2. Arten von Korpora

2.1 Merkmale zur Klassifizierung von Korpora

Zwar handelt es sich bei jedem Korpus um eine gezielt zusammengestellte Sammlung an Texten, dennoch gilt: Korpus ist nicht gleich Korpus. Da ein Korpus immer im Hinblick auf einen bestimmten Verwendungszweck erstellt wird, hat jedes Korpus eigene Charakteristika. Ein Korpus der geschriebenen Sprache unterscheidet sich von einem Korpus der gesprochenen Sprache, ein Korpus der Gegenwartssprache von einem Korpus des Mittelhochdeutschen, ein Korpus der Jugendsprache von einem Korpus des Standarddeutschen.

Anhand von formalen Kriterien lassen sich Korpora, die computerlesbar sind, unterscheiden von solchen, die es nicht sind, Gesamtkorpora von Teilkorpora, Probenkorpora von Volltextkorpora sowie unveränderliche, statische Korpora von nicht abgeschlossenen Monitorkorpora. Nach dem Sprachmedium kann man differenzieren zwischen Korpora der gesprochenen Sprache und Korpora der geschriebenen Sprache. Die zeitliche Nähe oder Distanz zur Gegenwart ist relevant für die Einteilung in Korpora der Gegenwartssprache und historische Korpora. Nach dem Verwendungszweck werden Referenzkorpora von Spezialkorpora unterschieden, und für die Einteilung in einsprachige und mehrsprachige Korpora ist die Zahl der im Korpus enthaltenen Sprachen ausschlaggebend. Abbildung 1 gibt einen Überblick über die Arten von Korpora. Sie werden im Folgenden anhand von deutschsprachigen Korpora ausführlicher beschrieben. Ein Verzeichnis der in diesem Buch behandelten Korpora findet sich im Informationsteil am Ende des Buches.

Die genannten Kriterien zur Klassifizierung von Korpora sind jedoch nicht erschöpfend. So sind für die Arbeit mit einem fremden Korpus etwa auch die Größe und Zugänglichkeit des Korpus relevant. Neben kleineren (bis 1 Mio. Textwörter), mittleren (mehrere Mio. Textwörter) und großen Korpora (über 100 Mio. Textwörter) existieren inzwischen auch sehr große Korpora (über eine Milliarde Textwörter). Korpora können vor Ort oder online frei zugänglich sein, manche Korpora kann man nach vorheriger Registrierung nutzen, bei anderen Korpora ist eine externe Nutzung aus verschiedenen Gründen nicht möglich.

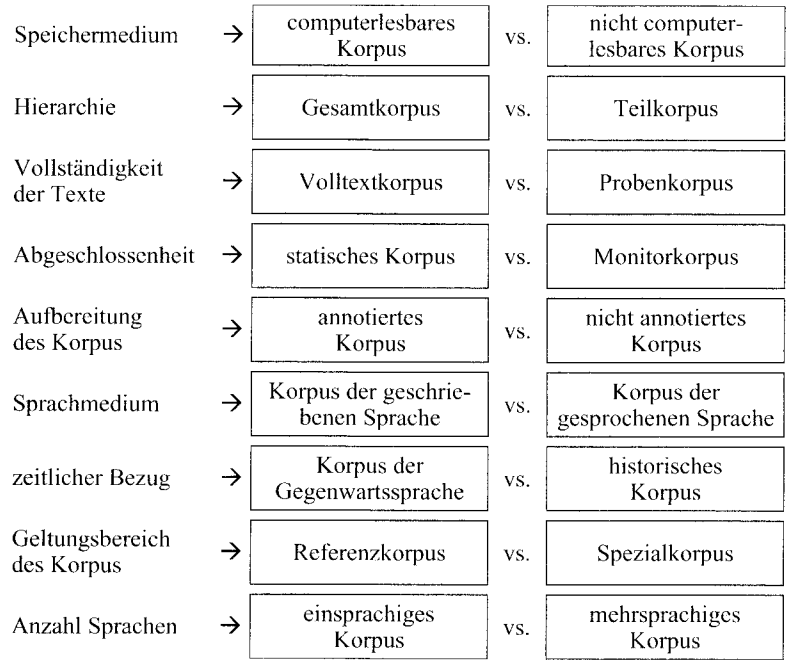

Abbildung 1: Arten von Korpora

2.2 Computerlesbare und nicht computerlesbare Korpora

In der Linguistik wird der Terminus "Korpus" traditionell als Oberbegriff für Sammlungen von Texten und Textteilen verwendet, die unter explizit sprachwissenschaftlichen Gesichtspunkten zusammengestellt wurden, um eine bestimmte sprachliche Gesamtheit abzubilden. In jüngerer Zeit wird zudem häufig die Computerlesbarkeit als Kriterium für die Korpusdefinition herangezogen. Nach diesem engeren Verständnis handelt es sich nur bei jenen Textsammlungen um Korpora, die auch computerlesbar sind. **Computerlesbare Korpora** werden auch elektronische Korpora oder Computerkorpora genannt.

Es ist wichtig, im Hinterkopf zu behalten, dass es zwei verschiedene Auffassungen darüber gibt, was als Korpus bezeichnet werden kann. Unter den traditionellen weiteren Korpusbegriff fallen eine Vielzahl von Textsammlungen, die nicht computerlesbar sind. Diese werden, um sie von computerlesbaren Korpora abzugrenzen

häufig als Belegsammlungen, Textarchive oder Ähnliches bezeichnet. Ich verstehe den Begriff Korpus jedoch im traditionellen Sinn als Oberbegriff für computerlesbare und nicht computerlesbare Korpora und werde ihn in diesem Buch entsprechend verwenden.

Prinzipiell kommen Korpora in unterschiedlichen Stadien der maschinellen Bearbeitbarkeit vor. Zum einen gibt es Korpora wie das Mainzer Zeitungskorpus, die in reiner Papierform vorliegen und aus denen die benötigten Informationen manuell herausgesucht werden. Diese Korpora werde ich im Folgenden als **Papierkorpora** bezeichnen. Papierkorpora bilden den Schwerpunkt meiner Ausführungen in Abschnitt 4, wo die Arbeit mit einem eigenen Korpus erläutert wird. Zum anderen gibt es Korpora, die als reine Textdateien gespeichert sind, und wo man mithilfe einfacher Suchbefehle nach einzelnen Zeichenfolgen wie *autovermietung*, *vermietung*, *vermiet* oder *ung* suchen kann. Dies ist der Fall bei den Texten im Projekt Gutenberg, bei den Texten im Trierer Mittelhochdeutschen Textarchiv oder ganz allgemein, wenn elektronische oder Internet-Editionen von Texten verwendet werden. Darüber hinaus gibt es Computerkorpora, die wie die Korpora des IDS oder das TIGER-Korpus annotiert sind. Annotierte Korpora enthalten über den reinen Text hinausgehende grammatische oder strukturelle Informationen, die über standardisierte oder speziell entwickelte Programme abfragt werden können (vgl. Kapitel 2.6, 4.5). Schließlich gibt es Korpora, die weder auf Papier noch als Textdatei vorliegen, sondern ausschließlich in Form von Sprach- oder Videoaufnahmen. Diese Korpora werden jedoch im Folgenden nicht weiter behandelt.

2.3 Korpus und Teilkorpora

Korpora enthalten im Normalfall eine Vielzahl von Texten oder Textteilen, da ein einzelner Text nicht als repräsentativer Ausschnitt für eine sprachliche Gesamtheit angesehen werden kann. Will man etwa die Sprache des Sturm und Drang untersuchen, so reicht es nicht aus, ein Gedicht von Goethe oder ein Drama von Schiller zu untersuchen. Vielmehr wird man eine Vielzahl von Gedichten und Dramen, daneben aber auch Romane, Briefe und andere Textsorten einbeziehen. Man wird sich auch kaum auf Goethe und Schiller beschränken, sondern zudem Texte anderer Autoren wie Lenz, Klinger und Herder berücksichtigen. Korpora bestehen demnach aus einer bestimmten Anzahl an kleineren sprachlichen Einheiten, die zusammen eine übergeordnete Gesamtheit bilden.

Es kann jedoch sinnvoll sein, innerhalb eines Gesamtkorpus einzelne Texte nach bestimmten Kriterien zu **Teilkorpora** zusammenzufassen. Teilkorpora können nach dem Medium (gesprochen versus geschrieben), nach bestimmten Sprechereigenschaften (Texte von Frauen bzw. Männern), nach den enthaltenen Textsorten (Zeitungstexte versus literarische Texte) oder nach historischen Epochen (Texte aus dem 19. bzw. 20. Jahrhundert) gebildet werden.

Die Teilkorpora können bereits vor der Erstellung des Korpus definiert werden wie im Fall des Mannheimer Wendekorpus, das sich aus den Teilkorpora "Wendekorpus Ost", "Wendekorpus West" und dem Teilkorpus zur Wiedervereinigung zusammensetzt. Teilkorpora können aber auch auf der Grundlage bereits bestehender Korpora nachträglich definiert werden. Das DWDS-Kernkorpus erlaubt es z.B., bestimmte Textsorten (Zeitung, Belletristik, Wissenschaft, Gebrauchsliteratur) und Zeiträume auszuwählen und so beliebige Teilkorpora zusammenzustellen (vgl. Kapitel 5.2). Noch komfortabler sind die Möglichkeiten, die das IDS bei der Definition von Teilkorpora bietet (vgl. Kapitel 5.3).

Aufgabe 8: Versuchen Sie, die in den Aufgaben 3, 5 und 6 konzipierten Korpora zur Fachsprache des Rechts bzw. der Medizin und der Jugendsprache sinnvoll in Teilkorpora zu gliedern (vgl. Kapitel 1.3, 1.4).

2.4 Volltextkorpus und Probenkorpus

Korpora unterscheiden sich darin, ob sie aus vollständigen Texten oder aus Textteilen aufgebaut sind. **Volltextkorpora** wie das Mannheimer-Morgen-Korpus oder das DWDS-Korpus enthalten Texte in ihrer gesamten Länge. **Probenkorpora** wie das Bonner Frühneuhochdeutsch-Korpus oder das LIMAS-Korpus hingegen bestehen aus Textausschnitten einer genormten Größe. Das Bonner Frühneuhochdeutsch-Korpus umfasst 40 Textproben von jeweils 30 Normalseiten Länge. Das LIMAS-Korpus hingegen enthält 500 Textproben mit einer Länge von jeweils 2.000 Textwörtern.

Die Verwendung von Textteilen ist zum größten Teil historisch bedingt. Als Vorbild für den Aufbau eines Korpus diente lange Zeit das amerikanische Brown-Korpus aus den frühen siebziger Jahren, das sich aus 500 Textproben mit einer Länge von 2.000 Textwörtern zusammensetzt. Auch was die Korpusgröße betrifft, war das Brown-Korpus mit einer Million Textwörtern lange Zeit wegweisend. Textproben zu verwenden hat aber eine Reihe von Nachteilen:

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 bestimmtem 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 andererseits mehrere Texte zusammengefasst werden, um die vorgegebene Normlänge zu erreichen.

Vor diesem Hintergrund erscheint es sinnvoll, mit Volltexten zu arbeiten und, sofern eine bestimmte Textlänge wichtig ist, gegebenenfalls bereits im Vorfeld die Länge der einzelnen Texte als Auswahlkriterium zu berücksichtigen. Sollen die einzelnen Texte etwa gleich lang sein, so ist es sicherlich sinnvoll, aus der Zahl der möglichen Texte jene auszuwählen, die in ihrer Länge am wenigsten differieren.

2.5 Statische Korpora und Monitorkorpora

Korpora dienen als Abbild einer Sprache oder einer Varietät. Bei diesem Abbild handelt es sich üblicherweise um eine sprachliche Momentaufnahme, bei der bestimmte Texte anhand von vorab definierten Kriterien ausgewählt werden. So enthält das Bochumer Mittelhochdeutsch-Korpus ebenso wie das Mannheimer Wendekorpus eine bestimmte Auswahl an Texten. Beim Mannheimer Wendekorpus sind dies 3.387 Texte unterschiedlicher Textsorten aus den Jahren 1989 und 1990 wie Zeitungsartikel, Flugblätter, Protokolle, Reden usw., die sich mit der Wende und der deutschen Wiedervereinigung beschäftigen. Das Bochumer Mittelhochdeutsch-Korpus enthält hingegen je zwei Vers- und Prosatexte aus sechs verschiedenen Dialektgebieten und fünf verschiedenen Zeiträumen. Für die letzten beiden Zeiträume kommen ergänzende Urkundentexte hinzu. Im Bochumer Mittelhochdeutsch-Korpus ebenso wie im Mannheimer Wendekorpus stehen folglich nicht nur die Anzahl und die Merkmale der enthaltenen Texte fest, sondern auch die Texte als solche. Beide Korpora sind in ihrer Gesamtheit unveränderlich und erfüllen somit das Kriterium der Beständigkeit. Ein Korpus, dessen Zusammensetzung konstant bestehen bleibt, wird als **statisches Korpus** bezeichnet.

Anders ist dies beim Mannheimer-Morgen-Korpus des IDS. Dieses Korpus ist ein so genanntes **Monitorkorpus**, dessen Zusammensetzung sich mit der Zeit verändert. Beim Mannheimer-Morgen-Korpus geschieht dies dadurch, dass regelmäßig aktuelle Zeitungsartikel in das Korpus aufgenommen werden. Dies bedeutet, dass die Größe des Korpus kontinuierlich wächst. Umfasste das Mannheimer-Morgen-Korpus Ende 1999 noch knapp 64 Millionen Textwörter, so waren es zwei Jahre später bereits knapp 108 Millionen und Ende 2003 rund 141,7 Millionen Textwörter. Ähnliches gilt auch für die Bank of English, die gegenwärtig über 550 Millionen Textwörter umfasst.

Monitorkorpora nehmen aber nicht zwangsläufig an Umfang zu. Auch bei einem Korpus, bei dem die ursprüngliche Struktur und Größe erhalten bleibt, wo jedoch regelmäßig ältere Texte gegen neuere Texte mit denselben Strukturmerkmalen wie Textsorte, Medium oder Länge ausgetauscht werden, handelt es sich um ein Monitorkorpus.

2.6 Annotierte und nicht annotierte Korpora

Neben Korpora, die nur Primärdaten, also reinen Text beinhalten, gibt es auch Korpora, die zusätzlich zum Text grammatische und/oder strukturelle Informationen etwa zu Wortart oder Flexion enthalten. Diese über den Text hinausgehende Information in einem Korpus nennt sich **Annotation**. Sie wird mithilfe von speziellen Markierungen im Text kodiert (vgl. Kapitel 4.5). Der Vorteil einer Annotation ist, dass sie implizite Information explizit macht und somit eine einfachere und schnellere Erhebung der benötigten Informationen ermöglicht. Ein wichtiger Grundsatz ist jedoch, dass die Annotation den Originaltext nicht zerstören darf. Es muss also jederzeit möglich sein, die Markierungen zu entfernen und die Rohfassung wiederherzustellen. Korpora, die über den Text hinausgehende Informationen enthalten, nennt man **annotierte Korpora**.

Bei den über den Text hinausgehenden Informationen handelt es sich zum einen um Metadaten, d.h. Angaben zu den einzelnen Texten (vgl. Kapitel 1.4), zum anderen werden linguistische Informationen über die im Text enthaltenen Einheiten kodiert.

Die linguistische Annotation kann auf den unterschiedlichsten Ebenen der Sprache erfolgen. So können Informationen auf Wort-, Satz-, Text-, Laut- oder Bedeutungsebene eingefügt werden. Auf der Lautebene können Merkmale der Aussprache (phonetische An-

notation) sowie der Betonung und Intonation (prosodische Annotation) kodiert werden. Phonetisch oder prosodisch annotierte Korpora sind jedoch selten. Eine Ausnahme stellen die Korpora des Bayerischen Archivs für Sprachsignale (BAS) dar. Auf Wortebene werden Informationen über Flexionsmerkmale oder Wortarten eingefügt (morphologische Annotation), auf Satzebene Informationen über Phrasentypen oder syntaktische Funktionen (syntaktische Annotation). Auf der Bedeutungsebene werden Bedeutungsmerkmale von Wörtern oder inhaltliche Beziehungen zwischen Elementen im Text kodiert (semantische Annotation). Die diskurs- und textlinguistische Annotation hingegen erfasst Phänomene wie die sprachliche Markierung von Höflichkeit oder die sprachliche Wiederaufnahme bzw. Vorwegnahme bestimmter Sachverhalte im Text. Zudem ist auch die Annotation spezifischer Fragestellungen möglich wie die Annotation bestimmter Fehlertypen in einem Korpus mit Texten von Fremdsprachlernern (problemorientierte Annotation).

Am weitesten verbreitet ist jedoch die grammatische Annotation auf Wort- und auf Satzebene (vgl. Kapitel 4.5). Korpora, die auf Satzebene annotiert sind, nennt man **Baumbanken**, da die Informationen auf Satzebene häufig in Form von Strukturbäumen dargestellt werden. Im Unterschied zu reinen Textkorpora und auf Wortebene annotierten Korpora stellt in Baumbanken nicht das Wort die zentrale Analyseeinheit dar, sondern der Satz bzw. kleinere syntaktische Einheiten wie die Phrase.

Die bekannteste deutschsprachige Baumbank ist das TIGER-Korpus, ein Gemeinschaftsprojekt der Universitäten Potsdam, Saarbrücken und Stuttgart. Es umfasst gegenwärtig rund 50.000 Sätze mit knapp 900.000 Textwörtern (vgl. Kapitel 5.4). Weitere wichtige deutschsprachige Baumbanken sind das Saarbrücker NEGRA-Korpus und die drei Tübinger Baumbanken. Das NEGRA-Korpus ist ein Vorläufer des TIGER-Korpus und die erste deutschsprachige Baumbank überhaupt. Wie das TIGER-Korpus basiert es auf Zeitungstexten aus der *Frankfurter Rundschau*. Das NEGRA-Korpus umfasst in seiner aktuellen zweiten Version 20.602 Sätze mit rund 355.000 Textwörtern. Auch der Tübinger Baumbank des Deutschen/Schriftsprache (TüBa-D/Z) und dem Tübinger Partiell Geparsten Korpus des Deutschen/Schriftsprache (TüPP-D/Z) liegen Zeitungstexte zugrunde. Im Gegensatz dazu basiert die Tübinger Baumbank des Deutschen/Spontansprache (TüBa-D/S) auf gesprochener Sprache. Sie enthält rund 38.000 transkribierte Sätze bzw. 360.000 Textwörter, die einer Sammlung von Dialogen zur Terminvereinbarung entstammen.

2.7 Korpora der geschriebenen und gesprochenen Sprache

Ein wichtiges Merkmal von Korpora ist, welchem sprachlichen Medium die Texte in einem Korpus entstammen, ob es sich also um Texte der gesprochenen oder der geschriebenen Sprache handelt. Insgesamt überwiegt bei Weitem die Zahl der schriftsprachlichen Korpora, es gibt jedoch auch Korpora, die ausschließlich gesprochene Sprache oder beides, sowohl Laut- als auch Schriftsprache, enthalten. Für die Dominanz von **Korpora der geschriebenen Sprache** gibt es zwei Gründe, die eng miteinander zusammenhängen: den Aufwand bei der Korpuserstellung und die Verfügbarkeit des Materials.

Prinzipiell ist der Aufwand bei der Erstellung eines schriftsprachlichen Korpus im Vergleich deutlich geringer. Dies liegt daran, dass geschriebene Texte bereits auf Papier, teilweise auch elektronisch in Form einer Textdatei, vorliegen. Sie sind somit leichter zu handhaben und in ein Korpus zu integrieren als Gespräche, Telefonanrufe oder Reden. Um diese und andere Texte der gesprochenen Sprache in ein Korpus aufzunehmen, müssen die einzelnen Monologe oder Dialoge zuerst aufgenommen und anschließend transkribiert, d.h. in schriftliche Form gebracht, werden. Die Verschriftung des Materials erfordert einen hohen Zeitaufwand, der mit der Genauigkeit der Transkription zusätzlich steigt. Allerdings machen es erst die Transkripte möglich, gesprochene Sprache in größerem Umfang systematisch zu erforschen. Bei Texten der geschriebenen Sprache entfällt hingegen der Aufwand für die Transkription.

Was die Verfügbarkeit der Sprachdaten betrifft, so können nur Texte in ein Korpus integriert werden, die in irgendeiner Form konserviert oder konservierbar sind. Dies bedeutet, dass es deutlich leichter ist, ein Korpus der Gegenwartssprache aufzubauen als ein Korpus mit Material aus früheren Sprachstufen. Dies betrifft insbesondere die Lautsprache. Gesprochenes Gegenwartsdeutsch lässt sich jederzeit aufnehmen, für älteres Material ist die Forschung auf Ton- und Filmaufzeichnungen angewiesen, deren Menge und Qualität abnimmt, je weiter man in der Geschichte zurückgeht. Für die Epochen, die vor der Erfindung von Phonograph, Grammophon und Tonband im späten 19. bzw. frühen 20. Jahrhundert liegen, existiert authentisches Material gesprochener Sprache überhaupt nicht. Will man hier lautsprachliche Daten untersuchen, muss man auf schriftliche Textsorten zurückgreifen, die nahe an der gesprochene Sprache sind wie Dramentexte, Predigten oder Gerichtsprotokolle.

Allerdings nimmt aufgrund der Überlieferungslage auch die Verfügbarkeit geschriebener Texte mit zunehmendem historischen Abstand ab, gleichzeitig steigt der Bearbeitungsaufwand. Lassen sich Texte, die in der heute gebräuchlichen Schrifttype Antiqua gedruckt sind, noch ohne größere Probleme einscannen und per Texterkennungsprogramm in eine Textdatei umwandeln, so stellen Texte in Fraktur die übliche Scannersoftware häufig vor Probleme. Sollen mittelalterliche Handschriften oder auch handgeschriebene Texte jüngeren Datums wie Privatbriefe oder Notizzettel in ein elektronisches Korpus aufgenommen werden, kommt man um ein Abtippen der Texte nicht mehr herum. Aber selbst wenn man eine Untersuchung zur Gegenwartssprache plant, kann es für bestimmte Varietäten wie Dialekte oder Kindersprache schwer sein, an computerlesbares Material zu kommen.

Ob ein Korpus Texte der geschriebenen und/oder der gesprochenen Sprache enthält, sollte in erster Linie davon abhängen, welche Varietät oder Varietäten einer Sprache das Korpus abbilden soll. In der Realität spielen häufig die Verfügbarkeit der Daten und der Aufwand bei deren Beschaffung eine limitierende Rolle. Ein Korpus wie das britische Nationalkorpus (BNC), das eine Sprache, in ihrer ganzen Breite repräsentieren soll, sollte demnach sowohl gesprochene als auch geschriebene Sprache enthalten. Tatsächlich beinhaltet das BNC 90% geschriebene und 10% gesprochene Sprache. Bei einem Gesamtumfang von 100 Millionen Textwörtern entspricht der Anteil der gesprochenen Sprache somit einer Anzahl von rund 10 Millionen Textwörtern (vgl. Kapitel 2.9).

Zwar basiert die Mehrheit der deutschsprachigen Korpora auf der Schriftsprache, es gibt aber auch eine ganze Reihe an **Korpora der gesprochenen Sprache**. Für das Deutsche bieten sich das Archiv für Gesprochenes Deutsch (AGD) des IDS und das Bayerische Archiv für Sprachsignale (BAS) als Anlaufstelle an. Das BAS verfügt über rund 20 Korpora, die für den nicht-kommerziellen Zweck öffentlich zugänglich sind. Das AGD verwaltet rund 40 Korpora, die in Form von Transkripten, Ton- und Videoaufnahmen vorliegen. Über die Hälfte dieser Korpora wie das Pfeffer-Korpus zur deutschen Umgangssprache oder das Saarbrücker Korpus der Kindersprache sind öffentlich zugänglich (vgl. Kapitel 5.3).

Wer sich speziell für Kindersprache interessiert, wird auch beim Child Language Data Exchange System (CHILDES) fündig. Hier sind neben Korpora in verschiedenen anderen Sprachen auch mehrere Korpora mit Transkripten deutscher Kindersprache zugänglich.

Schließlich sind noch das Kiel-Korpus des Instituts für Phonetik und digitale Sprachverarbeitung (IPdS) in Kiel und die Tübinger Baumbank des Deutschen/Spontansprache (TüBa-D/S) zu nennen. Das Kiel-Korpus ist für all jene interessant, die sich nicht nur für grammatische Merkmale der gesprochenen Sprache, sondern für die tatsächliche Aussprache der einzelnen Wörter, Sätze und Laute interessieren, da es Informationen zu den einzelnen lautlichen Segmenten enthält. Die Tübinger Baumbank der Spontansprache hingegen ist eines der wenigen Korpora der gesprochenen Sprache, das auf Satzebene annotiert ist (vgl. Kapitel 2.6). Weitere Korpora des gesprochen Deutschen sind über das Linguistic Data Consortium (LDC) erhältlich. Jedoch ist nur ein Teil dieser Korpora transkribiert. Zudem fällt für die Nutzung der Korpora eine Lizenzgebühr an, die mehrere Hundert Euro betragen kann.

Es gibt aber nicht nur Korpora der gesprochenen und der geschriebenen Sprache. Auch für die Gebärdensprache existieren Korpora, wie sie beispielsweise vom Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser in Hamburg zur Erstellung von Fachgebärdenlexika genutzt werden.

2.8 Korpora der Gegenwartssprache und historische Korpora

Ob man ein Korpus als Korpus der Gegenwartssprache bezeichnet oder nicht, hängt davon ab, wie man den Begriff Gegenwartssprache fasst. Sicherlich würde man ein Korpus mit Texten aus den Jahren 2000 bis heute als **Korpus der Gegenwartssprache** bezeichnen. Dasselbe gilt wohl auch für Texte aus den neunziger Jahren des 20. Jahrhunderts. Was aber ist mit Texten aus den Fünfzigern oder Siebzigern? Was ist mit Texten aus dem späten 19. oder dem frühen 20. Jahrhundert? Kann man diese noch als Gegenwartssprache bezeichnen oder nicht?

Ein Korpus ist in der Regel statisch, d.h. es erfasst prinzipiell nur einen bestimmten Zeitabschnitt der Sprache (vgl. Kapitel 2.5). So war das Bonner LIMAS-Korpus, das hauptsächlich Texte aus den frühen siebziger Jahren enthält, anfangs ein Korpus der Gegenwartssprache. Heute, gut dreißig Jahre nach seiner Fertigstellung, ist fraglich, ob das LIMAS-Korpus als repräsentativ für das Deutsche zu Beginn des 21. Jahrhunderts gelten kann. Die Antwort hängt davon ab, welche Fragestellung untersucht werden soll. Zu bedenken ist, dass manche sprachlichen Veränderungen wie etwa Neuerungen im Wortschatz innerhalb relativ kurzer Zeit stattfinden,

wohingegen sich andere Veränderungen, etwa in der Satzstruktur, deutlich langsamer vollziehen. Aktuelle, neue Wörter wie *bloggen* oder *Wellness*, auch Neologismen genannt, wird man im LIMAS-Korpus vergeblich suchen, sodass das Korpus für eine Untersuchung zu Neologismen in der Gegenwartssprache veraltet wäre. Die Stellung des Verbs im Satz hingegen ist seit Jahrhunderten unverändert geblieben, sodass das LIMAS-Korpus hinsichtlich der Verbstellung durchaus den gegenwärtigen Zustand der Sprache repräsentieren kann. Was Untersuchungen zur Verbstellung betrifft, spiegeln sogar noch Texte aus dem 19. und frühen 20. Jahrhundert den Zustand des Gegenwartsdeutschen wider.

Unabhängig davon, wo man die Grenze zwischen Gegenwart und Vergangenheit zieht, ist klar, dass Korpora altern. Je größer der zeitliche Abstand zwischen der Entstehung eines Textes und der Gegenwart wird, umso eher ist ein Text und das Korpus, in dem er enthalten ist, als historisch einzustufen. Auch wenn es seltsam erscheint, eine Zeit, die nur kurz vor oder sogar nach der eigenen Geburt liegt, aus historischer Perspektive zu betrachten, so sind Korpora, die zwanzig oder dreißig Jahre alt sind, je nach Fragestellung als **historische Korpora** oder als Korpora der Gegenwartssprache anzusehen.

Ein weiteres Problem bei der Unterscheidung zwischen gegenwartsprachlichen und historischen Korpora ist, dass es durchaus Texte aus früheren Jahrhunderten gibt, die wie literarische Texte von Goethe und Schiller, philosophische Texte von Kant oder die Bibel in ihrer Übersetzung von Luther bis heute gelesen werden. Obwohl es sich von der Entstehung her um historische Texte handelt, sind diese Texte aufgrund ihrer Rezeption in der Gegenwart in gewisser Weise Bestandteil der Gegenwartssprache.

Fraglos wird man jedoch jene Korpora als historisch bezeichnen, deren Texte bereits ein oder mehrere Jahrhunderte alt sind. Da es oftmals sehr schwierig ist, ein historisches Korpus zu erstellen, ist die Zahl der Korpora, die Material aus früheren Jahrhunderten enthalten, im Vergleich zu Korpora mit Texten aus den letzten fünfzig Jahren deutlich geringer. Dennoch gibt es eine ganze Reihe kleinerer und größerer historischer Korpora. Das IDS in Mannheim ist gegenwärtig mit dem Aufbau eines historischen Korpus beschäftigt, das den Zeitraum von 1700 bis 1945 abdecken soll (vgl. Kapitel 5.3). Texte aus allen Dekaden des 20. Jahrhunderts umfasst das DWDS-Korpus (vgl. Kapitel 5.2).

Neben Korpora zum Neuhochdeutschen gibt es jedoch auch Korpora, die die Sprache des Mittelalters, das Althochdeutsche (ca. 750-1050) bzw. Mittelhochdeutsche (ca. 1050-1350), und die Sprache der frühen Neuzeit, das Frühneuhochdeutsche (ca. 1350-1600), erfassen. Alt- und mittelhochdeutsche Texte findet man in der Bibliotheca Augustana, beim Internetportal Mediaevum und in der Datenbank Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS). Auf mittelhochdeutschen Texten basieren auch das Digitale Mittelhochdeutsche Textarchiv in Trier und das Bochumer Mittelhochdeutsch-Korpus. Korpora des Frühneuhochdeutschen sind das Bonner Frühneuhochdeutsch-Korpus oder das Erlanger Dürer-Korpus. Ein historisches Referenzkorpus für das Deutsche, das Texte aus dem 9. bis 19. Jahrhundert umfassen soll, ist im Rahmen des Projekts Deutsch Diachron Digital (DDD) in Planung.

Was die zeitliche Dimension bei Korpusstudien betrifft, muss man zusätzlich entscheiden, ob das Korpus verwendet wird, um einen Zeitpunkt oder eine Zeitspanne zu untersuchen. Im ersten Fall handelt es sich um eine synchrone Untersuchung, im zweiten Fall um eine diachrone. Beide Arten von Untersuchung können denselben Zeitraum, z.B. das 20. Jahrhundert, umfassen. Der Unterschied besteht darin, dass bei einem synchronen Vorgehen die Gemeinsamkeiten der Texte im Korpus erforscht werden, wohingegen bei diachroner Betrachtung der Sprachwandel, d.h. die Veränderungen, die sich in den Texten des Korpus zeigen, im Mittelpunkt des Interesses stehen.

2.9 Referenzkorpora und Spezialkorpora

Unterschiedliche Korpora erlauben, eine unterschiedliche Zahl und Art von Fragestellungen zu beantworten. Grundsätzlich sollte man trennen zwischen Referenzkorpora und Spezialkorpora. Ein **Referenzkorpus** ist ein Korpus, das dazu bestimmt ist, eine Sprache in ihrer Gesamtheit zu repräsentieren und eine Vielzahl von sprachlichen Informationen zu liefern. Ein Referenzkorpus sollte so groß sein, dass es als Grundlage für die Untersuchung aller wichtigen Varietäten der Sprache dienen kann, seien es nun regionale oder gruppenspezifische Varietäten. Das deutsche Referenzkorpus (DEREKO) am IDS in Mannheim umfasst gegenwärtig rund zwei Milliarden Textwörter. Im Gegensatz zum britischen Nationalkorpus (BNC) enthält es allerdings ausschließlich Texte der geschriebenen Sprache. Eine Untersuchung zur gesprochenen Sprache ist somit im Rahmen des deutschen Referenzkorpus nicht möglich. Streng genommen handelt es sich bei DEREKO also nicht um ein

Referenzkorpus der deutschen Sprache, sondern um ein Referenzkorpus der deutschen Schriftsprache.

Das britische Nationalkorpus (BNC) hingegen, das das britische Englisch repräsentieren soll, beinhaltet zu 90% Texte der geschriebenen Sprache und zu 10% transkribierte gesprochene Sprache. Die Texte der Schriftsprache umfassen unter anderem Auszüge aus regionalen und überregionalen Zeitungen, Fachzeitschriften und Zeitschriften für alle Alters- und Interessensgruppen, wissenschaftliche und nichtwissenschaftliche Bücher, veröffentlichte wie unveröffentlichte Briefe und Notizen, Aufsätze von Schülern und Studenten. Bei den Texten der gesprochenen Sprache handelt es sich überwiegend um informelle spontane Gespräche, die von Freiwilligen aufgenommen wurden, daneben aber auch um gesprochene Sprache aus anderen Kontexten wie Regierungsgespräche oder Radiosendungen. Wenn man sich allerdings klar macht, welchen Anteil an der tagtäglichen Kommunikation die gesprochene Sprache einnimmt, erscheint die gesprochene Sprache auch mit einem Anteil von 10% noch deutlich unterrepräsentiert zu sein. Für eine Untersuchung zur gesprochenen Sprache reichen die im BNC enthaltenen 10 Millionen Textwörter dennoch gut aus.

Spezialkorpora erheben hingegen nicht den Anspruch, repräsentativ für eine Sprache in ihrer Gesamtheit zu sein. Sie dienen vielmehr dazu, eine bestimmte Varietät der Sprache wie die Jugendsprache, die deutsche Rechtssprache, die Zeitungssprache, das Hessische oder die Sprache von Deutsch-als-Fremdsprache-Lernern zu erforschen. Diesen Teilbereich der Sprache sollten sie jedoch hinreichend repräsentieren.

Viele bekannte Korpora sind Spezialkorpora, die auf bestimmten Textsorten basieren. Insbesondere Korpora der Zeitungssprache sind weit verbreitet. Eine große Zahl der IDS-Korpora wie das Mannheimer-Morgen-Korpus oder das Bonner Zeitungskorpus bestehen aus Zeitungstexten. Auch verschiedene Baumbanken wie das TIGER-Korpus und das NEGRA-Korpus sind Korpora der Zeitungssprache. Daneben gibt es Korpora, die auf andere Textsorten spezialisiert sind, etwa auf Lyrik, Romane, Gebrauchsanweisungen oder Bibelübersetzungen.

Als Spezialkorpora einzuordnen sind aber auch die bereits erwähnten Lerner- und Spracherwerbskorpora. Lernerkorpora wie das Fehler-annotierte Lernerkorpus des Deutschen als Fremdsprache (FALKO), das derzeit in Berlin aufgebaut wird, enthalten Texte, die von Schülern und Studenten in einer Fremdsprache verfasst wurden (vgl. Kapitel 1.5). Spracherwerbskorpora wie die CHILDES-Korpora und das Saarbrücker Korpus der Kindersprache umfassen Transkripte und Aufzeichnungen von – in der Regel gesprochener – Kindersprache (vgl. Kapitel 2.7).

Eine große Zahl an Spezialkorpora sind Fachtextkorpora wie das Darmstädter Korpus deutscher Fachsprachen, das rund 2,8 Millionen Textwörter aus den Gebieten Bauingenieurwesen, Elektrotechnik, Maschinenbau und Wirtschaft enthält. Weitere Fachtextkorpora sind die bereits erwähnten Korpora zum Computerdiskurs und zur Sprache der Bochumer Stadtverwaltung (vgl. Kapitel 1.5). Ein mehrsprachiges Fachtextkorpus ist das OPUS-Korpus, das Teilkorpora zur Verwaltungssprache und zu verschiedenen technischen Disziplinen umfasst (vgl. Kapitel 2.10). Als Beispiele für historische Fachtextkorpora sei an dieser Stelle lediglich auf zwei frühneuhochdeutsche Korpora, nämlich das Olmützer medizinische Korpus sowie das Erlanger Dürer-Korpus mit mathematisch-technischen Texten, verwiesen.

2.10 Einsprachige und mehrsprachige Korpora

Die meisten Korpora enthalten nur Daten aus einer Sprache. Dies gilt auch für Referenzkorpora wie DEREKO oder BNC, die zwar Material zu einer Vielzahl von Varietäten beinhalten, die jedoch alle derselben Sprache, in diesem Fall dem Deutschen bzw. Englischen, zuzuordnen sind (vgl. Kapitel 2.9). Daneben gibt es aber auch Korpora wie das International Sample of English Contrastive Texts (INTERSECT) oder das Chemnitzer German/English-Translation-Korpus, die sowohl deutsche als auch englische Texte enthalten. Das Verbmobil-Korpus umfasst neben deutschen und englischen Texten auch japanische Texte. Eine besonders große sprachliche Vielfalt bietet das OPUS-Korpus. Die Anzahl der enthaltenen Sprachen schwankt in den fünf Teilkorpora zwischen sechs (OpenOffice-Korpus) und 61 (KDE-Korpus). Eine Vielzahl mehrsprachiger, allerdings kostenpflichtiger Korpora bietet die Evaluations and Language Resources Distribution Agency (ELDA).

Bei **mehrsprachigen Korpora** ist zu unterscheiden zwischen so genannten Parallelkorpora und vergleichbaren Korpora. **Parallelkorpora** wie das Chemnitzer German/English-Translation-Korpus zeichnen sich dadurch aus, dass sie Originaltexte in einer Sprache und deren Übersetzung in eine oder mehrere andere Sprachen beinhalten. Das Chemnitzer Korpus umfasst insgesamt rund zwei Millionen Textwörter, je eine Million Textwörter für das Deutsche

und das Englische, die aus den Bereichen Politik, Wissenschaft und Tourismus stammen. Zugänglich ist das Korpus über die Internetseite der Chemnitzer Internet-Grammatik.

Wie das Chemnitzer German/English-Translation-Korpus ist auch das OPUS-Korpus ein Parallelkorpus, das im Internet frei zugänglich ist. Es enthält Gebrauchsanweisungen und Dokumente der Europäischen Union mit den jeweiligen Übersetzungen. Dabei kann ausgewählt werden, in wie viele Sprachen eine Textpassage übersetzt werden soll. In Abbildung 2 werden die deutsche, englische, französische und finnische Version eines Satzes angezeigt.

11014334	Als Katalane würde ich mir für die Zukunft wünschen, dass auch meine **Sprache**, die von zehn Millionen europäischen Bürgern gesprochen wird, hier im Hause offiziell anerkannt wird.
en	As a native Catalan, it is my hope that, in the future, my language, which is spoken by 10 million European citizens, may also be recognised in this Chamber.
fi	Syntyperäisenä katalonialaisena toivon, että tulevaisuudessa äidinkieleni, jota puhuu 10 miljoonaa Euroopan kansalaista, tunnustetaan myös tässä parlamentissa.
fr	En tant que Catalan, je souhaiterais que, demain, ma langue, qui est celle de dix millions de citoyens européens, ait également droit de cité dans cette maison.

Abbildung 2: OPUS-Korpus: Konkordanz für *Sprache* (Ausschnitt)

Charakteristisch für **vergleichbare Korpora** ist, dass alle Teilkorpora denselben Aufbauprinzipien folgen. Die Teilkorpora verfügen also über eine identische Struktur. Im Gegensatz zu Parallelkorpora, die prinzipiell mehrsprachig sind, können vergleichbare Korpora sowohl ein- als auch mehrsprachig sein.

Ein Beispiel für ein mehrsprachiges vergleichbares Korpus ist das PAROLE-Korpus. Es besteht aus zwölf Teilkorpora, unter anderem einem deutschen, englischen und französischen Teilkorpus, mit jeweils rund 20 Millionen Textwörtern. Drei weitere Teilkorpora enthalten eine geringere Anzahl an Textwörtern. Alle Teilkorpora wurden nach einheitlichen Kriterien aufgebaut und mit zusätzlichen grammatischen Informationen versehen. In jedem Teilkorpus wurden dieselben Anteile an Texten aus Büchern, Zeitungen und Zeitschriften erhoben. Ein Beispiel für ein mehrsprachiges vergleichbares Korpus der gesprochenen Sprache ist das Verbmobil-Korpus, das deutsche, englische und japanische Spontansprache aus dem Bereich der Terminvereinbarung enthält.

Als vergleichbare einsprachige Korpora sind das Brown-Korpus, das Lancaster-Oslo/Bergen-Korpus (LOB) und das Kolhapur-Korpus zu nennen. Alle drei Korpora verfügen über dieselbe Struktur, nämlich die des Brown-Korpus, bilden jedoch unterschiedliche regionale Varianten des Englischen – amerikanisches, britisches und indisches Englisch – ab.

Aufgabe 9: Das International Corpus of English (ICE) enthält insgesamt 20 Millionen Textwörter. Das Korpus besteht aus zwanzig Teilkorpora aus Ländern, in denen Englisch die einzige oder eine der offiziellen Nationalsprachen ist. Jedes der Teilkorpora enthält zu 60% gesprochene und zu 40% geschriebene Sprache, die nach denselben Kriterien erhoben wurden.
Handelt es sich beim ICE Ihrer Meinung nach um ein einsprachiges oder ein mehrsprachiges Korpus? Handelt es sich bei den Teilkorpora um parallele oder vergleichbare Korpora? Bitte begründen Sie Ihre Ansicht.

2.11 Zusammenfassung

Korpora werden nach formalen Kriterien eingeteilt in computerlesbare und Papierkorpora, in Gesamt- und Teilkorpora, in Proben- und Volltextkorpora, in statische und Monitorkorpora sowie in annotierte und nicht annotierte Korpora.
Im Hinblick auf ihren Inhalt werden Korpora der gesprochenen und der geschriebenen Sprache, Korpora der Gegenwartssprache und historische Korpora, Referenz- und Spezialkorpora sowie ein- und mehrsprachige Korpora unterschieden.

Grundbegriffe: Annotation, annotiertes Korpus, Baumbank, computerlesbares Korpus, einsprachiges Korpus, historisches Korpus, Korpus der Gegenwartssprache, Korpus der geschriebenen Sprache, Korpus der gesprochenen Sprache, mehrsprachiges Korpus, Monitorkorpus, Papierkorpus, Parallelkorpus, Probenkorpus, Referenzkorpus, Spezialkorpus, statisches Korpus, Teilkorpus, vergleichbares Korpus, Volltextkorpus

Weiterführende Literatur
Sinclair (1998) gibt einen Überblick über verschiedene Arten von Korpora. Eine weitere Korpustypologie sowie eine systematische Übersicht über deutschsprachige Korpora bieten Lemnitzer/Zinsmeister (2006, Kapitel 5). Hinweise auf weitere Korpora des Deutschen, insbesondere historische und mehrsprachige Korpora, finden sich in dem Sammelband von Schwitalla/Wegstein (Hgg.) (2005).