Melanie Andresen

Computerlinguistische Methoden für die Digital Humanities

Eine Einführung für Geisteswissenschaftler:innen

narr STUDIENBÜCHER

narr\f
ranck
e\atte
mpto

Dr. Melanie Andresen hat über neun Jahre an den Universitäten Hamburg und Stuttgart in der Linguistik, Computerlinguistik und den Digital Humanities gelehrt und geforscht. Seit 2024 arbeitet sie bei DeepL an der Verbesserung maschineller Übersetzung.

narr **STUDIENBÜCHER**

Melanie Andresen

Computerlinguistische Methoden für die Digital Humanities

Eine Einführung für Geisteswissenschaftler:innen

narr\f
ranck
e\atte
mpto

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

DOI: https://doi.org/10.24053/9783823395799

Internet: www.narr.de
eMail: info@narr.de

CPI books GmbH, Leck

ISSN 0941-8105
ISBN 978-3-8233-8579-0 (Print)
ISBN 978-3-8233-9579-9 (ePDF)
ISBN 978-3-8233-0505-7 (ePub)

Inhalt

Vorwort

Dieses Buch basiert auf der Lehrveranstaltung „Computerlinguistische Methoden für die Digital Humanities“, die ich in den Wintersemestern 2020/21 bis 2023/24 an der Universität Stuttgart im Masterstudiengang „Digital Humanities“ unterrichtet habe. Ich danke allen Studierenden dieser Lehrveranstaltung ganz herzlich für ihre hochmotivierte Teilnahme, wertvolle Rückmeldungen und die zahlreichen Impulse aus ihren geisteswissenschaftlichen Disziplinen, die den Austausch in der Lehrveranstaltung wie auch in den Digital Humanities im Ganzen für mich so facettenreich und spannend machen.

Mein Dank gilt meinem Lektor Tillmann Bub, der mir genau zum richtigen Zeitpunkt den Anstoß gegeben hat, dieses lange erwogene Projekt tatsächlich in die Tat umzusetzen. Jonas Kuhn danke ich herzlich dafür, dass er mir die Umsetzung ermöglicht hat.

Für ihre Zeit zum Korrekturlesen, ihr hilfreiches Feedback und gute Gespräche danke ich (in alphabethischer Reihenfolge) Johanna Binnewitt, André Blessing, Lisa Dücker, Agnieszka Faleńska, Lina Franken, Sarah Ihden, Sarah Jablotschkin, Nora Ketschik, Roman Klinger, Janis Pagel, Axel Pichler, Nils Reiter, Evelyne Roth, Michael Roth, Nadja Schauffler, Eleonore Schmitt, Lena Schnee, Carla Sökefeld, Anna Tilmans und Michael Vauth.

Stuttgart, im Februar 2024 Melanie Andresen

1 Einleitung

> In diesem Kapitel werden Inhalt und Aufbau dieses Buches vorgestellt. Wir klären außerdem, was genau die Computerlinguistik ist und welche Gemeinsamkeiten und Unterschiede zwischen der Computerlinguistik und dem eng verwandten Gebiet der Korpuslinguistik bestehen. Zuletzt führen wir die Grundbegriffe Korpus, Metadaten und Annotation ein, die im ganzen Buch zentral sind.

1.1 Über dieses Buch

Computerlinguistische Methoden durchdringen heute unseren Alltag: Wir stellen Anfragen an Suchmaschinen, die ermitteln, welche Webseiten am besten zu unserem Anliegen passen. Wir nutzen automatische Übersetzer, damit wir uns bei einer geschäftlichen E-Mail auf Englisch nicht allein auf unser Sprachgefühl verlassen müssen. Unser Textverarbeitungsprogramm korrigiert unsere Rechtschreibfehler. Das Tippen von Textnachrichten auf dem Handy wird dadurch erleichtert, dass uns jederzeit die wahrscheinlichsten nächsten Wörter vorgeschlagen werden. Bei Bedarf können wir unser Smartphone auch mündlich beauftragen, den Wecker für morgen früh zu stellen oder Mama anzurufen. Und vielleicht fragen wir Chatbots nach den richtigen Antworten für die heutigen Hausaufgaben oder lassen sie ganze Essays für uns schreiben.

Auch für die geisteswissenschaftliche Textanalyse bieten computerlinguistische Methoden ein großes Potenzial. Sie ermöglichen uns die Auswertung von Textmengen, die mit manuellen Methoden nicht realistisch bearbeitet werden können. Denn auch in den Geisteswissenschaften stehen uns immer größere Datenmengen zur Verfügung, die wir nicht mehr manuell sichten können. Stattdessen sind wir auf das sog. Distant Reading angewiesen, d. h. die computerbasierte Erschließung großer Textmengen. Die Computerlinguistik bietet uns zudem neue, datengeleitete Zugänge zu unseren Gegenständen. Dies ist insbesondere bei explorativen Fragestellungen hilfreich, wenn wir unsere Daten zunächst erschließen und nicht direkt eine bestimmte, aus der Theorie abgeleitete Hypothese prüfen wollen. Computerlinguistische Methoden ergänzen den traditionellen geisteswissenschaftlichen Blick auf Texte um Quantifizierungen, die unter anderem präzise Vergleiche und die Anwendung statistischer Methoden ermöglichen. Ein Teilschritt der Analyse wird dadurch reproduzierbar, auch wenn die Interpretation der Daten am Ende in der Regel uns Menschen und unserer subjektiven Perspektive überlassen bleibt.

Dieses Buch richtet sich an alle, die Interesse an der Anwendung computerlinguistischer Methoden auf geisteswissenschaftliche Fragestellungen und an der Reflexion ihrer Potenziale haben. Es setzt kein linguistisches, technisches oder mathematisches Vorwissen voraus und bietet dadurch einen niedrigschwelligen Einstieg in ein span-

nendes und interdisziplinäres Forschungsfeld, das an der Schnittstelle von ganz unterschiedlichen, textbasiert arbeitenden Geisteswissenschaften und der Computerlinguistik liegt.

Die Computerlinguistik hat sich in den letzten Jahrzehnten methodisch massiv verändert.[1] Frühe Ansätze haben vor allem menschliche Expert:innen genutzt, die ihr Wissen über den Gegenstand in maschinell lesbare Regeln übersetzt haben, die der Computer dann anwenden konnte. Für den Anwendungsfall der Spamerkennung in E-Mails ließe sich beispielweise als Regel formulieren, dass das Wort *gratis* im Betreff möglicherweise auf eine Spamnachricht hinweist und diese dann entsprechend behandelt wird. Durch die stark gestiegene (und weiterhin steigende) Verfügbarkeit von Sprachdaten und Rechenkapazitäten zu ihrer Verarbeitung setzen die meisten Ansätze der Gegenwart auf statistische Verfahren des maschinellen Lernens und Deep Learnings. Hierbei gibt es keine von Menschen formulierten Regeln. Stattdessen muss eine ausreichende Menge bereits korrekt klassifizierter Daten zum Training zur Verfügung stehen, anhand derer der Algorithmus die (teilweise sehr komplexen) Zusammenhänge zwischen den Merkmalen der sprachlichen Oberfläche und den Zielkategorien ermittelt.

Für die Anschlussfähigkeit computerlinguistischer Methoden an die Geisteswissenschaften stellen sich durch diese Entwicklung ganz neue Fragen. Insbesondere die Interpretierbarkeit der automatischen Analyse und ihrer Ergebnisse ist ein entscheidender Faktor für die Einsatzfähigkeit computerlinguistischer Modelle in den Geisteswissenschaften. Für die geisteswissenschaftlichen Erkenntnisinteressen ist es in der Regel nicht ausreichend, zum Beispiel die Unterscheidung zwischen zwei Gruppen von Texten erfolgreich automatisieren zu können. Stattdessen wollen wir durch die Analyse vor allem etwas über unseren Gegenstand lernen. Die erfolgreichsten Methoden der Computerlinguistik sind deshalb nicht unbedingt auch die mit dem größten Potenzial für die Geisteswissenschaften.

In dieser Einführung werden deshalb zwei Strategien verfolgt: Erstens liegt ein Schwerpunkt auf Methoden, die mit linguistischen Grundlagen in Verbindung stehen und sich durch gute Nachvollziehbarkeit durch den Menschen auszeichnen. Diese entsprechen aus computerlinguistischer Perspektive nicht immer dem allerneusten Stand der Technik, sind für geisteswissenschaftliche Fragestellungen aber vielfach geeigneter. Zweitens wird mit dem maschinellen Lernen und den künstlichen neuronalen Netzen in die aktuellen Methoden der Computerlinguistik einführt. Schließlich ist auch für manche geisteswissenschaftlichen Anliegen vor allem die erfolgreiche Automatisierung das Ziel. Der computerlinguistische Stand der Technik ist in einem sehr zügigen Wandel begriffen und mag sich zum Zeitpunkt der Veröffentlichung dieses Buches bereits weiterentwickelt haben, ohne dass die hier vermittelten Grundlagen ihre Gültigkeit verlieren würden.

---

1 Für einen detaillierteren Einblick in die Geschichte der Computerlinguistik siehe Menzel (2010) und Lobin (2010), einen aktuelleren Überblick über das Fach bietet Munro (2022).

Der Hauptteil dieses Buches ist in drei Teile gegliedert: Die Kapitel in **Teil I** gehen von linguistischen Beschreibungsebenen aus und stellen dar, welche computerlinguistischen Zugänge uns jeweils zu dieser Ebene von Sprache zur Verfügung stehen und wie wir damit praktisch arbeiten können. Im Rahmen der Lexik (Kapitel 2) geht es darum, was für den Computer (und für uns) ein Wort ist und wie wir den Wortschatz eines Korpus mit Methoden wie Kollokations- oder Keywordanalyse untersuchen können. Die Kapitel zu Wortarten (Kapitel 3) und Syntax (Kapitel 4) beschreiben, wie wir diese linguistischen Grundkategorien modellieren, manuell oder automatisch annotieren und für geisteswissenschaftliche Fragestellungen nutzen können. Im Gebiet der Semantik betrachten wir Wege, den Inhalt eines Korpus über Wortfelder zu erschließen (Kapitel 5), Möglichkeiten, im Rahmen der Sentimentanalyse Bewertungen oder Stimmungen zu erfassen (Kapitel 6) und mit den Konzepten der distributionellen Semantik, insbesondere den populären Word Embeddings, zu arbeiten (Kapitel 7). Im Bereich der Pragmatik blicken wir auf die Referenten von Texten und darauf, wie wir sie anhand von Named Entity Recognition und Koreferenzanalyse erfassen können (Kapitel 8). Am Ende jedes Kapitels zeigen Beispielstudien, welche Anwendungspotenziale sich aus den jeweiligen Methoden für die Digital Humanities ergeben. **Teil II** setzt einen methodischen Schwerpunkt quer zu den linguistischen Teilgebieten. Wir widmen uns der Frage, wie wir in Korpora nach Wörtern und Mustern suchen und die Ergebnisse durch statistische Kennzahlen und Visualisierungen präsentieren können (Kapitel 9). Kapitel 10 fokussiert die manuelle Annotation von Daten, die für viele Automatisierungen der wichtige erste Schritt ist. In zwei Kapiteln zum maschinellen Lernen (Kapitel 11) und spezifischer dem Deep Learning (Kapitel 12) geht es um Möglichkeiten der Automatisierung von Annotationen. In **Teil III** betrachten wir computerlinguistische Methoden im Kontext der Gesellschaft und widmen uns den ethischen Fragen, die bei der Anwendung computerlinguistischer Methoden berücksichtigt werden müssen (Kapitel 13).

Am Ende der meisten Kapitel gibt es Übungen, zu denen im digitalen Anhang des Buches Musterlösungen zur Verfügung stehen. Sie können im Online-Shop des Narr Verlags aufgerufen werden (https://files.narr.digital/9783823385790/Zusatzmaterial.zip). Zu manchen Aufgaben gehören außerdem Beispielskripte in *Python*, die ebenfalls im digitalen Anhang zu finden sind. Die Skripte sind so gestaltet, dass sie auch ohne fundierte Programmierkenntnisse ausprobiert werden können. Um die Skripte auf Ihrem eigenen Rechner ausführen zu können, müssen Sie eine möglichst aktuelle Version von → *Python 3* installiert haben. Zusätzlich empfiehlt sich eine (kostenlose) Programmierumgebung wie → *PyCharm* oder → *Visual Studio Code*. Mögliche Probleme bei der Installation und Einrichtung lassen sich in einem Buch nur schwer abdecken. Über die Suchmaschine Ihres Vertrauens finden Sie aber bei Bedarf zahlreiche Anleitungen in Text- und Videoform.

Alle Tools und Ressourcen, die mit einem Pfeil (→) versehen sind, finden Sie im Ressourcenverzeichnis am Ende des Buches mit allen wichtigen Informationen zum Zugriff wieder. Zur Veranschaulichung der in diesem Buch vorgestellten Methoden

wird häufig das → *Foodblogkorpus* als Beispiel verwendet. Es umfasst 150 deutschsprachige Texte aus 15 Foodblogs und steht als freier Download zur Verfügung. Alle URLs in diesem Buch wurden zuletzt am 12. Dezember 2023 überprüft.

1.2 Korpus- und Computerlinguistik

Dieses Buch ist eine Einführung in computerlinguistische Methoden. In den Grundlagen des Faches ergibt sich aber eine Überschneidung mit dem Gebiet der Korpuslinguistik. Einige Konzepte und Methoden, die in diesem Buch präsentiert werden, sind auch Teil des korpuslinguistischen Werkzeugkoffers. Deshalb wollen wir die beiden Fächer zu Beginn vergleichend nebeneinanderstellen. Die **Korpuslinguistik** kann definiert werden als:

> die Gesamtheit aller Tätigkeiten, die darauf gerichtet sind,
> (1) umfangreiches authentisches Sprach- oder Textmaterial (gesprochen oder geschrieben) zu sammeln, zusammen zu stellen [sic], aufzubereiten, mit Informationen zu annotieren, zu verwalten und zu warten sowie verfügbar zu machen,
> (2) solches Material für wissenschaftliche oder technische Zwecke oder andere Anwendungen systematisch auszuwerten. (Köhler 2005: 1)

Diese Definition betont in Punkt 1, dass das Textmaterial „umfangreich" sein muss. Dies hängt damit zusammen, dass die Korpuslinguistik im Normalfall zu quantitativen, generalisierenden Aussagen kommen möchte und dazu Muster analysiert, die sich erst ab einer gewissen Menge von Material beobachten lassen. Wie groß die Menge an Daten zu diesem Zweck sein muss, lässt sich nur für den Einzelfall beantworten. Weiterhin wird auch die Authentizität des Sprachmaterials hervorgehoben. Dies erfolgt insbesondere in Abgrenzung zu in der Sprachwissenschaft historisch häufig genutzten Verfahren, der Introspektion, also der Befragung des eigenen, subjektiven Sprachgefühls, sowie der Konstruktion von Beispielsätzen, die unter Umständen zwar grammatisch möglich sind, aber in der wirklichen Sprachverwendung nicht vorkommen.

Die Definition erwähnt außerdem, dass in Korpora gesprochene oder geschriebene Sprache gesammelt werden kann. Beide Modi bringen ihre eigenen Herausforderungen mit sich. Gesprochene Sprache muss zunächst aufgezeichnet und dann transkribiert, also verschriftlicht werden. Auch wenn die automatische Erkennung gesprochener Sprache große Fortschritte macht, erfordert dieser Vorgang in den meisten Fällen erheblichen manuellen Aufwand. Geschriebene Sprache ist oft besser verfügbar, wenn sie von vornherein digital ist oder bereits digitalisiert wurde. Bauen wir hingegen ein Korpus aus mittelalterlichen handschriftlichen Dokumenten auf, ist auch hier mit einem erhöhten Arbeitsaufwand aus Scannen, automatischer Texterkennung, Nachbearbeitungen usw. zu rechnen.

Neben der Erstellung des Korpus wird auch die Verwaltung und Verfügbarmachung als Aufgabe der Korpuslinguistik angeführt. Während nicht bei jedem Korpus eine

Veröffentlichung möglich ist (insbesondere im Rahmen studentischer Arbeiten mit geringen Kapazitäten oder bei Korpora aus datenschutzrechtlich sensiblen oder urheberrechtlich geschützten Texten), ist es für die wissenschaftliche Gemeinschaft von großer Bedeutung, dass Daten allgemein zur Verfügung stehen, sodass Aufbereitungsarbeit nicht mehrfach geleistet werden muss. Öffentlich verfügbare Daten ermöglichen außerdem eine unabhängige Überprüfung von Ergebnissen und tragen so dazu bei, dass das Fach den Ansprüchen an die gute wissenschaftliche Praxis gerecht wird (siehe z. B. Deutsche Forschungsgemeinschaft 2022).

Punkt 2 der Definition trägt der Tatsache Rechnung, dass Korpora nicht nur zu linguistischen Zwecken analysiert werden, sondern für ganz unterschiedliche Wissenschaften interessant sein können. Dies ist gerade im Kontext der Digital Humanities von Bedeutung, wo sich potenziell alle geisteswissenschaftlichen Fächer korpus- und computerlinguistischer Methoden bedienen. Die zusätzliche Erwähnung von technischen Zwecken leitet bereits in den Zuständigkeitsbereich der **Computerlinguistik** über:

> Die Computerlinguistik ist diejenige Wissenschaft, die ganz allgemein die maschinelle Verarbeitung von Sprache mit dem Computer in den Blick nimmt. Im Mittelpunkt stehen dabei Prozesse, die die Erzeugung oder Analyse von gesprochener oder schriftlich fixierter Sprache erlauben. Aber auch die Beschreibung der Sprache selbst in einer Weise, dass der Computer damit umgehen kann, ist Gegenstand der Computerlinguistik. Und schließlich verfolgt man mit der maschinellen Verarbeitung von Sprache meist ein bestimmtes praktisches Ziel, so dass auch die Entwicklung von Software, von sprachverarbeitenden Systemen, ein wichtiges Teilgebiet der Computerlinguistik darstellt. (Lobin 2010: 10)

Wie bei der Korpuslinguistik wird Sprache auch hier in geschriebener und gesprochener Form erwähnt. Während in der Korpuslinguistik gesprochene Sprache für die Analyse zunächst ins geschriebene Medium übertragen wird, befassen sich Teilbereiche der Computerlinguistik auch mit Sprache als akustischem Phänomen, etwa im Kontext von Sprachassistenten. In diesem Buch wird es nur um die schriftliche Form von Sprache gehen.

Gegenüber der Korpuslinguistik wird in der Definition eine neue Unterscheidung gemacht: Die Computerlinguistik befasst sich wie die Korpuslinguistik mit der Analyse, aber auch mit der Erzeugung von Sprache. Während wir uns bei der Analyse mit vorhandenen Sprachdaten befassen, können wir bei der Erzeugung ganz neue Sprache, ebenfalls in gesprochener oder geschriebener Form, generieren. Sprachgenerierung kommt zum Beispiel bei Chatbots zum Einsatz und wird im Kontext von Produktbeschreibungen genutzt. Sie ist außerdem Teil der bereits erwähnten Sprachassistenzsysteme, die in gesprochener Sprache auf unsere Fragen antworten.

Die Definition erwähnt die Herausforderung, Sprache überhaupt in einer Form zu modellieren, mit der ein Computer etwas anfangen kann. Das betrifft zum Beispiel die Segmentierung: Arbeiten wir mit Wörtern, Sätzen oder Texten als Analyseeinheiten? Was für Kategorien benötigen wir für unsere Analyse? Und in was für Datenstrukturen

können wir all das im Computer abbilden? Im Laufe des Buchs werden wir uns mit einigen Möglichkeiten hierzu befassen.

Ein wichtiger Unterschied gegenüber der Korpuslinguistik ist das am Ende erwähnte praktische Ziel: Computerlinguistische Entwicklungen erfolgen mehrheitlich in Hinblick auf ein bestimmtes Anwendungsszenario oder ein spezifisches Problem, das mithilfe von Software gelöst werden soll. Das kann zum Beispiel darin bestehen, für beliebige Sätze eine vollständige syntaktische Analyse zu produzieren oder zuverlässig positive Bewertungen zu einem Produkt von negativen zu unterscheiden.

Betrachtet man nun Korpus- und Computerlinguistik im Vergleich, zeigen sich Gemeinsamkeiten und Unterschiede: Beide Felder betreiben computergestützte Forschung zu Sprache mithilfe von Korpora. Aber sie verfolgen dabei ganz unterschiedliche Erkenntnisinteressen: Die Korpuslinguistik beschreibt die Verwendung von Sprache anhand von Korpora und ist an sprachlichen Mustern um ihrer selbst willen interessiert. Die Computerlinguistik demgegenüber versucht, Sprache mit dem Computer zu modellieren und so praktische Probleme technisch zu lösen (McEnery & Hardie 2012: 228).

Folglich unterscheidet sich auch, was in den beiden Fächern jeweils als interessantes Forschungsergebnis betrachtet wird. Durch die hohe Anwendungsorientierung in der Computerlinguistik ist die entscheidende Frage häufig: Wie gut funktioniert dieses System, das bestimmte sprachliche Muster oder Handlungen erkennen soll? Zum Beispiel: Mit welcher Genauigkeit kann das automatische System Hatespeech in den sozialen Medien erkennen? Ist es genau genug, um in der Praxis eingesetzt werden zu können? Die Korpuslinguistik legt den Fokus hingegen auf die Beschreibung und Erklärung von sprachlichen Phänomenen und fragt: Was können wir aus den Ergebnissen über den sprachlichen Gegenstand lernen? Im Beispiel interessiert sich die Korpuslinguistik etwa für die Frage: Welche sprachlichen Merkmale zeichnen Hatespeech in den sozialen Medien aus und welche Funktionen haben sie?

Viele Ergebnisse computerlinguistischer Forschung werden heute in der Korpuslinguistik und den Digital Humanities eingesetzt und einige davon werden wir in diesem Buch betrachten. Dazu gehört etwa die Tokenisierung, also die Segmentierung von Zeichenketten in Wörter, die Annotation von Wörtern mit ihrer Wortart oder ihrer syntaktischen Funktion sowie die Erkennung von Eigennamen oder im Text ausgedrückten Sentiments. Diese Analyseoptionen sind nicht nur in den Sprachwissenschaften relevant. Sprache ist auch in vielen anderen Geisteswissenschaften Gegenstand oder ermöglicht zumindest einen Zugang zum Forschungsgegenstand: „[E]xperience of the human world is largely a textually mediated experience, and to that extent, human beings live in a textually mediated world“ (McEnery & Hardie 2012: 230).

1.3 Grundbegriffe

In diesem Kapitel war bereits vielfach von Korpora die Rede, die sowohl in der Korpus- als auch in der Computerlinguistik eine entscheidende Rolle spielen. In diesem

Abschnitt werfen wir einen genaueren Blick auf die Grundbegriffe Korpus, Metadaten und Annotationen.

Das Wort „Korpus“ gibt es im Deutschen als Maskulinum und als Neutrum:[2] Während wir „der Korpus“ sagen, wenn es zum Beispiel um den Körper eines Menschen, eines Schranks oder einer Gitarre geht, nutzen wir „das Korpus“ in der Korpuslinguistik, um von einer wissenschaftlich untersuchbaren Textsammlung zu sprechen. Genauer lässt sich das **Korpus** (Plural: Korpora) wie folgt definieren:

> Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. (Lemnitzer & Zinsmeister 2015: 13)

Die meisten Korpora umfassen schriftliche Äußerungen. Der Aufbau von Korpora gesprochener Sprache ist meist aufwendiger, da er zunächst die Transkription der gesprochenen Sprache, also ihre Übertragung in den schriftlichen Modus, erfordert. Generell ist die gesprochene Sprache deshalb korpuslinguistisch weniger erforscht. Aber auch die Aufbereitung schriftlicher Texte kann sehr aufwendig sein, wenn sie nicht schon von sich aus maschinenlesbar sind. Für die korpuslinguistische Analyse von handschriftlichen Aufzeichnungen einer historischen Persönlichkeit etwa ist viel manuelle Aufbereitung notwendig. Die maschinenlesbare Form des Korpus ist Voraussetzung für die effiziente (oder überhaupt realistische) Durchführung aller korpus- und computerlinguistischen Verfahren. Neben den Primärdaten, also den Texten selbst, werden in der Definition noch Metadaten und Annotationen als Teile von Korpora genannt.

Der Begriff der **Metadaten** ist manchen vielleicht aus dem öffentlichen Diskurs um Datenschutz und die Vorratsdatenspeicherung bekannt. Hier ist oft die Rede davon, dass beispielsweise Metadaten von Telefongesprächen erfasst werden können. Es werden also nicht die Gespräche selbst aufgezeichnet (das wären hier die Primärdaten), aber alle Informationen dazu erfasst, wer wann wie lange mit wem telefoniert hat – Informationen, die ebenfalls bereits weiterreichende Schlüsse zulassen. Metadaten sind also ihrem Präfix entsprechend „Daten über die Daten“. In der Korpuslinguistik beantworten Metadaten die Frage: Was ist eigentlich drin in diesem Korpus? Diese Information ist essenziell, um wissenschaftlich mit den Daten arbeiten zu können. Welche Metadaten wichtig sind und zu den Texten des Korpus zur Verfügung stehen sollten, hängt von der Fragestellung ab, zu deren Beantwortung sie beitragen sollen. Beispiele für häufig erfasste Metadaten sind etwa die Textsorte, die Autorin oder der Autor des Textes (bzw. demografische Daten wie Alter und regionale Herkunft), der

2 Genaugenommen gibt es auch das Femininum „die Korpus“. Dabei handelt es sich um einen Fachbegriff aus dem Druckwesen für einen Schriftgrad von 10 Punkt (https://www.duden.de/node/83098/revision/1413078).

Modus (geschrieben/gesprochen), der Entstehungszeitpunkt und ggf. die Erhebungsbedingungen.

Metadaten sind unheimlich wichtig, um die Daten zu verstehen und beurteilen zu können, ob ein verfügbares Korpus zu unserer Fragestellung passt. Anhand der Metadaten können wir außerdem erkennen, ob es im Korpus möglicherweise Teilgruppen gibt, in denen die Antwort auf unsere Frage unterschiedlich ausfällt und die getrennt analysiert werden sollten. Wenn wir uns etwa für die Satzlänge in der deutschen Schriftsprache interessieren und unser Korpus Zeitungstexte, wissenschaftliche Texte und Social-Media-Posts enthält, ist eine separate Analyse dieser Gruppen empfehlenswert.

Wenn wir ein bereits verfügbares Korpus nutzen, müssen wir uns deshalb immer ausführlich über die Metadaten informieren und prüfen, ob die Daten zur Bearbeitung unserer Fragestellung geeignet sind. Metadaten werden zum Beispiel über eine begleitende Webseite veröffentlicht oder können in wissenschaftlichen Publikationen enthalten sein. Sollten relevante Informationen fehlen, besteht vielleicht die Möglichkeit, direkt bei den Ersteller:innen nachzufragen. Wenn wir selbst ein Korpus erstellen, müssen wir neben den Texten selbst auch möglichst viele Metadaten erheben. Es empfiehlt sich, dies frühzeitig anzugehen, bevor eventuell Informationen verloren gehen. Im Zweifelsfall lohnt es sich, alle verfügbaren Metadaten zu erfassen, falls sie sich erst später als wichtig herausstellen oder Forscher:innen mit anderen Interessen das Korpus nachnutzen wollen. Metadaten sollten in maschinenlesbarer Form erfasst werden, zum Beispiel in einer Tabelle. Bei der Erfassung sollte man von vornherein auf Einheitlichkeit achten, um spätere Nachbearbeitungen zu vermeiden (z. B. gibt es sehr viele unterschiedliche Möglichkeiten, ein Datum zu schreiben). Suchen wir zum Beispiel im Kernkorpus des 20. Jahrhunderts des → *Digitalen Wörterbuchs der deutschen Sprache (DWDS)* nach Verwendungsbelegen für ein bestimmtes Wort, bekommen wir zu jedem Treffer die Information, aus welcher Publikation er stammt (inkl. Titel, Autor:in, Veröffentlichungsjahr, Seitenzahl), zu welcher Textklasse der Text gehört (Belletristik, Wissenschaft, Gebrauchsliteratur oder Zeitung) und welcher Lizenz der Text unterliegt.

Neben den Primärdaten und den Metadaten ist in der oben angeführten Definition von Korpus noch von (optionalen) Annotationen die Rede. Bei **Annotationen** handelt es sich um Anreicherung des reinen Textes eines Korpus mit zusätzlichen Informationen. Oft sind das linguistische Informationen wie Wortarten, syntaktische Strukturen, Eigennamen oder Koreferenzrelationen. Grundsätzlich kann aber jede Art Information annotiert werden, die am Text beobachtbar ist, zum Beispiel das Thema eines Absatzes oder die Erzählebene in literarischen Texten. Metadaten liefern in der Regel Informationen über den Text als Ganzes, Annotationen können sich auf sprachliche Einheiten beliebiger Größe beziehen: Laute, Morpheme, Wörter, Wortgruppen, Sätze oder Absätze.

Annotationen ermöglichen es, das Korpus gezielter nach Phänomenen zu durchsuchen. Wenn wir uns zum Beispiel für die Verwendung von Adjektiven interessieren, ist

es hilfreich, wenn zu jedem Wort eines Textes die Wortart hinterlegt ist und wir direkt danach suchen können. Annotationen haben außerdem den Vorteil, dass sie unsere Interpretation der Daten wiederauffindbar und kritisierbar machen. Wenn wir in einem Text annotiert haben, in welchen Sätzen es unserer Meinung nach um Krankheit geht, kann eine andere Person sich diese Annotationen später ansehen und unter Umständen feststellen, dass sie selbst manche Entscheidungen anders getroffen hätte. So tragen Annotationen zur Wissenschaftlichkeit des Forschungsprozesses bei.

Annotationen können manuell oder automatisch vorgenommen werden. Mit beiden Formen werden wir uns in diesem Buch ausführlich beschäftigen. Die manuelle Annotation (Kapitel 10) erfordert eine klare Ausformulierung von Regeln zur Annotation (Annotationsrichtlinien), damit die Annotationen nicht subjektiv ausfallen, sondern mehrere Personen anhand der Regeln zu ähnlichen Annotationsergebnissen kommen. Die manuelle Annotation kann durch zahlreiche digitale Tools unterstützt werden. Das Ziel der Computerlinguistik ist in der Regel die automatische Annotation durch den Computer. Für einige linguistische Kategorien ist das bereits mit hoher Qualität möglich, etwa für die Wortarten. Andere Kategorien, die mehr Wissen über den größeren sprachlichen Kontext oder die Welt erfordern, sind weniger gut automatisierbar. Für viele Kategorien stehen bereits Tools zur automatischen Annotation zur Verfügung. Wenn wir für unsere Analyse individuellere Kategorien benötigen, für die das nicht der Fall ist, können wir uns auch selbst an der Automatisierung der Annotationsaufgabe versuchen (Kapitel 11 und 12).

Die Kategorien oder Label, die bei der Annotation vergeben werden, bezeichnet man auch als **Tags**. Eine Sammlung von Tags, die gemeinsam einen Phänomenbereich abdecken, heißt **Tagset**. Das STTS (Schiller et al. 1999) ist zum Beispiel ein Tagset zur Annotation von Wortarten, das aus 54 unterschiedlichen Tags besteht. Von einem Tagset erwarten wir, dass es das zu annotierende Phänomen mehr oder weniger vollständig abdeckt. Ein Tagset ist exhaustiv, wenn für alle denkbaren Phänomene ein Tag vorgesehen ist. Tagsets sollten außerdem disjunkt sein, d. h., die Kategorien sind trennscharf und jeder Instanz wird genau ein Tag zugewiesen. Im Beispiel der Wortarten sollte jedes Wort des Deutschen genau einer Kategorie des Tagsets zuzuweisen sein (und nicht keiner oder mehreren).

Annotationen können in ganz unterschiedlichen Formen vorgenommen werden. Allen von uns bekannt sind sicherlich handschriftliche Annotationen auf Papier. So können wir etwa beim Lesen einer Kurzgeschichte anhand von Markierungen formale oder inhaltliche Besonderheiten hervorheben, zum Beispiel alle Textstellen, die zur Charakterisierung der Hauptfigur beitragen. Annotationen auf Papier sind leicht anzufertigen und oft für einen ersten Zugang zu einem Text geeignet, wenn die zu annotierenden Kategorien möglicherweise noch gar nicht feststehen. Die Auswertung von Annotationen auf Papier ist allerdings mühsam und kaum automatisierbar, weshalb es sich immer empfiehlt, frühzeitig auf eine digitale Form umzusteigen.

Mit welchen Annotationstools und in welchem Annotationsformat wir sinnvollerweise arbeiten, hängt von einer Reihe von Faktoren ab, insbesondere davon, welche

Art Information wir annotieren und wie wir sie später analysieren wollen. In den folgenden Kapiteln werden wir eine Reihe von Beispielen für Annotationskategorien und die Arbeit mit ihnen kennenlernen.

Teil 1: Linguistische Ausgangspunkte