1 Korpora

Heike Zinsmeister

Im vorangehenden Kapitel zu den computerlinguistischen Methoden wurden an mehreren Stellen linguistische Korpora erwähnt, die als empirische Datengrundlage dienen und zum Trainieren von (statistischen) Programmen oder allgemein zum Testen eingesetzt werden (siehe zum Beispiel die Unterkapitel **??** und **??**). Korpora können zudem als Zeugnisse für die Möglichkeiten computerlinguistischer Verarbeitung betrachtet werden, da sie oftmals bis zu einem gewissen Grad auf automatischer oder semi-automatischer Vorverarbeitung und Annotation basieren.

Unabhängig von der Verwendung in der Computerlinguistik kann ein linguistisches **Korpus** (neutrum: *das Korpus*) definiert werden als Sammlung gesprochener oder schriftlicher Äußerungen, die digital erfasst, also auf Rechnern gespeichert und maschinenlesbar sind, und für eine linguistische oder computerlinguistische Aufgabe aufbereitet wurden. Von den eigentlichen Korpora unterscheiden sich **Textarchive**, die ebenfalls digitalisierte Sprachdaten enthalten, welche aber nicht primär für linguistische Zwecke bereitgestellt werden. Ein bekanntes Textarchiv ist das internationale **Gutenberg Project** (Lebert 2008) mit Texten, deren Urheberrecht abgelaufen ist bzw. deren Autoren die Texte zur Nutzung freigegeben haben. Reine **Belegsammlungen** unterscheiden sich ebenfalls von Korpora, indem sie nur einzelne Sätze oder Paragraphen aufführen und nicht ganze Texte oder zumindest substanzielle Ausschnitte aus diesen. Belegsammlungen enthalten mitunter auch konstruierte und bewusst ungrammatische Beispielsätze – Korpora hingegen **authentische Sprachdaten**, die in einer linguistisch unreflektierten Kommunikationssituation produziert wurden. Belegsammlungen bieten Evidenz für bestimmte linguistische Phänomene. Die lexikalisch orientierte Sammlung CoDII (*Collection of distributionally idiosyncratic items*, Trawinski *et al.* 2008) zum Beispiel beinhaltet korpusbasierte Belege für Lizenzierungsbedingungen von negativ polaren Ausdrücken (wie *sich scheren um*), die nur zusammen mit einer Negation oder in anderen spezifischen Kontexten auftreten.

Neben der oben genannten Korpusdefinition wird in der Praxis auch ein anwendungsbezogener Korpusbegriff verwendet. Nach diesem wird jeder Text als Korpus bezeichnet, wenn er für linguistische oder computerlinguistische Aufgaben genutzt wird. Dies schließt auch Texte mit ein, die nicht speziell für die Linguistik/Computerlinguistik aufbereitet wurden, sondern in einer unstrukturierten Rohfassung vorliegen wie zum Beispiel Texte, die aus dem HTML-Code von Internetseiten extrahiert oder aus Textarchiven entnommen wurden. Unter diesen weiter gefassten Korpusbegriff fallen auch Sprachdaten, die nicht digital erfasst sind, sondern zum Beispiel nur als magnetische Tonbandaufnahmen oder in gedruckter Form vorliegen (z. B. Ruoff 1984).

Im Folgenden geht es vorwiegend um digitale, aufbereitete Korpora. In Abschnitt 1.1 wird die generelle Architektur eines einzelnen Korpus thematisiert,

in Abschnitt 1.2 eine umfassende Typologie, die dabei hilft, Korpora anhand verschiedener Eigenschaften zu systematisieren. Die Typologie soll auch dazu dienen, einige bekannte Korpusressourcen vorzustellen. Wozu Korpora in der Computerlinguistik verwendet werden, wird in Abschnitt 1.3 erklärt. Abschnitt 1.4 bietet schließlich Hinweise auf weiterführende Arbeiten und Ressourcen.

1.1 Aufbau eines Korpus

Ein aufbereitetes Korpus besteht aus drei Schichten: den Sprachdaten, den analysierenden Annotationen und den beschreibenden Metadaten. Um die Daten verschiedener Projekte austauschbar und vergleichbar zu halten, haben internationale Initiativen Standardisierungsempfehlungen für Annotationen, Metadaten und die allgemeine Datenstruktur und Enkodierung erarbeitet, auf die am Ende dieses Abschnitts kurz eingegangen wird.

1.1.1 Sprachdaten

Den Kern eines Korpus bilden die **Sprachdaten**, die aus Texten, Sprachaufnahmen oder deren Verschriftlichungen bestehen und die in digitalisierter Form abgespeichert sind. Sie können auf sprachlichen **Primärdaten** basieren, die zum Beispiel als Tonaufnahme oder Textveröffentlichung unabhängig vom Korpus existieren. Je nach Art der Primärdaten unterscheidet man **Textkorpora** von **Korpora der gesprochenen Sprache**. Liegen in einem Korpus der gesprochenen Sprache die Primärdaten selbst vor – die physikalisch messbaren Sprachsignale – und nicht nur eine **schriftliche Transkription**, kann das Korpus in einer **Sprachdatenbank** verwaltet werden (siehe Unterkapitel **??**). Textuelle Primärdaten können bereits in digitaler Form vorliegen, müssen es aber nicht. Sie können auch nur als gedruckte Texte, Handschriften oder ähnliches zur Verfügung stehen. Existieren die Primärdaten als konkrete Veröffentlichungen, besitzen sie nicht nur einen Wortlaut, sondern auch eine äußere Form: die Verteilung des Texts auf einer oder mehreren Seiten, die Größe, die Farbe und der Font der Buchstaben usw. Primärtextliche Eigenschaften dieser Art sind selten in den Sprachdaten kodiert, die im Korpus als Grundlage für die weitere Analyse genutzt werden.

1.1.2 Annotation

Als zweite Schicht lagern sich verschiedene **Annotationsebenen** um die Sprachdaten. Die erste Analyseebene besteht aus der **Segmentierung** (auch Segmentation), der Zerlegung des Sprachsignals oder der Zeichenkette in linguistisch definierte Einheiten wie Phoneme, Wörter oder Sätze (siehe auch Unterkapitel **??**). In Sprachdatenbanken bilden **Transkriptionen** die nächste Analyseebene, wobei der Wortlaut einer Äußerung orthographisch als Text wiedergegeben werden kann oder ihre Lautung in phonetischen oder phonemischen Symbolen. Gegebenenfalls werden auch non-verbale Geräusche notiert wie Räuspern oder Lachen. In Textkorpora können bei der Segmentierung

auch **textstrukturelle Einheiten** wie Paragraphen, Kapitel, Überschriften oder Fußnoten abgegrenzt werden. Die Segmentierung kann indirekt kodiert sein zum Beispiel durch die Konvention, dass Tokengrenzen durch Leerzeichen markiert sind, Satzgrenzen durch Zeilenumbrüche und Paragraphengrenzen durch Leerzeilen. Besser ist eine Annotation, welche die Sprachdaten von der Analyse explizit trennt, indem die textstrukturellen Einheiten durch Annotationslabel benannt werden und den Sprachdaten mit Hilfe einer **Auszeichnungssprache** wie **XML** (eXtensible Markup Language) zugeordet werden (vergleiche auch Unterkapitel **??**).

Auf der Basis der Segmentierung können weitere **linguistische und außerlinguistische Annotationsebenen** vorliegen. Sehr oft enthalten Korpora eine Annotation der Wortart (englisch *part of speech*, POS) und der Basiswortform (englisch *lemma*). Allgemeine Empfehlungen zur POS-Annotation wurden von der Text Encoding Initiative formuliert (TEI AI1W2 1991). Für das Deutsche wurden davon die STTS-Guidelines (Stuttgart-Tübingen-Tagset) abgeleitet (Schiller *et al.* 1999). **Syntaktische Annotationen** findet man zur Konstituentenstruktur und zu grammatischen Funktionen (Marcus *et al.* 1993, 1994, Brants *et al.* 2002), Dependenzen (Hajičová *et al.* 1999, Foth 2006) und für das Deutsche auch zu topologischen Feldern (Telljohann *et al.* 2006). **Semantische Annotationen** beinhalten Lesarten (*word senses*, Fellbaum 1998), semantische Rollen und semantische Frames (Palmer *et al.* 2005, Meyers *et al.* 2004, Burchardt *et al.* 2006) sowie Tempus und Aspekt (Pustejovsky *et al.* 2003). Zu den **diskursbezogenen Annotationen** gehören Koreferenzphänomene (Poesio 2000, Naumann 2006), Informationsstatus (Nissim *et al.* 2004, Riester 2008), Informationsstruktur (Calhoun *et al.* 2005, Götze *et al.* 2007), Diskursrelationen (Mann and Thompson 1988, Miltsakaki *et al.* 2004) und Dialogakte (Anderson *et al.* 1991, Carletta *et al.* 1997, Alexandersson *et al.* 1998). In Lernerkorpora zum Erst- oder Fremdspracherwerb werden **Fehler** annotiert (MacWhinney 1995, Granger 2002, Lüdeling 2008), ebenso in Korpora zu gestörter Sprache. Über die rein linguistische Analyse hinaus gehen zum Beispiel die Annotationen von **Emotionen** und **Meinungen** (Wiebe *et al.* 2004), ebenso die Analyse von **Mimik** (Foster 2007) und **Gestik** (Martell 2002, Kipp *et al.* 2007).

Um eine konsistente Annotation zu erreichen und auch für jede spätere Nutzung ist es sehr wichtig, dass die Annotationen ausführlich dokumentiert sind. Die Bedeutung der **Annotationslabel** (*tags*) werden in einem **Tagset** eindeutig definiert und die **Annotationskriterien** in Richtlinien (*guidelines*) mit Beispielen belegt. Für die Dokumentation der **Annotationsqualität** wird die Übereinstimmung unter den Annotatoren festgehalten (*inter-annotator agreement*, Artstein and Poesio 2008).

Das Korpus in den Abbildungen 1 und 2 verwendet für die Wortartenannotation das **STTS-Tagset**. Hier steht das Label *ART* für *Artikel* und *NN* für *Normales Nomen*, welches als Appelativum definiert ist und sich von Eigennamen abgrenzt. Die STTS-Richtlinien geben neben der Definition auch eine Reihe von Beispielen an und verweisen auf jeweils bekannte Grenzfälle zu anderen Labeln.

In Abbildung 1 ist jedes Wort und auch der Satzpunkt in ein XML-Element

```
<sentence editor="shartung" date="2004083117:26:19" origin="T990430.196">
  <word form="Der" pos="ART"/>
  <word form="Scheibenwischer" pos="NN"/>
  <word form="quietscht" pos="VVFIN"/>
  <word form="." pos="$."/>
</sentence>
```

Figure 1: Ausschnitte aus der XML-Inline-Annotation von Satz 20209 der TüBa-D/Z (Telljohann *et al.* 2006): *Der Scheibenwischer quietscht.*

```
<body>Der Scheibenwischer quietscht.</body>

<mark id="tok_1" xlink:href="#xpointer(string-range(//body,'',1,3))"/>
<mark id="tok_2" xlink:href="#xpointer(string-range(//body,'',5,15))"/>
<mark id="tok_3" xlink:href="#xpointer(string-range(//body,'',21,9))"/>
<mark id="tok_4" xlink:href="#xpointer(string-range(//body,'',30,1))"/>

<feat xlink:href="#tok_1" value="stts.type_pos.xml#ART"/>
<feat xlink:href="#tok_2" value="stts.type_pos.xml#NN"/>
<feat xlink:href="#tok_3" value="stts.type_pos.xml#VVFIN"/>
<feat xlink:href="#tok_4" value="stts.type_pos.xml#DOLLAR_PERIOD"/>

<mark id="s_20209" xlink:href="#xpointer(id('tok_1')/range-to(id('tok_4')))"/>
```

Figure 2: Ausschnitte aus der XML-Standoff-Annotation von Satz 20209 der TüBa-D/Z im PAULA-Format (Dipper 2005)

word als Wert des Attributs *form* eingebettet. Informationen über Leerzeichen gehen in dieser Kodierung verloren. Abbildung 2 stellt ein alternatives Format (vereinfacht) dar. Hier werden die Token durch Bezugnahme auf Buchstabenpositionen in der Zeichenkette definiert. Das erste Token *tok_1* (*Der*) beginnt an Position 1 und ist drei Zeichen lang. Das zweite Token *tok_2* (*Schweibenwischer*) beginnt an Position 5 und ist fünfzehn Zeichen lang. Die Wortarten-Label werden über Links den Token zugeordnet. Ebenso ist die Satzausdehnung in Relation zu den Token definiert. Beim ersten Beispiel sind verschiedene Annotationsebenen in einem **Inline-Format** gekoppelt. Die Wortform und das Wortarten-Label sind Attribute eines gemeinsamen Elements *word* und die syntaktische Hierarchie (hier vereinfacht nur die Satz- und Wortebene) ist durch die Einbettung der XML-Elemente (*sentence, word*) nachgebildet. Anders im zweiten Beispiel. Dort werden in einem **Standoff-Format** alle Informationsebenen (Text, Token, Wortart, Satz) getrennt aufgeführt. Sie sind lediglich über *Pointer* und *Links* indirekt miteinander verbunden.

1.1.3 Metadaten

Metadaten werden auch als *Daten über Daten* bezeichnet. In ihnen werden die Primärdaten, die im Korpus enthaltenen Sprachdaten und die Annotationen beschrieben. Sie erfassen zum Beispiel, welchen Textgattungen die Daten zugehören, wie groß der Datenumfang ist und wie die Sprachdaten kodiert sind. Außerdem werden kontextuelle Aspekte der Entstehung des Korpus dokumentiert, zum Beispiel die Entstehungs- und Publikationszeiten der Primärdaten, der Publikationsort, beteiligte Personen, die Entstehungszeit der Annotation und die Namen der Annotatoren. Zusätzlich findet man Verweise auf externe Quellen wie die Definitionen der Annotationslabel (der Tagsets), Annotationsrichtlinien und Publikationen, die das Korpus beschreiben. Eine nicht unerhebliche Information ist die Angabe von **urheberrechtlichen Eigenschaften** des Korpus und seiner Primärdaten. Außer diesen Angaben über die Daten und die Annotationen findet man auch Informationen über die Metadaten selbst, zum Beispiel, ob die Metadaten manuell oder automatisch erstellt wurden und ob sie einem bekannten **Standard** folgen.

1.1.4 Standardisierung

Immer kürzere Zyklen in der Hard- und Softwareentwicklung gefährden die nachhaltige Nutzbarkeit von Korpora. Deshalb sind projektübergreifende, einheitliche Beschreibungen und Formate wichtig, die es erleichtern, Korpusdaten auch über die jeweilige Projektlaufzeit hinaus nutzbar zu halten (Schmidt *et al.* 2006, Zinsmeister *et al.* 2008). Ressourcen, die sich am selben Standard orientieren, können zudem besser miteinander verglichen und kombiniert werden.

Von der **Dublin Core Metadata Initiative** (DC) wurde Mitte der 1990er Jahre erstmals eine Kernmenge von Metadaten für die Beschreibung elektronischer Ressourcen definiert. Im Standard der **Open Language Archive Community** (OLAC) wurde der Dublin Core für mehrsprachige und multimodale Ressourcen, die Text-, Bild- und Audiomaterial verbinden, erweitert. Alternativ hat die **ISLE Meta Data Initiative** (IMDI) ebenfalls einen Metadatenstandard auf der Basis des Dublin Core vorgeschlagen. Die seit mehr als zwanzig Jahren bestehende **Text Encoding Initiative** (TEI) definiert im TEI-Header einen eigenen Satz von Metadaten, mit dem Ziel, speziell Textdokumente und Korpora zu archivieren. Da, wo das DC-Metadatenset zum Beispiel nur ein unspezifisches Element *source* vorsieht, bietet das TEI-Set spezielle Elemente für bibliographische Angaben, so dass deren Bestandteile wie *editor* und *edition* in spezifischen Feldern abgelegt werden können.

Neben den Metadaten schlug die TEI auch ein Standardformat für die Korpusannotation vor. Darauf aufbauend hat ein internationales Gremium zur Standardisierung von sprachtechnologischen Ressourcen, die Expert Advisory Group on Language Engineering Standards, **EAGLES**, Empfehlungen erarbeitet (zum Beispiel Leech *et al.* (1996) für POS-Annotation), die im **Corpus Encoding Standard** (CES bzw. dem XML-basierter Nachfolger XCES) umgesetzt wurden und die TEI-Kategorien um sprachtechnologisch relevante Kate-

gorien erweiterten.

1.2 Typologie

Korpora lassen sich anhand einer Reihe von Kriterien klassifizieren. In Anlehnung an die Korpustypologie in Lemnitzer and Zinsmeister (2006, Kap. 5) werden in den folgenden Abschnitten eine Reihe von Ressourcen vorgestellt.

Weil die Erstellung von Korpora relativ zeit- und kostenintensiv ist, besteht der Anspruch, dass ein Korpus möglichst **wiederverwendbar** und **multifunktional** einsetzbar sein sollte. Der **ursprüngliche Verwendungszweck** eines Korpus legt zwar dessen weitere Nutzung nicht fest, kann aber bestimmte Eigenschaften des Korpus erklären. Im Projekt **Verbmobil** zum Beispiel wurden für die Entwicklung eines Übersetzungssystems für Spontansprache mehrsprachige Korpora erstellt und annotiert (Burger *et al.* 2000, Jekat and v. Hahn 2000). Um das Vorhaben handhabbar zu halten, wurde die sprachliche Domäne auf Terminverhandlungen zwischen Geschäftspartnern, Reiseplanungen und Hotelreservierungen beschränkt. Hierfür wurden spontane Dialoge auf Deutsch, Englisch und Japanisch aufgenommen, von denen Teilkorpora für die Entwicklung und das Testen einer integrierten Grammatikkomponente mit syntaktischer Information annotiert wurden (TüBa-D/S, TüBa-E/S und TüBa-J/S, Hinrichs *et al.* 2000). Obwohl die daraus resultierenden Baumbanken nur die genannten Domänen abdecken, können sie unabhängig vom Verbmobil-Projekt für andere Forschungsfragen zur Syntax bei gesprochener Sprache und Dialogen eingesetzt werden. Das Brown University Standard Corpus of Present-Day American English (kurz: **Brown Corpus**) wurde anders als die Verbmobilkorpora von vornherein als **repräsentatives Korpus** geplant (Francis and Kučera 1979). Es sollte die Gesamtheit des schriftlich veröffentlichten amerikanischen Englisch des Jahres 1961 repräsentieren und umfassende Analysen und computerbasierte Auswertungen erlauben. Dafür wurden nach systematischen Kriterien Exzerpte von bis zu 2000 Wörtern aus 155 Texten unterschiedlicher Textgenres entnommen. Das Brown-Korpus etablierte sich als Standard und wurde in vielen computerlinguistischen Arbeiten genutzt. Für das britische Englisch wurde nach den selben Kriterien das Lancaster-Oslo/Bergen (LOB) Corpus erstellt und für das Deutsche das LIMAS-Korpus.

Die **Sprachenauswahl** bezieht sich auf die Sprache der Primärdaten. **Monolinguale Korpora** enthalten nur eine Sprache, **bi- und multilinguale Korpora** zwei oder mehrere Sprachen. Handelt es sich um Quelltexte einer Sprache und deren Übersetzungen in eine oder mehrere andere Sprachen, spricht man von **Parallelkorpora**. Mehrsprachige Sammlungen zu vergleichbaren Diskursbereichen, bei denen die Texte keine unmittelbaren Übersetzungen von einander sind, werden als **Vergleichskorpora** bezeichnet. In den Übersetzungswissenschaften wird der Begriff *Vergleichskorpus* etwas anders verwendet. Er beschreibt dort monolinguale Korpora, welche sowohl Originaltexte als auch übersetzte Texte in derselben Sprache enthalten. Für computerlinguistische Anwendungen sind besonders solche bi- und multilingualen Parallelkorpora relevant, bei denen die parallelen Texte auf Paragraphen-, Satz-

oder Wortebene **aligniert** vorliegen, so dass die Texteinheiten der Übersetzung den jeweiligen Texteinheiten des Quelltexts zugeordnet werden. Ein häufig zitiertes Korpus ist das European Parliament Proceedings Parallel Corpus (kurz: **Europarl Corpus**, Koehn 2005), das auf Mitschriften und Übersetzungen von Debatten des Europäischen Parlaments beruht. Es umfasst Sprachpaare von elf europäischen Sprachen.

Ein weiteres Kriterium für die Klassifizierung von Korpora ist das **Medium**, in dem die Primärdaten entstanden bzw. erfasst wurden. Man unterscheidet Korpora der **geschriebenen Sprache**, Korpora der **gesprochenen Sprache** und **multimodale Korpora**. Bei multimodalen Korpora werden die Primärdaten mit verschiedenen Medien erfasst, oft werden Audio- mit Videoaufnahmen kombiniert, so dass auch non-verbale Kommunikationsaspekte ausgewertet werden können wie beim **Smartkom-Korpus** im **Bayerischen Archiv für Sprachsignale** (BAS), bei dem Gestik, Mimik und Augenbewegung mit einbezogen wurden, um verschiedene Interaktionen zwischen Mensch und Maschine zu untersuchen (Schiel *et al.* 2002). Die Einordnung eines Korpus in geschriebene oder gesprochene Sprache ist im Einzelfall nicht trivial. Ist die Aufnahme einer ausformulierten Ansprache ein Beleg für gesprochene Sprache? Sind E-Mails oder Protokolle aus Chat-Räumen wie im **Dortmunder CHAT-Korpus** Belege für geschriebene Sprache? Um die Daten angemessen beschreiben zu können, bedarf es einer detaillierteren Klassifikation, die nicht nur das Medium der sprachlichen Realisierung berücksichtigt, sondern auch deren konzeptuellen Hintergrund. Die Text Encoding Initiative sieht daher für die Angabe des Mediums in den Metadaten eines Korpus die Werte *spoken to be written* und *written to be spoken* vor. Die Transkription der Aufnahme einer ausformulierten Ansprache wäre demnach *written to be spoken*.

Die **Annotation** ist ein weiteres Unterscheidungskriterium. Eine Reihe von Korpora basieren auf den selben Sprachdaten und unterscheiden sich nur durch ihre Annotationsebenen. Das markanteste Beispiel dafür sind Daten aus dem **Wall Street Journal** (WSJ) Subkorpus, das einen Teil der **Penn Treebank** bildet (welche zusätzlich u. a. die Daten des Brown-Korpus beherbergt). Das WSJ-Subkorpus beinhaltet Annotation der Wortart, eine syntaktische Analyse, die von der Rektions- und Bindungstheorie (Chomsky 1981) inspiriert ist sowie die Angabe von grammatischen Funktionen (siehe auch Unterkapitel **??**). Die selben Daten wurden im **PropBank-Projekt** und im **NomBank-Projekt** mit Prädikat-Argumentstrukturen für Verben bzw. Nomen versehen. Teile davon sind auch in der **Penn Discourse Bank** und der **TimeBank** enthalten.

Korpora variieren stark in ihrer **Größe**. Neben vielen kleinen Korpora existieren langfristig angelegte Großprojekte. Die erste Generation digitaler Korpora wie das Brown Corpus beinhaltet eine Million Wortformen. Die zweite Generation, zu der das **British National Corpus** (BNC) gehört, umfasst bis zu 100 Millionen Wortformen. Korpora der dritten Generation gehen weit über die bisherigen Größenordnungen hinaus. Die aus dem Web extrahierten und automatisch aufbereiteten Korpora von **WaCky** (*Web as Corpus kool ynitiative*) beinhalten jeweils mehr als eine Milliarde Token. Die Texte sind automatisch vom HTML-Code und von Duplikaten bereinigt, segmentiert und POS-getaggt.

Das deutsche deWaCky-Korpus zum Beispiel hat eine Größe von 1 278 177 539 Token oder 25,9 GB (Baroni *et al.* 2009). Ein anderes Beispiel ist die ein Terabyte große Sammlung von Google (Brants and Franz 2006), auch wenn sie nach der eingangs genannten Definition kein Korpus im eigentlichen Sinn darstellt, weil sie nur eine Sammlung von Wort-Quintupeln mit Frequenzangaben ist und keine fortlaufenden Texte enthält. Für viele statistische Anwendungen in der Computerlinguistik sind diese Wortketten vollkommen ausreichend. Viele Algorithmen arbeiten sogar nur auf der Basis von Dreierketten (Trigrammen, siehe Unterkapitel **??**). Andere Megakorpora werden von Wörterbuchverlagen verwaltet. Das englische COBUILD-Korpus, genannt die **Bank of England**, mit über einer halben Milliarde Token, ist ein gemeinsames Produkt der Universität Birmingham und des Verlages Harper-Collins. Der Duden-Verlag pflegt ein deutschsprachiges Megakorpus mit mehr als 1,3 Milliarden Token, welches aber nicht frei verfügbar ist. Die größte deutschsprachige Korpussammlung findet sich am **Institut für Deutsche Sprache** in Mannheim (IDS). Dort stehen Korpora mit insgesamt mehr als zwei Milliarden Token zur Verfügung. Eine Teilmenge davon ist auch online durchsuchbar. Linguistisch nicht aufbereitet, aber frei verfügbar sind die **XML-Dumps** von Wikipedia, in denen einzelsprachliche Versionen der Internet-Enzyklopädie gespeichert werden.

Korpora unterscheiden sich in Bezug auf die **Persistenz** ihrer Daten. Nicht alle Korpora basieren auf einem **statischen Datensatz**. **Monitorkorpora** wie das Mannheimer Morgen-Korpus des Instituts für Deutsche Sprache oder die bereits genannte Bank of England wachsen permanent, weil immer neue Daten eingepflegt werden. Ein Monitorkorpus der anderen Art liegt der Belegsammlung Wortwarte zugrunde, in der seit dem Jahr 2000 Wortneubildungen dokumentiert werden. Das zugrundeliegende Korpus besteht aus dem täglichen Online-Angebot von Zeitschriften und wird aus urheberrechtlichen Gründen nicht gespeichert.

Das nächste Kriterium ist der **Sprachbezug**, der wieder stark mit dem ursprünglich intendierten Anwendungszweck zusammenhängt. Man unterscheidet **Referenzkorpora**, die versuchen, eine Sprache in ihrer Gesamtheit zu vertreten, wie das British National Corpus oder das **Kerncorpus** des Digitalen Wörterbuchs der deutschen Sprache (DWDS), von **Spezialkorpora**, die sich auf Sprachdaten bestimmter eingeschränkter Domänen beschränken. Um ein **ausgewogenes, repräsentatives Korpus** zu erhalten, werden sorgfältige **Designkriterien** (*sampling criteria*) entwickelt. Allerdings können auch Referenzkorpora immer nur eine Annäherung an eine Sprache sein, da man die Grundgesamtheit einer Sprache nicht wirklich erfassen kann. Welche Belege sollte man für ein Korpus der *deutschen Sprache* zusammenstellen? Umgangssprachliche und dialektale Äußerungen, offizielle Statements, Zeitungsartikel, Romane, Gesetzestexte, E-Mails und Chat-Konversationen, Lyrik, Texte aus dem 18. oder 19. Jahrhundert, die Bibelübersetzung von Martin Luther (als Beginn des Neuhochdeutschen)? Neben Spezial- und Referenzkorpora gibt es **opportunistischen** Sammlungen, bei denen aus pragmatischen Gründen auf Designkriterien verzichtet wird. Ein prominentes Beispiel dafür ist das bereits genannte Wall Street Journal Subkorpus der Penn Treebank.

1.3 Anwendungen

Allgemein sind linguistische Korpora eine wertvolle empirische Datenressource. In der (Computer-)Linguistik kommen sie auf verschiedene Arten zum Einsatz, von denen die gängisten kurz vorgestellt werden.

Die **Qualität** von computerlinguistischen Analyseprogrammen wird oft durch einen automatischen Abgleich mit manuell erstellten Referenzdaten, dem sogenannten **Goldstandard**, getestet. In industrienahen Projekten bezeichnet man das Referenzkorpus auch als Benchmark-Korpus. Je nach Anwendung werden unterschiedliche Qualitätsmaße verwendet. Oft handelt es sich um Varianten der ursprünglich aus dem Information Retrieval stammenden Maße **Präzision** (englisch *precision*) und **Abdeckung** (englisch *recall*). Um verschiedene Evaluationsergebnisse besser vergleichen zu können, wird in der Literatur oft der sogenannte **F-Wert** angegeben, der harmonische Mittelwert aus Präzision und Abdeckung.

Bei der **Entwicklung von korpusbasierten Programmen**, insbesondere beim Einsatz von maschinellen Lernverfahren, wird das Korpus dazu anfänglich geteilt. Ein Teil des Korpus wird als **Trainingskorpus** genutzt, auf dessen Basis zum Beispiel die Regeln des Programms und Wahrscheinlichkeiten (manuell oder automatisch) abgeleitet werden können. Ein zweiter Teil wird als **Entwicklungskorpus** eingesetzt, der zum Testen während der Programmerstellung und zur Erweiterung bzw. Verbesserung des Programms dient. Ein dritter und letzter Teil des Korpus bildet schließlich das **Testkorpus**. Diese Daten sollten während der Progammentwicklung nicht betrachtet werden, so dass das fertig gestellte Programm auf diesen bis dahin ungesehenen Daten objektiv getestet werden kann (die Testdaten werden auch als **Heldout Data** bezeichnet). Eine mögliche Teilung ist 80% Trainings-, 10% Entwicklungs- und 10% Testkorpus (Jurafsky and Martin 2008, S. 92). Eine alternative Teilungsmöglichkeit besteht in der **k-fachen Kreuzvalidierung** (englisch *k-fold cross validation*). Bei der zehnfachen Kreuzvalidierung (*ten-fold cross validation*) zum Beispiel wird das Korpus in zehn gleich große Teile geteilt. Das Programm wird auf neun der zehn Teile trainiert und auf dem zehnten getestet. Dies wird insgesamt zehnmal durchgeführt, so dass jedes Korpusteil einmal als Testkorpus zum Einsatz kommt. Als Ergebnis wird dann der Mittelwert aller zehn Evaluierungsergebnisse angegeben. Es wird davon abgeraten, bei der Korpusteilung zusammenhängende Textblöcke zu wählen. Man sollte besser eine zufällig gestreute Auswahl treffen und zum Beispiel nur jeden zehnten Satz für das Testkorpus extrahieren.

Neben der Entwicklung und dem Testen von korpusbasierten Programmen, dienen Korpora auch als Testbett zur **empirischen Untermauerung** von (computer)linguistischen Theorien. Generative Grammatiktheorien wie die Head Driven Phrase Structure Grammar (**HSPG**) und die Lexical-Functional Grammar (**LFG**) wurden in Parsern implementiert und an realen Korpusdaten getestet (Oepen *et al.* 2002, Zinsmeister *et al.* 2002).

In der **Lexikographie** spielen Korpora seit je her eine große Rolle, um Lesarten von Wörtern zu identifizieren und Beispiele für den Wortgebrauch

zu finden. "You shall know a word by the company it keeps", fasst Firth (1968, S. 179) den Ansatz des Kontextualismus zusammen, aus dem das für die Lexikographie wichtige Konzept der **Kollokation** stammt: das habituelle gemeinsame Auftreten von zwei oder mehreren Wörtern. Im Deutschen *putzt* man zum Beispiel *seine Nase*, während man sie im Englischen *bläst* (*blow your nose*). Kollokationen können über relative Auftretenshäufigkeiten in Korpora ermittelt werden, vgl. Unterkapitel **??**.

Im **Sprachunterricht** wurden Korpora traditionell nur indirekt genutzt zum Beispiel als Datenressource für die Erstellung von Unterrichtsmaterialien. Der unmittelbare Einsatz von Korpora im Unterricht ist eine neue Entwicklung (Mukherjee 2002, Nesselhauf 2004, Bick 2005).

In der Linguistik besteht eine lange Tradition der Verwendung von Korpora zum Beispiel in Teildisziplinen wie der historischen Linguistik und der Spracherwerbsforschung. In der Mitte des zwanzigsten Jahrhunderts grenzten sich theoretisch arbeitende Linguisten von den damals etablierten korpusbasierten Methoden ab und argumentierten, dass **Korpusdaten als empirische Evidenz** für linguistische Erkenntnis ungeeignet wären (Chomsky 1962, nach McEnery and Wilson 2001, S. 10). Zur Zeit erlebt die Verwendung von Korpora jedoch auch in der theoretischen Linguistik eine Renaissance (z. B. Bresnan *et al.* 2007). Es stehen dafür inzwischen leicht zugängliche Ressourcen zur Verfügung, die mit entsprechenden Such- und Analysewerkzeugen ausgewertet und visualisiert werden können (Baayen 2008, Johnson 2008, Gries 2008, 2009).

1.4 Weiterführende Informationen

Weiterführende Informationen zu allen Themen dieses Unterkapitels bietet das Handbuch Corpus Linguistics (Lüdeling and Kytö 2008). Die Corpora Mailing List (gandalf.aksis.uib.no/corpora/) hilft bei allen Fragen rund um Textkorpora und liefert Informationen zu Konferenzen und Veröffentlichungen. Es ist ratsam, das umfangreiche Archiv der Liste zu konsultieren, bevor man eine eigene Frage an die Listengemeinschaft stellt. Eine umfassende Linksammlung zu Korpora und Tools wird von David Lee gepflegt (devoted.to/corpora). Speziell an Computerlinguisten wendet sich die Sammlung der Stanford Natural Language Processing Group (www-nlp.stanford.edu/links/statnlp.html). Zusätzliche Verweise auch auf deutschsprachige Seiten findet man auf der Linksammlung des Lehrstuhls von Anke Lüdeling (www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links/). Die europäische Organisation Evaluations and Language Resources Distribution Agency (ELDA) veranstaltet alle zwei Jahre die International Conference on Language Resources and Evaluation (LREC). Die Special Interest Group for Annotation der Association for Computational Linguistics (ACL-SIGANN) führt in unregelmäßigen Abständen ebenfalls Workshops zum Thema durch. Zuletzt sei noch auf das Natural Language Toolkit verwiesen (www.nltk.org), ein Open Source-Projekt, das computerlinguistisch relevante Python-Module zusammenstellt. In das NLTK-Paket integriert ist eine Sammlung von Korpora mehrerer Sprachen, unter anderem Teile der syntaktisch annotierten englischen Penn Treebank.

References

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Koch, S., Maier, E., Maier, E., Reithinger, N., Schmitz, B., and Schmitz, B. (1998). Dialogue Acts in VERBMOBIL-2. verbmobil-report 226. Technical report, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes. Second Edition.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC MapTask Corpus. *Language and Speech*, **34**(4), 351–366.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge University Press, Cambridge.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta., E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*. Online-First: 10.2.2009, http://www.springerlink.com.

Bick, E. (2005). Grammar for fun: It-based grammar learning with visl. In P. J. Henriksen, editor, *CALL for the Nordic Languages*, Samfundslitteratur (Copenhagen Studies in Language), pages 49–64, Copenhagen.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER Teebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT)*, pages 24–41, Sozopol, Bulgaria.

Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. (2007). Predicting the Dative Alternation. In G. Bouma, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., and Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 969–974, Genoa, Italy.

Burger, S., Weilhammer, K., Schiel, F., and Tillmann, H. (2000). Verbmobil Data Collection and Annotations. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 539–551. Springer, Berlin.

Calhoun, S., Nissim, M., Steedman, M., and Brenier, J. (2005). A framework for annotating information structure in discourse. In A. Meyers, editor, *Proceedings of the ACL'05 Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddo, G., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, **23**, 13–31.

Chomsky, N. (1962). A Transformational Approach to Syntax. In Hill, editor, *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English on May 9-12, 1958*, pages 124–158, Texas. (Reprinted in Structure of Language, edited by Fodor and Katz. New York: Prentice-Hall, 1964; reprinted as "Une Conception Transformationelle de la Syntaxe." Language 4 (December 4, 1966): 39-80; Reprinted in Classics in Linguistics, edited by Hayden, Alworth and Tate, 337-71. New York: Philosophical Library, 1967).

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter.

Dipper, S. (2005). XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation Schema. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Firth, J. R. (1968). A synopsis of Linguistic Theory. In *Selected Papers of J.R. Firth, 1952-1959*, pages 168–205. F.R. Palmer, London.

Foster, M. E. (2007). Associating facial displays with syntactic constituents for generation. In *Proceedings of the Linguistic Annotation Workshop*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

Foth, K. (2006). Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Technical report, Universität Hamburg, Hamburg.

Francis, W. and Kučera, H. (1979). Brown Corpus Manual – Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. revised edition, Brown University, http://khnt.hit.uib.no/icame/manuals/brown.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., and Stoel, R. (2007). Information Structure. In S. Dipper, M. Götze, and S. Skopeteas, editors, *Information Structure in Cross-Linguistic Corpora*, number 07 in Interdisciplinary Studies on Information Structure (ISIS), pages 147–187.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. P.-T. Sylviane Granger, Joseph Hung, editor, *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pages 3–33. John Benjamins, Amsterdam / Philadelphia.

Gries, S. (2008). *Statistik für Sprachwissenschaftler*. Number 13 in Studienbücher zur Linguistik. Vandenhoeck & Ruprecht, Göttingen.

Gries, S. (2009). *Quantitative corpus linguistics with R: a practical introduction*. Routledge, Taylor & Francis Group, London, New York.

Hajičová, E., Kirschner, Z., and Sgall, P. (1999). A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V., and Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 552–576. Springer, Berlin.

Jekat, S. and v. Hahn, W. (2000). Multilingual Verbmobil-dialogs: Experiments, data collection and data analysis. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 577–584. Springer, Berlin.

Johnson, K. (2008). *Quantitative Methods in Linguistics*. Blackwell Publishing, Malden / Oxford / Victoria.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2nd edition.

Kipp, M., Neff, M., and Albrecht, I. (2007). An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation*, **41**(3), 325–339.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86.

Lebert, M. (2008). Project Gutenberg (1971-2008). http://www.gutenberg.org/etext/27045.

Leech, G., Wilson, A., *et al.* (1996). EAGLES Guidelines: Recommendations for the Morphosyntactic Annotation of Corpora. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

Lemnitzer, L. and Zinsmeister, H. (2006). *Korpuslinguistik. Eine Einführung*. narr studienbücher. Narr, Tübingen.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In P. Grommes and M. Walter, editors, *Fortgeschrittene Lernervarietäten*, pages 119–140. Niemeyer, Tübingen.

Lüdeling, A. and Kytö, M., editors (2008). *Corpus Linguistics. An International Handbook*. Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1. Mouton de Gruyter, Berlin/New York.

MacWhinney, B. (1995). *The CHILDES-Project: Tools for Analyzing Talk*. Erlbaum, Hillsdale, NJ, 2nd edition.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structures. In *ARPA Human Language Technology Workshop*, pages 114–119, San Francisco. Morgan Kaufmann.

Martell, C. (2002). FORM: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of ICSLP*, pages 353–356.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*. Edinburgh University Press, Edinburgh, 2nd edition.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC-2004*, pages 803–806, Lisbon, Portugal.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber., B. (2004). Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, MA.

Mukherjee, J. (2002). *Korpuslinguistik und Englischunterricht: Eine Einführung*. Peter Lang, Frankfurt am Main.

Naumann, K. (2006). Manual of the annotation of in-document referential relations. http://www.sfs.uni-tuebingen.de/resources/tuebadz_relations_man.pdf.

Nesselhauf, N. (2004). Learner Corpora and their Potential for Language Teaching. In J. Sinclair, editor, *How to use corpura in Language Teaching*, pages 125–152. John Benjamins, Amsterdam.

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An annotation scheme for information status in dialogue. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC2004)*, Lisbon.

Oepen, S., Toutanova, K., Shiebe, S., Manning, C., Flickinger, D., and Brants, T. (2002). The LinGO Redwoods treebank: Motivation and preliminary applications. In *In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1253–1257, Taipei.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics*, **31**(1), 71–106.

Poesio, M. (2000). The GNOME Annotation Scheme Manual. http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*, pages 647–656.

Riester, A. (2008). A semantic explication of *Information Status* and the underspecification of the recipients' knowledge. In A. Grønn, editor, *Proceedings of SuB-12*, pages 508–522, Oslo.

Ruoff, A. (1984). *Alltagstexte I. Transkriptionen von Tonbandaufnahmen aus Baden-Württemberg und Bayrisch-Schwaben mit zwei Karten.* IDIOMATICA 10. Veröffentlichungen der Tübinger Arbeitsstelle "Sprache in Südwestdeutschland". Niemeyer, Tübingen.

Schiel, F., Steininger, S., and Türk, U. (2002). The smartkom multimodal corpus at bas. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC2002)*, pages 200–206.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung, Stuttgart.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan.

TEI AI1W2 (1991). List of Common Morphological Features for Inclusion in TEI Starter Set of Grammatical-Annotation Tags. http://www.w3.org/People/cmsmcq/1991/ai1w02.html.

Telljohann, H., Hinrichs, E., Kübler, S., and Zinsmeister, H. (2006). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Universität Tübingen. http://www.sfs.uni-tuebingen.de/resources/sty.pdf.

Trawinski, B., Söhn, J.-P., Sailer, M., and Richter, F. (2008). A multilingual electronic database of distributionally idiosyncratic items. In E. Bernal and J. DeCesaris, editors, *Proceedings of the XIII Euralex International Congress*, volume 20 of *Activitats*, pages 1445–1451, Barcelona, Spain.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, **30**(3), 277–308.

Zinsmeister, H., Kuhn, J., and Dipper, S. (2002). TIGER TRANSFER – Utilizing LFG Parses for Treebank Annotations. In M. Butt and T. Holloway King, editors, *Proceedings der LFG02 Conference*, pages 427–447, Athens. CSLI Publications.

Zinsmeister, H., Witt, A., Kübler, S., and Hinrichs, E. (2008). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In A. L. üdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1, chapter 37. Mouton de Gruyter, Berlin/New York.