## Korpuslinguistik

Die Korpuslinguistik ist eine sprachwissenschaftliche Arbeitsweise und umfasst verschiedene Tätigkeiten, die mit der **Erstellung** und der **Auswertung** von **Textkorpora** (Köhler 2005) sowie mit der Erstellung von **Werkzeugen** für die beiden ersten Tätigkeiten zu tun haben. McEnery (2003) (Hauptquelle) definiert ein Textkorpus als "**a large body of linguistic evidence**". Damit sind Korpora eine sehr wertvolle Quelle für die linguistische Arbeit.

Korpuslinguistische Methoden werden in verschiedenen **Bereichen** der Sprachwissenschaft verwendet: etwa in der allgemeinen Sprachwissenschaft, der Computerlinguistik und dem *Natural Language Processing* (NLP). Somit ist Korpuslinguistik kein eigener Teilbereich der Linguistik sondern eine **Arbeitsweise**, die innerhalb der genannten Teilbereiche verwendet wird und auf verschiedene **Ebenen** der Sprache (wie Morphologie, Syntax, Semantik oder Pragmatik) angewendet werden kann (McEnery & Wilson 1996 : 2).

Korpuslinguistik ist eine Form von Evidenz- oder **datenbasierter** (Köhler 2005) Linguistik. Im Grunde ist alle Linguistik vor Chomsky datenbasiert. McEnery & Wilson (1996) bezeichnen diese dementsprechend als **frühe Korpuslinguistik**. Eine Besonderheit der Arbeit mit Korpora im Gegesatz zu anderen Formen linguistischer Evidenz ist, dass nicht nur empirisch untersucht werden kann, ob eine Konstruktion möglich ist, sondern auch, **wie häufig** diese auftritt (Kennedy 1998 : 8).

## Korpora

Korpora können aus **verschiedenen Quellen** stammen, etwa aus aufgenommener Konversation (gesprochener Teil des British National Corpus BNC), Radionachrichten (IBM/Lancaster Spoken English Corpus) oder schriftlichen Veröffentlichungen (schriftlicher Teil des BNC).

Korpora sind durch verschiedene Eigenschaften gekennzeichnet. Zunächst sind Korpora heute typischerweise **maschinenlesbar**. Da Korpora nicht nur in den oben genannten Bereichen der Linguistik sowie etwa zur literarischen Stilanalyse eingesetzt werden, handelt es sich bei Korpora um eine **multifunktionale** Ressource (McEnery 2003). Ein Korpus ist darüber hinaus im Hinblick auf eine bestimmte Fragestellung **organisiert**. Diese Ausrichtung auf eine bestimmte Fragestellung bezeichnet McEnery (2003) als *sampling frame*, innerhalb dessen das Korpus **ausgewogen** und **repräsentativ** sein soll.

Als **Beispiel** nennt McEnery (2003) die Entwicklung eines Dialogsystems zum Verkauf von Eintritts- und Fahrkarten. Wenn hierfür ein Korpus zusammengestellt werden soll, sollten zum einen nur relevante Texte verwendet werden, d.h. Dialoge von Kartenverkäufen (dies entspricht dem gewählten ***sampling frame***). Es sollten jedoch verschiedene Arten von Verkaufsgesprächen verwendet werden, etwa von Bus- und Flugzeugtickets sowie von Telefon- und Schalterverkäufen. Ausserdem sollten in jedem Bereich Gespräche verschiedener Personen verwendet werden, um Idiosynkrasien einzelner Sprecher zu vermeiden. So erhält man ein **ausgewogenes** und **repräsentatives** Korpus.

## Korpusarten

McEnery (2003) unterscheidet verschiedene Arten von Korpora: **Monolinguale** Korpora sind Korpora mit Texten in einer einzigen Sprache. **Vergleichbare** Korpora sind Korpora aus Texten in verschiedenen Sprachen, die in Bezug auf *sampling frame*, Ausgewogenheit und Repräsentativität vergleichbar sind. Diese Korpora eignen sich etwa für den kontrastiven Sprachvergleich. **Parallele** Korpora sind Korpora, die zunächst aus Texten einer einzigen Sprache zusammengestellt und dann in andere Sprachen übersetzt werden. Solche Korpora eignen sich etwa für Systeme zur maschinellen Übersetzung, die aus Beispielübersetzungen lernen. **Monitorkorpora** werden laufend aktualisiert und dienen etwa zur Beobachtung von Sprachwandel (z.B. *Bank of English*).

Neben diesen Unterscheidungen können gesprochene von reinen Schriftkorpora unterschieden werden. **Gesprochene** Korpora (die monolingual, vergleichbar oder parallel sowie Monitorkorpora sein können) enthalten Informationen über die gesprochene Form des Textes. Dies kann bedeuten, dass es sich um **akustisches** Material handelt (solche Korpora sind schwerer zu verwenden, etwa zur Suche nach bestimmten Wörtern), oder auch um **transkribierte** Sprachdaten, wie im Fall des gesprochenen Teils des BNC (was u.a. zum Verlust prosodischer Feinheiten führt). Da diese beiden Formen Vor- und Nachteile haben, gibt es zunehmend Korpora, die aus akustischem und aus transkribiertem Material bestehen, deren Inhalte aufeinander abgestimmt sind, wodurch es möglich ist, auf die jeweils andere Form zuzugreifen (McEnery 2003). Solche **multimodale** Korpora (Evert & Fitschen 2001) können darüber hinaus auch Informationen über Mimik oder Gestik enthalten, etwa in Form von Videomaterial.

## Geschichte

Eine frühe Form der Korpuslinguistik sind etwa Bibelkonkordanzen, die alle Wörter der Bibel alphabetisch sortiert in ihrem Kontext auflisten und ab dem **13. Jahrhundert** entstanden. Käding erstellte und beschrieb **1897** ein großes, von Hand zusammengestelltes Korpus mit einem Umfang von 11 Mio. Wörtern (McEnery & Wilson 1996 : 3). Seit den **1920ern** gab es insbesondere in den USA und im UK eine Tradition des Wortzählens, um die häufigsten Wörter zu finden; die Anwendung lag hier vor allem im Bereich des Sprachlernens. In den **1930ern** wurden von Vertretern der Prager Schule quantitative Untersuchungen etwa zur Häufigkeit von bestimmten grammatikalischen Konstruktionen unternommen (Kennedy 1998 : 10).

In der heutigen, computerisierten Form gibt es die Korpuslinguistik seit den späten **1940er** Jahren (McEnery 2003). Das Beispiel der Bibelkonkordanzen zeigt den radikalen Wandel durch die Entwicklung des Computers: Während das Erstellen der ersten Bibelkonkordanz 14 Jahre dauerte,[1] ist es durch die **Entwicklung des Computers** möglich geworden, innerhalb von Sekundenbruchteilen Konkordanzen beliebiger Texte zu erstellen; das Erstellen eines solchen Programms selbst ist innerhalb von Tagen oder Stunden möglich. Aufgrund dieser Unterschiede hat die Erfindung des Computers die Korpuslinguistik im heutigen Sinn erst ermöglicht.

In den **1950ern** wurde das erste *vergleichbare* Korpus zusammengestellt, das auch schon den Grundlagen der *sampling frames*, der Ausgewogenheit und der Repräsentativität entsprach. In den **1980ern** wuchsen Anzahl und Größe von Korpora, was sich in den **1990ern** fortsetzte, als das BNC und die *Bank of English* Größen von 100 Mio. und 300 Mio. Wörtern erreichten. Zudem wuchs in den 1990ern die Anzahl paralleler Korpora. Köhler (2005) erwähnt eine "lange, vor allem europäische Tradition" der quantitativen Analyse empirischer Daten in der zweiten Hälfte des 20. Jahrhunderts, die jedoch aufgrund der dominierenden "**formalen, Kompetenz-orientierten**" Linguistik in den USA kaum rezipiert wurde. Erst seit einigen Jahren sei die Korpuslinguistik auch in den USA auf dem Vormarsch und wirke von dort auch wieder auf Europa zurück. McEnery (2003) ist der Ansicht, dass mit der **fortschreitenden Entwicklung** in der Korpuslinguistik zukünftig auch Probleme lösbar werden, die heute mit Mitteln der Korpuslinguistik nicht lösbar sind.

## Korpusannotation

Annotierte Korpora sind Korpora, die mit verschiedenen Arten linguistischer Information **angereichert** sind (McEnery & Wilson 1996 : 24). Ein Beispiel für Korpusannotationen ist etwa die Kennzeichnung von **Wortarten** (Part-of-Speech-Tags, POS-Tags). Andere mögliche Annotationen enthalten etwa Informationen über **Stammformen** (Stemming oder Lemmatisierung), **syntaktische** Struktur (solche Korpora werden auch Baumbanken genannt) sowie **semantische**, **stilistische** oder Informationen zur **Diskursstruktur**[2] (in abnehmender Häufigkeit und Verbreitung). Diese Anreicherung basiert immer auf einer bestimmten **Interpretation** der Daten. Diese stellt kein Hinzufügen neuer informationen dar, sondern macht lediglich **implizit vorhandene Informationen explizit** (McEnery 2003).

[1] Quelle: http://de.wikipedia.org/wiki/Bibelkonkordanz

[2] Beispiele verschiedener Korpusannotationen und weiteres Material gibt es unter: http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm

McEnery (2003) nennt vier **Vorteile** von Korpusannotationen: Zunächst eine **verbesserte Nutzbarkeit** der Korpora für eine größere Anzahl von Benutzern, seien dies Menschen, die einer bestimmten Fremdsprache unkundig sind oder Computerprogramme, die Sprache generell nicht verstehen können. Hier werden die Annotationen benötigt, können jedoch von dem, der sie benötigt nicht selbst erstellt werden. Darüber hinaus ist ein annotiertes Korpus auch für Benutzer, die die Analysen selbst vornehmen könnten viel schneller zu verwenden als ein nicht-annotiertes Korpus. Ein zweiter Vorteil ist die **Wiederverwertbarkeit** der bei der Analyse gewonnen Daten, etwa der ermittelten POS-Tags, der Stammformen, syntaktischen Strukturen etc. Als weiteren Vorteil nennt McEnery (2003) die **Multifunktionalität** der Daten, so kann bei einer Wiederverwertung ein ganz anderer Zweck verfolgt werden. Schließlich ist die explizite Formulierung und objektive **Erfassung von Ergebnissen** einer Analyse ein weiterer Vorteil.

Korpusannotationen können automatisch, von Hand oder in einer kombinierten Form erstellt werden: Für Aufgaben wie das POS-Tagging oder Lemmatisieren für Sprachen wie Englisch, Französisch und Spanisch können Annotationen **automatisch** erstellt werden. Häufig aber wird das Ergebnis der maschinellen Annotation von Hand kontrolliert und erfolgt damit **semi-automatisch**, was immer noch ein schnelleres Annotieren als allein von Hand ermöglicht. Einige Bereiche erfordern aber auch rein **manuelle** Erstellung der Annotationen, etwa zur Darstellung anaphorischer und cataphorischer Relationen (McEnery 2003).

Als **Argument gegen die manuelle oder semi-automatische Annotation** ist vorgebracht worden, dass diese nur eine geringe Konsistenz aufweisen könne, da menschliche Annotatoren vergleichbare Stellen im Korpus nicht immer übereinstimmend annotieren. Es wurden jedoch Untersuchungen unternommen, die gezeigt haben, dass bei geschulten Annotatoren der leichte Rückgang der Konsistenz mehr als kompensiert wurde durch eine gesteigerte Genauigkeit der Annotationen (McEnery 2003 : 457).

## NLP und Quantitative Linguistik

McEnery (2003) bezeichnet Textkorpora als “the raw fuel of NLP” und damit als Grundlage und Voraussetzung des NLP. Der Grund hierfür besteht darin, dass annotierte Korpora es Coputerprogrammen ermöglichen, die Intuitionen von Experten, die die Annotationen erstellt oder verbessert haben, in Bezug zu den Texten zu setzen und so menschliche Intuitionen zu reproduzieren. Ein Beispiel wäre etwa ein POS-Tagger, der aus annotierten Korpora lernt und dadurch in die Lage versetzt wird, neue Texte zu annotieren. So bilden annotierte Korpora die Grundlage für **maschinelles Lernen** im NLP-Bereich (McEnery 2003 : 459).

Ein weiterer wichtiger Einsatzbereich für Korpora bildet die gemeinsame **Evaluation** von NLP-Systemen durch die Verwendung eines gemeinsamen Korpus, wie etwa bei den *Message Understandig Conferences*. Im Bereich des NLP und der Computerlinguistik gibt es **zahlreiche Bereiche**, die Verbindungen zur Korpuslinguistik aufweisen oder als Teilbereich der Korpuslinguistik verstanden werden können. Dies umfasst etwa die großen Teilbereiche *Information Retrieval*, *Text Data Mining* bzw. *Text Mining* (Mustersuche in Texten) und verwandte Gebiete wie die Informationsextraktion. Hier werden Korpora nicht als linguistische Evidenz zur Theoriebildung verwendet, sondern **pragmatisch** als Grundlage für verschiedene Problemlösungen in der Sprachverarbeitung.

Da ein großer Teil der Auswertung von Korpora mithilfe von statistischen Verfahren erfolgt, kann diese Art der Arbeit mit Korpora auch als **Quantitative Linguistik** charakterisiert werden (Köhler 2005), einem eigenständigen Bereich des NLP (siehe etwa Manning & Schütze 1999). Köhler (2005) betont dabei die Notwendigkeit, linguistische Fragestellungen in die Sprache der **Statistik** zu überführen und Ergebnisse in die Sprache der Linguistik zurückzuübersetzen und beklagt ein bislang zu wenig ausgeprägtes Methodenbewusstsein im Bereich statistischer korpuslinguistischer Arbeit.

Architekturen zur Erstellung und Verwendung von korpuslinguistischen Werkzeugen, auch SALE genannt (**Software Architecture for Language Engineering**, siehe auch Cunningham & Bontcheva 2006 und Köhler 2005), sind etwa UIMA oder GATE. An der Abteilung für Sprachliche Informationsverarbeitung[3] hier am Institut wird unter dem Namen Tesla (Text Engineering Software Laboratory) ein vergleichbares System

[3] http://www.spinfo.uni-koeln.de

entwickelt. Schwerpunkt der Arbeit bilden hier Wiederverwertbarkeit von Analysekomponenten (z.B. Tokenisierung) und Ergebnissen sowie die Verteilbarkeit der Analysearbeit zur Durchführung rechenaufwändiger Verfahren.

## Standards

Zentrale **technische Aspekte** bei der Korpuserstellung sind **Zeichenkodierung** und **Dateiformat** der Daten. In den vergangenen Jahre haben sich **Unicode** als Standard zur Zeichenkodierung (hiermit können alle Zeichen aller bekannten Schrift- und Zeichensysteme eindeutig dargestellt werden) sowie SGML oder dessen Nachfolger **XML** als Standard-Dateiformat etabliert, z.B. im *Corpus Encoding Standard* (CES) und der *Text Encoding Initiative* (TEI). Intern werden auch binäre Formate verwendet, etwa zur Speicherung eines erstellten Index zum effizienteren Zugriff auf die Daten. Detailliertere Informationen zu technischen Aspekten der Korpuserstellung finden sich etwa in Evert & Fitschen (2001).

## Korpusabfrage

Man kann drei Arten der **Abfrage** von Korpora unterscheiden: Die **Konkordanzsuche**, die Wörter in ihrem Kontext zeigt (*key word in context*, KWIC), **musterbasierte** Suche, etwa mit regulären Ausdrücken, die eine Suche nach bestimmten Mustern ermöglichen, in einem Korpus mit POS-Tags etwa einfache Nominalphrasen des Musters

(DET)? ( (ADV)? ADJ )* NN

sowie **statistische** Suche, die etwa die Bestimmung von Häufigkeiten bestimmter Folgen von Wortformen (Kollokationen, etwa zum Vergleich der Häufigkeiten von *different from*, *different to* und *different than*) oder Wortarten (Kolligationen) ermöglicht (Evert & Fitschen 2001 : 376 und Kennedy 1998 : 11).

Einen Überblick über verschiedene Korpora gibt es bei der *European Language Resources Association* oder dem *Linguistic Data Consortium*. Einige relevante englischsprachige **Korpora** sind etwa das *British National Corpus*, das *IBM/Lancaster Spoken English Corpus*, das *Leverhulme Corpus of Children's Writing*, das *Survey of English Dialects* (McEnery 2003) sowie das *Brown Corpus*. **Abfragewerkzeuge** für das BNC sind etwa *Sara* zur Konkordanzsuche oder *BNCweb* zur statistischen Suche (Evert & Fitschen 2001).

## Literatur


Cunningham, H. & K. Bontcheva: 2006, 'Computational Language Systems, Architectures', in K. Brown, A. H. Anderson, L. Bauer, M. Berns, G. Hirst & J. Miller (eds.), *The Encyclopedia of Language and Linguistics*, second edn., Elsevier, München.

Evert, S. & A. Fitschen: 2001, 'Textkorpora', in K. U. Carstensen, C. Ebert, E. Endriss, S. Jekat, R. Klabunde & H. Langer (eds.), *Computerlinguistik und Sprachtechnologie*, Spektrum, Heidelberg, Berlin, pp. 369–376.

Kennedy, G.: 1998, *An Introduction to Corpus Linguistics*, Longman, London, New York.

Köhler, R.: 2005, 'Korpuslinguistik - zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven', *GLDV-Journal for Computational Linguistics and Language Technology* **20**(2), 1–16.

Manning, C. D. & H. Schütze: 1999, *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, USA.

McEnery, T.: 2003, 'Corpus Linguistics', in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, Oxford University Press, Oxford, pp. 448–463.

McEnery, T. & A. Wilson: 1996, *Corpus Linguistics*, Edinburgh University Press, Edinburgh.