mit einer Beschreibung der jeweiligen Sammlung und den Zugriffsmöglichkeiten gibt W. Hoffmann in seinem Aufsatz 'Probleme der Korpusbildung in der Sprachgeschichtsschreibung und Dokumentation vorhandener Korpora'[11]. Hinweise auf sprachwissenschaftliche Belegsammlungen zur Gegenwartssprache geben L. Lemnitzer/H. Zinsmeister.[12]

28.2.4. Materialgewinnung für eigene Untersuchungen

Für manche sprachwissenschaftliche Fragestellungen sind die hier beschriebenen Belegsammlungen, Textarchive oder Corpora nicht geeignet, so dass eine eigene Sammlung von Sprachdaten erstellt werden muss. Für die Zusammenstellung dieser Sammlung sind Kriterien wie Textmenge, zeitliche und räumliche Ausdehnung, Textsorten, Sprachschichten und Sprachvarietäten usw. zu berücksichtigen, damit das Material sich zur Beantwortung der jeweiligen Fragestellung eignet. Gegenstand und Ziel einer Untersuchung bestimmen also die Anlage eines Textcorpus oder einer Belegsammlung.

Will man beispielsweise ein Phänomen der Gegenwartssprache untersuchen, können nur gegenwartssprachliche Texte im Corpus enthalten oder in Form von Belegen verzettelt sein. Der Begriff Gegenwartssprache bedarf allerdings zunächst der Definition, wofür es in der Sprachwissenschaft divergierende Ansätze gibt. Das 'Wörterbuch der deutschen Gegenwartssprache' definierte 1964 seinen Objektbereich folgendermaßen: „Unter deutscher Gegenwartssprache wird außer der [...] heute geschriebenen und gesprochenen Sprache [...] auch

---

[9] http://www.wortwarte.de

[10] http://www.uni-koblenz.de/~compling/Forschung/Gtu/gtu.html

[11] in: Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung, 1. Teilband, S. 882-886

[12] in: Korpuslinguistik. Eine Einführung, S. 42-43

die Sprache der in unserer Zeit noch gelesenen lebendigen deutschen Literatur der Vergangenheit verstanden. Daher fußt das Wörterbuch zwar vornehmlich auf dem Wortschatz des 20. Jahrhunderts, zieht aber auch den der Literatur des 19. Jahrhunderts und in gewissem Umfang des letzten Drittels des 18. Jahrhunderts heran.“ Für das ‘Lexikon der Germanistischen Linguistik’ hingegen ist „es sinnvoll, die sprachliche Gegenwart im Jahr 1945 beginnen zu lassen“.

Ein Corpus, aber auch eine Belegsammlung soll verallgemeinernde Aussagen über die beschriebene Sprache erlauben. Es muss daher eine gewisse Repräsentativität für die untersuchte Sprache besitzen. Bei der deutschen Gegenwartssprache müssen also je nach Zielsetzung ihre geografischen Varianten, ihre Schichtung in Hochsprache und Umgangssprache, ihre Gliederung in Fach- und Sondersprachen und ihre Varietäten in den Textsorten (z.B. Zeitungstexte, Werbetexte, Kochrezepte) berücksichtigt werden.

Gegenstand und Ziel der Untersuchung bestimmen auch die Art und Weise, in der die einzelnen Vorkommen erfasst werden. Sind in Corpora vollständige Texte enthalten oder zumindest sehr lange Textausschnitte, umfassen Belege je nach der Zielsetzung gegebenenfalls Satzglieder, Teilsätze, Satzgefüge oder größere Textausschnitte. Belegsammlungen können entweder auf Papier (mit jeweils einem Beleg auf einer Karteikarte) oder elektronisch (mit jeweils einem Beleg pro Datensatz einer Datenbank) angelegt sein. Corpora stehen dagegen, wegen der großen Textmengen und aus Gründen der besseren Recherchierbarkeit, meist elektronisch zur Verfügung.

Sowohl Belegsammlungen wie Corpora enthalten neben den sprachlichen **Primärdaten** auch Metadaten. Auf jeder Karteikarte einer Belegsammlung bzw. in jedem Datensatz eines Beleges ist neben dem eigentlichen Belegtext die Belegstellenangabe (bestehend aus einer Kurzsigle für den Text, aus dem der Beleg entnommen wurde, und einer Seiten- und Zeilenangaben) enthalten. Metadaten eines Corpus umfassen Angaben dazu, ob es sich um einen gesprochenen oder geschriebenen Text handelt, wer den Text verfasst hat, wo er publiziert wurde, wann er entstanden ist, welchem Thema/Sachgebiet er zugeordnet werden kann usw. Diese Daten werden von den sprachlichen Primärdaten (den Texten) getrennt gespeichert, um in ihnen recherchieren zu können. So können beispielsweise alle Texte eines Corpus, die dem gleichen Thema zugeordnet werden, in einem Teilcorpus zusammengestellt werden, das als Untersuchungsgrundlage für den Wortschatz eines Sachgebietes dienen kann. Dass Corpora neben den Primärdaten auch Metadaten enthalten, unterscheidet sie im Übrigen von reinen Textarchiven.

28.3. Analysebeispiel: Kookkurrenzlisten als Basis für die Ermittlung von Synonymen

So vielfältig, wie die Erkenntnisinteressen der Sprachwissenschaft sind, so vielfältig werden Corpora erstellt und genutzt. Die oben in diesem Kapitel verwendeten Beispiele zeigen bereits, wie die Daten aus gesprochen- oder geschriebensprachlichen Corpora zur Grundlage sprachwissenschaftlicher Untersuchungen werden können. Ein Analysebeispiel aus dem Bereich der Lexikologie und Lexikografie soll diese Beispiele ergänzen.

Vorstellbar ist eine Situation, in der ein Verlag ein grundlegend neues Wörterbuch sinn- und sachverwandter Wörter erarbeiten möchte. Es könnte aber auch in einer lexikologischen Einzeluntersuchung um die Synonyme zu einem bestimmten Wort gehen. Hierfür wird ein Corpus aufgebaut, das mithilfe ausgereifter Korpusrecherche- und -analysetools untersucht werden kann, die nicht nur Suchen nach Wörtern und Wortformen erlauben, sondern auch den Kontext der Wörter einbeziehen und Informationen über **Kookkurrenzen** eines Wortes, also 'Mitspieler'-Wörter im Kontext, liefern. Für die Ermittlung der Synonyme eines Wortes wird von der Hypothese ausgegangen, dass die synonymen Wörter die gleichen Kookkurrenzen zeigen.

Ein Tool in der 'Kookkurrenzdatenbank CCDB' (Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. © 2001 – 2007 Institut für Deutsche Sprache, Mannheim) erlaubt es, die Kookkurrenzprofile verschiedener Wörter mit dem eines Suchwortes zu vergleichen. Aus dem in Abbildung 3 gezeigten Ausschnitt der verwandten Kookkurrenzprofile zu *Dorf* lassen sich Hinweise auf mögliche Synonyme wie *Städtchen*, *Örtchen* oder *Dörfchen* gewinnen. Allerdings ist für das Wörterbuch zu überlegen, ob Diminutiva eines Stichwortes (*Dörfchen* zu *Dorf*) tatsächlich als sinnverwandte Wörter aufgenommen werden sollen. Daneben scheinen in der Liste auch andere Relationen auf: Sind *Kleinstadt* oder *Gehöft* der Bezeichnung *Dorf* nebengeordnete Bezeichnungen unter einem Oberbegriff *Siedlung* oder *Ortschaft*? Sollen solche Relationen auch im Wörterbuch erfasst werden? Zu überlegen wäre daneben, ob Komposita wie *Bergdorf* oder *Fischerdorf* als Synonyme oder Hyponyme (also Unterbegriffe) zu *Dorf* im Wörterbuch erscheinen sollen.

Folgende verwandte Kookkurrenzprofile zu [redacted] wurden gefunden (anklickbar, absteigend nach Verwandtschaftsgrad sortiert):

Ortschaft
Dörfchen
Bergdorf
Städtchen
Kleinstadt
Gehöft
Örtchen
Fischerdorf
Heimatdorf
Landstrich
Bergregion
Provinzstadt
Kampfgebiet
Gegend
Anhöhe
entvölkern
Siedlung
Zivilist
Ort
Gebiet
Kaff

Abb. 3: Verwandte Kookkurrenzprofile zu *Dorf* in der Kookkurrenzdatenbank CCDB (Ausschnitt)

Generell gilt, dass für die Entscheidung der Frage, ob jedes dieser Wörter tatsächlich ein sinnverwandtes Wort ist und in welcher Relation es zum Stichwort *Dorf* steht, der einfache Blick in solche Listen nicht genügt. Unerlässlich ist die Überprüfung der Vorkommen in den Corpustexten, die erst klare Entscheidungen ermöglichen. So wird z.B. deutlich, dass *Ortschaft* und *Dorf* synonym zueinander verwendet werden können (vgl. Beispiel 1 aus dem 'Deutschen Referenzkorpus' DEREKO), aber auch in einer Beziehung der Über-/Unterordnung stehen (vgl. Beispiel 2 aus dem 'Deutschen Referenzkorpus' DEREKO).

Beispiel 1: Vom Flughafen gibt es einen Buspendelverkehr nach Saariselkä. Auf der Strecke hat man gute Chancen, seine ersten Rentiere zu sehen. Dann taucht das **Dorf** auf, hineingekauert in eine tief verschneite Postkartenlandschaft zwischen den Bergen Kaunispää und Kiilopää. Die **Ortschaft** wurde 1870 von Goldsuchern gegründet. (Die Zeit (Online-Ausgabe), 20.11.2003, Tanzen, bis der Hund kommt, S. 64.)

Beispiel 2: Angeklagt sind die beiden Brüder Zoran und Mirjan Kupreskic, ihr Cousin Vlatko Kupreskic sowie Drago Josipovic, Dragan Papic und Vladimir Santic. Sie sollen das **Dorf** Ahmici und andere **Ortschaften** überfallen haben. (Tiroler Tageszeitung, 18.08.1998, Sechs Kroaten vor Tribunal in Den Haag.)

Es ist daher zu überlegen, ob *Dorf* und *Ortschaft* als Synonyme im Wörterbuch zu verbuchen sind oder nicht. Die Nennung von Textbelegen aus dem Corpus kann in solchen Fällen dem Nutzer die Interpretation der Angaben erleichtern.

28.4. Zur Problematik corpusgesteuerter und corpusbasierter Arbeit

Der Einsatz von Corpora für sprachwissenschaftliche Untersuchungen (wie generell der empirische Ansatz, der eingangs postuliert wurde) ist nicht unumstritten. Er erscheint dann logisch, wenn linguistische Erkenntnis vom Sprachgebrauch und nicht von Sprecherurteilen ausgehen will. Ist für eine empirische sprachwissenschaftliche Untersuchung die Entscheidung, mit einem Corpus zu arbeiten, gefallen, also eine 'corpusgestützte' Untersuchung anzufertigen, stehen weitere Entscheidungen über den Einsatz verschiedener corpuslinguistischer Methoden an, die nicht einfach zu treffen sind.

Bei **corpusgestützten** Untersuchungen kommen nämlich zwei unterschiedliche Verfahren zum Einsatz: die **corpusgesteuerte** Methode und die **corpusbasierte** Methode. Beim corpusgesteuerten Verfahren benutzt ein Sprachwissenschaftler das seiner Untersuchung zugrundegelegte Corpus explorativ und exhaustiv, d.h. er befragt es ohne Vorannahme mithilfe verschiedener Korpusrecherche- und -analysetools. Die Ergebnisse analysiert, bewertet und beschreibt er dann. Die Beobachtung des Sprachgebrauchs bildet also für die jeweilige Untersuchung die wichtigste Quelle. Die Corpusdaten werden sowohl quantitativ wie qualitativ analysiert. Zum Beispiel könnte etwa eine Darstellung der Verbalflexion bei Goethe aufgrund der weiter oben genannten digitalen Version der Weimarer Ausgabe erarbeitet werden.

Wird das Corpus corpusbasiert ausgewertet, geht der Sprachwissenschaftler dagegen von einer bestimmten Annahme zu einem sprachlichen Phänomen aus und sucht im Corpus gezielt nach Belegen dafür. Das Corpus wird bei diesem Vorgehen als zusätzliche Quelle benutzt, um die Hypothesen zu einem sprachlichen Phänomen zu bestätigen oder zu widerlegen. Die Corpusdaten selbst werden dabei im Grunde weder quantitativ noch qualitativ ausgewertet. Für ein Beispiel kann auf Abschnitt 28.2.2. zurückverwiesen werden, wo Belege aus Goethes Werken für den vermuteten zeitweiligen Übergang des Verbs *rufen* zur schwachen Flexion herangezogen wurden.

Da beide Methoden Vor- und Nachteile haben, stellt sich für jede corpusgestützte sprachwissenschaftliche Untersuchung die Frage, welche Methode angemessen und daher zu bevorzugen ist.

28.5. Definitionen

Beleg	authentische sprachliche Äußerung in Form kürzerer Textauszüge, die als Basis für qualitative Analysen dient
Belegsammlung	Zusammenstellung authentischer sprachlicher Äußerungen als Grundlage für eine sprachwissenschaftliche Untersuchung
Corpus	gezielt ausgewählte und strukturierte Sammlung von Texten als Grundlage für eine sprachwissenschaftliche Untersuchung
corpusbasiert	an den Corpusdaten rückprüfend
corpusgesteuert	von den Corpusdaten ausgehend
corpusgestützt	auf der Basis eines Corpus
Kookurrenzen eines Wortes	Wörter, die als 'Mitspieler' mit anderen Wörtern in einem Kontext auftreten
Lemmatisierung	Zuordnung von Wortformen eines Textes zu ihren Grundformen (Lemmata)
Metadaten	Informationen zu den Primärdaten eines Corpus
Primärdaten	die Texte eines Corpus
Textarchiv	Sammlung digitalisierter Texte zum Zweck der Konservierung und Verfügbarmachung

28.6. Literaturhinweise

Kurzinformation

Metzler Lexikon Sprache. Artikel: Frequenz, Informant, Institut für deutsche Sprache (IdS), Korpus, Korpusanalyse, Lemmatisierung

Einführende Literatur:

L. *Lemnitzer*/H. *Zinsmeister*, Korpuslinguistk. Eine Einführung

S. *Lenz*, Korpuslinguistik

C. *Scherer*, Korpuslinguistik

Grundlegende und weiterführende Literatur:

Corpus Lingustics. Ein internationales Handbuch

D. *Biber* u.a., Corpus Linguistics. Investigating language structure and use

S. *Hunston*, Corpora in Applied Linguistics

E. *Tognini-Bonelli*, Corpus Linguistics at Work