Vilnius University Press

Vilnius University Press

Kalbotyra
2025 (78) 201–208

ISSN 1392-1517 eISSN 2029-8315
DOI: https://doi.org/10.15388/Kalbotyra.2025.78.9

Book reviews

Melanie Andresen. *Computerlinguistische Methoden für die Digital Humanities: Eine Einführung für Geisteswissenschaftler:innen*. Tübingen: Narr Francke Attempto (= narr STUDIENBÜCHER), 2024, 240 S., ISBN 978-3-8233-8579-0, EAN 978382338579.

Rezensiert von / Reviewed by **Christa Dürscheid**

Deutsches Seminar
Universität Zürich
Schönberggasse 9
8001 Zürich, Schweiz
E-Mail: duerscheid@ds.uzh.ch
ORCID iD: 0000-0001-9141-7562
https://ror.org/02crff812

Vorbemerkungen

„Die Digital Humanities ersetzen – selbstverständlich – die Geisteswissenschaften nicht, sondern erweitern ihr Methodenarsenal, und zwar dann, wenn Forschungsfragen durch empirische Arbeit mit großen Datensammlungen bearbeitet werden können“ (Krämer 2025, 11). Diese Aussage steht zu Beginn eines Essays, das den markanten Titel „Der Stachel des Digitalen: Geisteswissenschaften und Digital Humanities“ trägt. Darin setzt sich die Philosophieprofessorin Sybille Krämer kritisch mit dem Selbstbild der Geisteswissenschaften auseinander und zeigt auf, in welchem Verhältnis diese zu den Methoden der Digital Humanities stehen. Doch um welche Methoden handelt es sich dabei? Und welche Vorteile können sich daraus ergeben, wenn traditionelle Verfahren der Textanalyse, d. h. Textinterpretation und Hermeneutik, eine Erweiterung ihres „Methodenarsenals“ erfahren? In dem Studienbuch von Melanie Andresen, das hier zur Rezension

Submitted: 10/11/2025. **Accepted:** 17/11/2025

steht, werden diese Fragen nicht nur auf theoretischer Ebene erörtert; es wird eine praxisorientierte Einführung präsentiert, die im Unterschied zu ähnlich gelagerten Überblicksdarstellungen (z. B. Alpaydin 2022; Mitkov 2022) den Schwerpunkt explizit auf die Anwendung computerlinguistischer Methoden in den Geisteswissenschaften legt.

Allerdings ist der Untertitel des Buches – „Eine Einführung für Geisteswissenschaftler:innen" – etwas zu weit gefasst. Denn diese Formulierung lässt vermuten, dass sich die Verfasserin auf Anwendungsfelder aus verschiedenen geisteswissenschaftlichen Disziplinen bezieht (so z. B. auch auf die Geschichtswissenschaft, die Kunstgeschichte, die Theologie oder die Musikwissenschaft), tatsächlich aber hat sie vor allem die Sprach- und Literaturwissenschaft im Blick. Dass die Verfasserin einen solchen Schwerpunkt setzt, verwundert nicht. Sie hat an der Universität Hamburg ein Masterstudium in germanistischer Linguistik absolviert, und auch die Dissertation, die 2022 unter dem Titel *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen* erschien, hat einen germanistisch-linguistischen Hintergrund. Nach der Promotion war sie als Post-Doc am Institut für Maschinelle Sprachverarbeitung an der Universität Stuttgart, seit 2024 arbeitet sie bei der Firma *DeepL* an der Verbesserung maschineller Übersetzung (so die Autor:inneninformation auf der Webseite des Narr-Verlags). Aus dem Vorwort zum Buch entnimmt man zudem, dass Melanie Andresen während ihrer Tätigkeit an der Universität Stuttgart Lehrveranstaltungen unter dem Titel „Computerlinguistische Methoden für die Digital Humanities" durchgeführt hat. Diese dienten ihr als Grundlage für das vorliegende Studienbuch.

Der Werdegang von Melanie Andresen wurde an dieser Stelle auch deshalb kurz skizziert, um deutlich zu machen, dass die Verfasserin vertiefte Kenntnisse in der germanistischen Linguistik und in der Computerlinguistik hat und über Unterrichtserfahrung in den Digital Humanities verfügt. All das kommt ihrem Einführungswerk, wie noch zu zeigen sein wird, zugute. Doch im Titel, im Klappentext und in der Einleitung hätte deutlich werden müssen, dass es im Buch vor allem um die Anwendung computerlinguistischer Methoden in den Philologien (und hier insbesondere in der Germanistik) geht. Zwar mag man einwenden, dass die von Frau Andresen vorgestellten Methoden auch auf andere geisteswissenschaftliche Disziplinen übertragbar sind. Das trifft aber nur dann zu, wenn in diesen Disziplinen Forschungsfragen gestellt werden, die sich auf der Basis von Textanalysen beantworten lassen. Darüber hinaus wäre es interessant gewesen, im Buch etwas über solche Methoden zu erfahren, die zur Bildanalyse oder zur Analyse von Audiodateien (z. B. in der Musikwissenschaft) eingesetzt werden können. Darauf geht die Verfasserin nicht ein. Auch mit der Analyse gesprochener Sprache (z. B. in der Interaktion mit Sprachassistenten) befasst sie sich nicht. Eine Begründung dafür wird nicht gegeben, auf S. 15 heißt es dazu nur: „In diesem Buch wird es nur um die schriftliche Form von Sprache gehen."

Aufbau des Buches

Das Kernstück des Buches bilden zwei Teile, die insgesamt mehr als 170 Seiten umfassen (Teil 1: „Linguistische Ausgangspunkte“; Teil 2: „Methoden“). Alle Kapitel im ersten Teil enden mit „Beispielstudien“ (z. B. zu Wahlprogrammen, Interviews und literarischen Werken) und „Übungen“, wobei zu den Übungen auf der Webseite des Narr-Verlags auch Musterlösungen angeboten werden (siehe dazu aber weiter unten). Ergänzt werden diese beiden Themenblöcke um einen dritten, wesentlich kürzeren Teil (elf Seiten), der unter der etwas vagen Überschrift „Gesellschaft“ steht und nur ein Kapitel beinhaltet. Den drei Teilen vorangestellt sind ein Vorwort und eine Einleitung, die – obwohl insgesamt recht kurz (neun Seiten) – bereits in wichtige Grundbegriffe einführt. Ein Verzeichnis der Skripts und der vielen interessanten digitalen Ressourcen, auf die im Laufe des Textes verwiesen wird (Korpora, Tools zur Korpusanalyse, Programmierumgebungen für Python etc.), folgt am Ende des Buches (alle Quellen wurden zuletzt im Dezember 2023 geprüft). Auch finden sich hier ein umfangreiches, sorgfältig aufbereitetes Literaturverzeichnis und ein Sachregister, das vom Stichwort „Abkürzungen“ (warum im Plural?) bis zu „Zufallsstichprobe“ reicht. So informativ dieser Service-Teil (der insgesamt über 30 Seiten umfasst) auch ist, es zeigt sich spätestens jetzt, welche Nachteile damit einhergehen, wenn man für die Lektüre des Studienbuches nur die Printausgabe vorliegen hat. Will man die entsprechenden Webseiten im Internet aufrufen, müssen die Links manuell eingetippt werden. In der E-Book-Ausgabe des Buches stellt sich dieses Problem nicht. Als Open Access-Publikation steht das E-Book aber leider nicht zur Verfügung. Im Interesse einer kostenfreien Bereitstellung von Forschungsergebnissen hätte man sich das anders gewünscht.

Zu den einzelnen Kapiteln

In der Einleitung trägt die Verfasserin viele Informationen zusammen, die hilfreich für die Lektüre der Folgekapitel sind (Kap. 1). Dazu gehört nicht nur, wie in Einleitungen zu wissenschaftlichen Monographien üblich, dass sie Inhalt und Aufbau des Buches knapp skizziert, sie behandelt schon hier interessante thematische Aspekte. So geht sie auf die (nicht triviale) Frage ein, welche Gemeinsamkeiten und Unterschiede zwischen Computerlinguistik und Korpuslinguistik bestehen und was man unter Termini wie *Metadaten*, *Annotationen* und *Tagsets* versteht. Auch stellt die Verfasserin in Kap. 1 – allerdings nur sehr kurz – das Korpus vor, auf das sie sich im Folgenden immer wieder bezieht. Es ist das Foodblog-Korpus, das 150 Texte aus 15 verschiedenen Blogs umfasst (vgl. Andresen & Zinsmeister 2018). Die Einleitung ist insgesamt sehr informativ, sie ist – wie der gesamte Text – flüssig geschrieben und sie beginnt wie alle folgenden Hauptkapitel auch: Zunächst legt Melanie Andresen in einigen wenigen Zeilen dar, welche Themenschwerpunkte sie im Folgenden setzen wird, dann erst beginnt sie mit ihren inhaltlichen

Ausführungen. Das ist eine lesefreundliche Vorgehensweise, zudem ist dieser einleitende Passus jeweils recht eingängig formuliert – und dies konsequent im *Wir*-Duktus. Das liest sich dann so: „Wir klären außerdem, was genau die Computerlinguistik ist" (S. 11). Oder an späterer Stelle: „Wir schauen uns einerseits lexikonbasierte Ansätze an, […], andererseits betrachten wir Ansätze mit maschinellem Lernen" (S. 81).

Teil 1: Linguistische Ausgangspunkte

Wie die Überschrift zu Teil 1 bereits vermuten lässt, steht in diesem ersten Teil des Buches die Linguistik (und nicht die Computerlinguistik) im Zentrum. Als „Ausgangspunkte" dienen verschiedene linguistische Beschreibungsebenen, die von der Lexik über die Syntax bis zur Semantik und Pragmatik reichen (vgl. Kap. 2 bis 8). Die ersten Kapitel lesen sich fast wie eine Bachelor-Einführung in die Sprachwissenschaft – ergänzt um Informationen zu Grundbegriffen der Korpusanalyse (z. B. *Tokenisierung*, *Lemmatisierung*, *Type-Toke-Ratio*). Inhaltlich sind die Ausführungen in den Kap. 2 bis 4 sehr gut nachvollziehbar. Stellenweise sind sie sogar so elementar, dass sich zunächst nicht erahnen lässt, wie anspruchsvoll der Text in der Folge noch werden wird. So beginnt das Kapitel zur Lexik mit einem Überblick über verschiedene Definitionen zum Wort-Konzept (Kap. 2.1), das Wortarten-Kapitel listet einleitend eine Reihe von Kriterien auf, nach denen sich die Wortarten im Deutschen unterscheiden lassen (Kap. 3.1). Im Syntax-Kapitel informiert die Verfasserin über die Grundlagen der Konstituentengrammatik und Dependenzgrammatik, dann leitet sie über zu der Frage, welche Möglichkeiten sich auf Basis dieser beiden syntaktischen Theorien für die automatisierte Textanalyse bieten. Sie stellt verschiedene Annotationsschemata vor (z. B. das Schema der *Hamburg Dependency Treebank*, HDT), beschreibt die Speicherung und Weiterverarbeitung von Dependenzannotationen (z. B. im CoNLL-Format) und gibt einen Überblick zur computerbasierten Syntaxanalyse (Kap. 4.3). In diesem Zusammenhang geht sie auch auf das Dependenzparsing und die Graphentheorie ein und nennt Tools, mit denen syntaktische Strukturen analysiert werden können (z. B. CoreNLP). Zum Schluss präsentiert Melanie Andresen – wie in allen Kapiteln von Teil 1 – „Beispielstudien" (hier z. B. zur Analyse von literarischen Texten) und „Übungen". Letztere starten mit einfachen Fragen zur Satzgliedbestimmung, dann folgen Aufgaben zur automatischen Annotation und zur Arbeit mit Musterskripten (Kap. 4.5).

Die folgenden drei Kapitel zur Semantik tragen die Überschriften „Wortfelder" (Kap. 5), „Sentimentanalyse" (Kap. 6) und „Distributionelle Semantik" (Kap. 7). Zunächst erinnert die Lektüre wieder an eine Bachelor-Einführung in die Sprachwissenschaft. Zu Beginn von Kap. 5 kündigt die Verfasserin an, dass man sich nun „mit der linguistischen Teildisziplin der Semantik [befassen werde], die die Bedeutung von Wörtern und ihre semantischen Relationen behandelt" (S. 71). Im Anschluss daran geht sie ausführlich auf

Ähnlichkeitsbeziehungen zwischen Wörtern und auf die Wortfeldtheorie ein und weist auf nützliche Tools zur automatisierten Wortschatzanalyse hin (z. B. *GermaNet*). Anspruchsvoller wird die Lektüre in den Kapiteln 6 und 7, in denen Melanie Andresen das Erstellen von Sentimentwörterbüchern und die Möglichkeiten zur computergestützten Analyse von Emotionen thematisiert. Auch stellt sie hier die Grundlagen der distributionellen Semantik vor und erklärt, wie die semantischen Relationen zwischen Wörtern auf der Basis von Vektorwerten und Kosinusähnlichkeiten berechnet werden können. Damit schafft sie die Grundlagen, die nötig sind, um in den Kapiteln 7.3.1 bis 7.3.5 ihren Ausführungen zum *Word Embedding* folgen zu können. Veranschaulicht wird dieses Verfahren wieder mit Beispielen aus dem Foodblog-Korpus. Es wird gezeigt, wie die Wörter *kochen* und *Minuten* als Vektorwerte dargestellt werden können und wie sich automatisiert ermitteln lässt, welche Wörter im Kontext anderer Wörter vorkommen. Zum Schluss des Kapitels stellt die Verfasserin einige Arbeiten vor, in denen in den Geisteswissenschaften mit *Word Embeddings* gearbeitet wurde, um „Wordfelder" [sic] zu erstellen, um Synonyme zu einem Zielwort zu ermitteln oder die Veränderung von Wortbedeutungen zu untersuchen (vgl. S. 106 f.): Sie verweist auf eine Untersuchung zur automatischen Erkennung von Shakespeare-Referenzen in modernen Texten, auf eine diachrone Studie zum Wortschatz in der englischen Wissenschaftssprache und auf eine Arbeit zur Verwendung des Wortes *national* in verschiedenen Korpora. Die „Übungen" in Kap. 7.5 umfassen vier Aufgaben, von denen sich zwei abermals auf das Foodblog-Korpus beziehen (hier auf die Berechnung der Kosinusähnlichkeit zwischen Wortpaaren wie *backen* und *Butter*, *Mehl* und *Salz*).

Das Kapitel „Pragmatik" (so die Überschrift) steht am Ende des ersten Themenblocks. Die Verfasserin erläutert zunächst die Termini *Entität* und *Referenz* und geht dann auf das computergestützte Verfahren der Eigennamenerkennung (*Named Entity Recognition*) und das Vorgehen bei der Annotation von Koreferenzrelationen ein. Wie praktisch diese Methoden sind, um Texte inhaltlich zu erschließen, macht sie wieder in den „Beispielstudien" (Kap. 8.4) deutlich. Dieser Abschnitt fällt mit drei Seiten sogar etwas länger aus als die Beispielstudien in den vorangehenden Kapiteln. Die Verfasserin stellt hier eine Untersuchung zur Figurencharakterisierung in einem Roman von Juli Zeh vor, die sie selbst zusammen mit Michael Vauth durchgeführt hat, sie geht aber auch auf andere interessante Studien ein (z. B. Braun & Ketschik 2019).

Teil 2: Methoden

Der zweite Teil des Studienbuchs fällt im Vergleich zu Teil 1 wesentlich kürzer aus, er umfasst nur vier Kapitel und ca. 70 Seiten. Wer sich gezielt über aktuelle korpus- und computerlinguistische Methoden informieren möchte, sei direkt auf diesen Teil verwiesen. In Kap. 9 erklärt Melanie Andresen auf didaktisch geschickte Weise, was reguläre

Ausdrücke sind; auch die Informationen zur Berechnung von Frequenzen, zur deskriptiven Statistik und zur Interenzstatistik sind gut nachvollziehbar. Sehr nützlich ist weiter, dass sie verschiedene Arten von Visualisierungen vorstellt (Säulendiagramme, Tortendiagramme, Boxplots u. a.) und deutlich macht, wie wichtig solche Darstellungen für die Textanalyse und die Ergebnispräsentation sind. Auch hier stehen Übungen am Ende des Kapitels (etwa zur Bestimmung des Skalenniveaus und zur Analyse von Beispielskripten), doch nur zu drei der vier Aufgaben finden sich im Internet auch Musterlösungen. Auf eine Zusammenstellung von „Beispielstudien" wird verzichtet.

In Kap. 10 erläutert die Verfasserin die Unterschiede zwischen manueller und automatischer Annotation, die Funktion von Annotationsrichtlinien und die verschiedenen Verfahren zur Qualitätskontrolle (z. B. das Inter-Annotator-Agreement). Thematisch schließt das Kapitel an Teil 1 der Einführung an, zum Verständnis ist es aber nicht nötig, die vorangehenden Ausführungen im Detail gelesen zu haben. Zwar fehlt auch hier ein separates Kapitel mit Beispielstudien, im Verlauf des Kapitels werden gelegentlich aber Bezüge zu spezifischen Forschungsfragen in den Geisteswissenschaften hergestellt (so etwa zu der Frage, wie der Mediendiskurs über ein bestimmtes Ereignis ausgewertet werden kann).

Die folgenden Kapitel befassen sich mit dem Maschinellen Lernen (Kap. 11) und seiner Weiterentwicklung, dem Deep Learning (Kap. 12), also mit solchen Methoden, die im Gegensatz zu regelbasierten Systemen mit KI-basierten Techniken arbeiten und auch in den Digital Humanities immer wichtiger werden. Insofern ist es besonders lobenswert, dass die Verfasserin dem maschinellen Lernverfahren und der Frage, wie künstliche neuronale Netze funktionieren, gleich zwei Kapitel widmet. Dabei ist sie durchweg bemüht, eine verständliche Einführung in diese komplexe Thematik zu geben. Das gelingt ihr über weite Strecken, doch wird diese Aufgabe immer anspruchsvoller – und das nicht zuletzt deshalb, weil ein solides mathematisches Grundlagenwissen erforderlich ist, um den Ausführungen folgen zu können. Die Inhalte der beiden Kapitel können hier nicht im Detail vorgestellt werden, es seien nur einige Aspekte genannt, die zur Sprache kommen und von der Verfasserin Schritt für Schritt erläutert werden. Dazu gehören (ausgewählten Überschriften in Kap. 11 und 12 folgend): die Unterscheidung zwischen überwachtem und unüberwachtem Lernen (Kap. 11.2), der Musterablauf einer Klassifikation (Kap. 11.3), das Training eines Deep-Learning-Modells (Kap. 12.3), Recurrent Neural Networks (Kap. 12.5) und Transformer (Kap. 12.6). Welche Forschungsfragen damit bearbeitet werden können, rückt hier allerdings in den Hintergrund, der Fokus liegt darauf, die Methoden so weit verständlich zu machen, dass man selbst künstliche neuronale Netze programmieren und trainieren oder auf einem Server damit arbeiten kann. Doch immerhin wird in den „Übungen" in Kap. 12.8 danach gefragt, welche Anwendungsfälle es für Deep Learning in den Geisteswissenschaften geben könnte. Auf der Webseite des

Narr-Verlags findet man dazu aber keine Hinweise. Das gilt auch für die anderen beiden Übungen, die hier angegeben sind (z. B. zum Modell *German BERT*). Auf der Webseite steht dazu nur, dass keine Musterlösungen vorhanden seien.

Teil 3: Gesellschaft

Dieser Teil besteht nur aus einem Kapitel; es trägt die Überschrift „Computerlinguistik und Ethik" (Kap. 13). Die Verfasserin nimmt hier vor allem solche Fragen in den Blick, die ethische Probleme beim Einsatz von KI-basierten Tools betreffen (etwa die Reproduktion von Diskriminierungen). Sie geht aber – anders als es die Überschrift erwarten lässt – auch auf ökologische Aspekte ein (so z. B. auf den enormen Energieverbrauch, der aus dem Training und der Nutzung von Sprachmodellen resultiert). In Kap. 13.5 problematisiert sie abschließend den Umstand, dass gesellschaftliche Entscheidungsprozesse davon abhängen können, welche Daten überhaupt für die Textanalyse zur Verfügung stehen. Wenn in historischen Textkorpora bestimmte Personengruppen über- oder unterrepräsentiert sind (z. B. Frauen), könne dies, so legt sie dar, dazu führen, dass in den Digital Humanities bestimmte Texte weniger Berücksichtigung finden (vgl. S. 208). So seien im literarischen Kanon Männer stark überrepräsentiert, Texte von Frauen würden kaum Beachtung finden. Mit solch kritischen Überlegungen leitet die Verfasserin zum Ende ihres Studienbuchs über und plädiert im letzten Satz „für eine differenzierte Sicht auf Daten und Analysen" (S. 208).

Fazit

Melanie Andresen ist es mit ihrem Studienbuch hervorragend gelungen, in ein „Methodenarsenal" (siehe das obige Zitat) einzuführen, das in den Geisteswissenschaften immer mehr an Bedeutung gewinnt. Sie führt ihre Leserinnen und Leser souverän durch alle 13 Kapitel und achtet dabei immer darauf, dass der Text, der zunehmend an Komplexität gewinnt, verständlich bleibt. Auf formaler Ebene gibt es nichts zu beanstanden; das gesamte Buch wurde sorgfältig gestaltet, der Text ist sehr gut ausformuliert, mit vielen anschaulichen Abbildungen und interessanten Zusatzmaterialien. Die wenigen Kritikpunkte, die zu nennen sind, beziehen sich vor allem auf das, was fehlt: Nicht zu allen Übungsaufgaben gibt es Musterlösungen; auch hätte man sich ein Abkürzungsverzeichnis gewünscht, um die vielen Kürzel (TTR, STTR, STTS, POS, HMM, HDT, CoNLL, NER etc.) nachschlagen zu können. Auf inhaltlicher Ebene wäre es bereichernd gewesen, wenn mehr Bezüge zu geisteswissenschaftlichen Forschungsfragen hergestellt worden wären und auch auf die Analyse nicht-geschriebener Daten eingegangen worden wäre.

Diese Anmerkungen sollen aber nicht die enorme Leistung schmälern, die hinter dem Studienbuch steht, sondern lediglich als Hinweise darauf verstanden werden, was in ei-

ner zweiten Auflage noch ergänzt werden könnte. Und eine zweite Auflage ist dem Buch zu wünschen. Bis hinein in die vielen weiterführenden Fußnoten stellt Melanie Andresen ihr profundes Fachwissen unter Beweis und arbeitet auch schwierige Inhalte didaktisch geschickt auf. Im Resultat liegt ein Werk vor, das mehr ist als ein Studienbuch; es ist ein Grundlagentext für die Digital Humanities. Die Lektüre wird allen empfohlen, die sich über aktuelle computerlinguistische Methoden informieren möchten.

Literatur

Alpyadin, Ethem. 2022. *Maschinelles Lernen*. 3., aktualisierte und erweiterte Auflage. Berlin: De Gruyter.

Andresen, Melanie. 2022. *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen*. Tübingen: Narr Francke Attempto.

Andresen, Melanie & Michael Vauth. 2020. Figurenrelationen und Figurencharakterisierung. Interdisziplinarität zwischen Literaturwissenschaft und Computerlinguistik am Beispiel der Text- und Genreanalyse. *Kultur und Technik. Interdisziplinäre Perspektiven*. Dominik Orth & Margarete Jarchow, Hrsg. Kiel/Hamburg: Wachholtz. 43–62.

Andresen, Melanie & Heike Zinsmeister. 2018. *Foodblog-Korpus. Zenodo*. https://doi.org/10.5281/zenodo.1410445

Braun, Manuel & Nora Ketschik. 2019. Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen. *Das Mittelalter* 24 (1), 54–70.

Krämer, Sybille. 2025. *Der Stachel des Digitalen. Geisteswissenschaften und Digital Humanities*. Frankfurt: Suhrkamp.

Mitkov, Ruslan, Hrsg. 2022. *The Oxford Handbook of Computational Linguistics*. 2. Auflage. Oxford: Oxford University Press.