jfml Journal für Medienlinguistik
Journal for Media Linguistics

Vol 7 (2025), No 1: 38–42
DOI: 10.21248/jfml.2025.92

Rezension

Andresen, Melanie (2022): Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen. Tübingen: Narr Francke Attempto (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 10). € 88,00 ISBN 978-3-8233-8514-1

Marcella Palladino

Die Arbeit von Andresen befasst sich in 11 Kapiteln (einschließlich des Anhangs) mit syntaktischen Annotationen in germanistischen Wissenschaftssprachen und untersucht Dissertationen aus der Literaturwissenschaft und der Linguistik. Kernpunkt des Werkes ist das methodische Verfahren bei der Erstellung und Analyse syntaktisch annotierter Korpora. Die Autorin geht den Fragen nach, welches Potenzial datengeleitete Forschung bietet und welche Möglichkeiten der Einsatz automatischer syntaktischer Annotationen bei der Sprachbeschreibung eröffnet.

Die erste methodische Grundlegung, die die Autorin in der Einleitung (**Kap. 1**) vornimmt, ist, dass diese Arbeit im methodischen Bereich der datengeleiteten Forschung zu verorten ist, wobei linguistische Annotationen laut der Autorin in der datengeleiteten Korpuslinguistik nur selten eingesetzt wurden. Datengeleitete Methodik bedarf eines statistischen Vorgehens, das die Daten auf Auffälligkeiten untersucht, anstatt eine Anfangshypothese zu formulieren. Darin sieht die Autorin das Potenzial syntaktischer Annotationen für eine datengeleitete Methodik und ebenso automatischer syntaktischer Annotationen. Diese sollten aber ihrer Ansicht nach aufgrund der unzureichenden Qualität üblicher automatisierter Annotationssystemen stets mit Reflexion begleitet werden.

Nach der Einleitung setzt sich die Autorin mit den beiden für das Korpus ausgewählten Disziplinen – Literaturwissenschaft und Linguistik – auseinander, deren Fachgeschichte sowie Charakteristika

Kontaktperson:
Marcella Palladino
Università degli Studi di Modena e Reggio Emilia
Largo Sant'Eufemia 19
41121 Modena MO,
ITALY
marcella.palladino@unimore.it

ISSN: 2569-6491

erläutert werden (**Kap. 2**). **Kap. 3** dient als theoretischer Ausgangspunkt. Die Autorin benennt zuerst das Problem, den Begriff *Wissenschaftssprache* eindeutig zu definieren, sowie die Schwierigkeit, eine Wissenschaftssprache von einem Register zu unterscheiden (vgl. S. 37). Nach einer theoretischen Auseinandersetzung mit dem Terminus entscheidet sich Andresen dafür, die Wissenschaftssprache als sprachliches Register nach der Definition von Biber (2006) zu übernehmen. Sie beschreibt sowohl die außersprachlichen als auch die sprachlichen Merkmale der Wissenschaftssprache und illustriert die Unterschiede zwischen Linguistik und Literaturwissenschaft (vgl. S. 58–63).

Kap. 4 ist der Methodologie gewidmet. Die Autorin präzisiert dabei sowohl die Begriffe, die in der Methodik und der Analyse verwendet werden, als auch deren Abgrenzung zu anderen Konzepten, die in der Sprachwissenschaft oft damit verknüpft sind, d. h. *induktiv* vs. *deduktiv* (S. 69 f.), *datengeleitet* vs. *theoriegeleitet* (S. 70–72) und *korpusgeleitet* vs. *korpusbasiert* (S. 72–79). Die Autorin hat sich für eine datengeleitete Methodik entschieden (vgl. S. 80): Lexikalische Daten werden in der Studie mit grammatischen Annotationen angereichert und Hypothesen werden aus den Daten statt aus der Theorie abgeleitet. Durch Annotationen wird die Analyse mit Informationen angereichert, die es ermöglichen, andere Phänomene in den Daten zu berücksichtigen, die bei einer rein wortbasierten Perspektive ausgeschlossen wären. Allerdings versteht die Autorin dies nicht als Kontamination der reinen Daten, sondern betrachtet die Annotationen als eine zusätzliche Perspektive, die an bestehende Forschungsarbeiten zu Wortarten und Dependenzen anschließt. Die Annotation erfolgt erst mit einem deduktiven Umgehen, während danach ein induktives Verfahren angewandt wird. Im induktiven Schritt stehen n-Gramme im Vordergrund und ihre Auswertung erfolgt durch Automatisierung (mittels des maschinellen Lernverfahrens). Die Interpretation der Unterschiede zwischen der Linguistik und der Literaturwissenschaft kombiniert die induktive (manuelle) Analyse distinktiver n-Gramme in ihren Verwendungskontexten im Korpus mit der Einordnung der Ergebnisse anhand Theorien zur Beschreibung der genannten Disziplinen und den Forschungsstand zur Wissenschaftssprache.

Andresen greift datengeleitete Forschungsmethoden zur Sprache im **Kap. 5** auf. Die Sprachmodellierung anhand von n-Grammen wird erläutert und der Unterschied zwischen linearen n-Grammen und syntaktischen n-Grammen wird ausführlich erklärt. Syntaktische n-Gramme erweitern lineare n-Gramme, da sie die syntaktischen Relationen zwischen Wörtern im Satz berücksichtigen. Dadurch kann eine weniger arbiträre linguistische Interpretation dargestellt werden, die über die lineare Oberflächenstruktur des Textes bzw. des Satzes hinausgeht (vgl. S. 85 f.).

Kap. 6 umfasst die Kriterien der Korpuserstellung und der Datenannotation. Gegenstand der Analyse sind 60 Dissertationen – 30 aus der Linguistik und 30 aus der Literaturwissenschaft – von verschiedenen Universitäten. Die Dissertationen sind im PDF-Format verfügbar und sind keine vom Verlag publizierten Texte. Dies bedeutet, dass keine verlagsspezifischen Formatierung- und Schreibrichtlinien vorhanden sind, sondern die disziplintypischen Merkmale übernommen werden (vgl. S. 118). Die Kriterien, die der Textauswahl zugrunde liegen, sind: die eindeutige disziplinäre Zugehörigkeit, ein geringer Fremdsprachanteil (obwohl keine Informationen zur L1 der Autor*innen vorhanden waren), wenige typographische Besonderheiten sowie eine Streuung über verschiedene Universitäten (vgl. S. 120). Die letzten zwei Kriterien lassen offen, was genau *wenig* bedeutet und warum die Verteilung der Texte auf die ausgewählten Universitäten vorgenommen wird. Pro Universität und Fach werden nicht mehr als drei (in einem Fall vier) Texte aufgenommen, jedoch nicht mehr als zwei Texte pro Erstgutachter*in. Man kann davon ausgehen, dass diese Entscheidungen zur Repräsentativität des Korpus beitragen sollen, aber detailliertere Begründungen werden nicht gegeben. Die Autorin kann mithilfe des HTML-Markups relevante Informationen extrahieren, indem sie formale Kategorien identifiziert (wie Zitate bzw. Beispiele, Fußnoten und Tabellen), die aus der Analyse ausgeschlossen werden. Für die Analyse wird ausschließlich der Haupttext verwendet und es folgt eine OCR-Prüfung und -Korrektur. Als letzter Schritt der Datenaufbereitung erfolgen die automatisierte Tokenisierung der Texte in Tokens und Sätze, die Annotation mit Lemmata und Wortarten sowie die syntaktische Dependenzannotation durch ein Dependenzparsing (vgl. Kübler/McDonald/Nivre 2009). Die Evaluation der automatischen Annotationen wird anhand einer zufälligen Stichprobe von 100 Sätzen aus dem Korpus getroffen. Andresen merkt an, dass die Qualität der Daten akzeptabel ist und dass nur eine automatische Annotation die syntaktische Analyse von Korpora dieser Größe ermöglicht (vgl. S. 133). Aus diesem Grund müsse eine gewisse Fehlerrate in Kauf genommen werden, die durch Stichproben überprüft werden kann.

Die Korpusbeschreibung erfolgt mittels Boxplots zur Darstellung der Distribution formaler Merkmale (bspw. Textlänge, Kapitelanzahl usw.) in den untersuchten Dissertationen. Inhaltliche Merkmale – d. h. das Thema und die Methode – werden ebenso berücksichtigt (vgl. S. 138 f.). Während Dissertationen aus der Linguistik einem Teilbereich der Disziplin zugeordnet werden können, folgt die Zuordnung der Dissertationen aus der Literaturwissenschaft eher einem zeitlichen Kriterium der behandelten literarischen Autor*innen. In Bezug auf die Methode wird in den Dissertationen aus der Linguistik eher zwischen qualitativer und quantitativer Methode unterschieden, wobei einige Texte eine Hybridform besitzen (vgl. S. 139). Die Dissertationen aus der Literaturwissenschaft entsprechen hingegen der Unterteilung in text-, autor-, leser- und kontextorientierte Ansätze (vgl. S. 140). Die Anwendung nicht homogener Zuordnungskriterien für die beiden Disziplinen weist darauf hin, dass zwischen ihnen grundlegende Unterschiede bestehen und dass die Analyse entsprechend angepasst werden sollte. Dies führt zu abweichenden Kategorien in Bezug auf die Zuordnung, erlaubt es aber, für jede Disziplin geeignete Kategorien zu verwenden. Dadurch erscheint die Zuordnung fundierter, als es bei der Anwendung einheitlicher, a priori festgelegter Kategorien der Fall wäre, die die disziplinspezifischen Charakteristika nicht angemessen berücksichtigen würden.

Im **Kap. 7** wird die Methodik illustriert, zu der sowohl die Ermittlung linearer n-Gramme als auch syntaktischer n-Gramme gehört, wobei die Arbeit ausschließlich auf satzinterne sprachliche Strukturen und nicht auf satzübergreifende Strukturen fokussiert. Zusätzlich zu den n-Grammen werden auch Tokens, Wortarten und syntaktische Relationen berücksichtigt. Aufgrund der Textauswahl bei der Erstellung (keine Zufallsstichprobe), aufgrund des Fehlens eines vollständigen Verzeichnisses aller Dissertationen der beiden Fächer und aufgrund des Umfangs des Korpus von 30 Dissertationen pro Disziplin entscheidet sich die Autorin gegen Signifikanztests (vgl. S. 147 f.). Um Muster in Daten zu erkennen, werden stattdessen Methoden des maschinellen Lernens – sowohl ein unüberwachtes als auch ein überwachtes Verfahren – eingesetzt.

Die Ergebnisse werden im **Kap. 8** dargestellt, in dem der Fokus auf lineare und syntaktische Uni- und Trigramme gelegt wird, um das methodische Potenzial unterschiedlicher n-Gramme-Typen zu vergleichen (S. 157). Linguistik und Literaturwissenschaft unterscheiden sich in zahlreichen Aspekten – sowohl auf lexikalischer als auch auf grammatischer Ebene. U. a. ist der unterschiedliche Stellenwert der Methodik besonders relevant: In den Dissertationen aus der Linguistik befinden sich viele Formulierungen auf Tokenebene, die mit dem empirischen Charakter der Disziplin zusammenhängen. Auch das Passiv wird öfters verwendet, was die Linguistik den Wissenschaftssprachen der naturwissenschaftlichen Disziplinen stilistisch annähert. Hingegen sind wenigere Hinweise auf Methoden in den literaturwissenschaftlichen Dissertationen vorhanden. Stattdessen sind Formulierungen zu finden, die mit interpretativen Vorgängen menschlichen Verstehens zusammenhängen (vgl. S. 199 f.). Andere Unterschiede sind bspw. ein höheres Type-Token-Ratio in der Literaturwissenschaft sowie eine stärker formelhafte Sprache in der Linguistik. Der datengeleitete Ansatz wurde in der Analyse angewandt und zeigte, dass die sprachlichen Unterschiede zwischen den Dissertationen der beiden Disziplinen mit außersprachlichen Unterschieden verbunden sind. Allerdings erscheint der Rückgriff auf Theorie im Zusammenhang mit den Ergebnissen weiterhin relevant, insbesondere in den Fällen, in denen die Grenzen eines datengeleiteten Verfahrens deutlich werden – z. B. wenn sprachliche Auffälligkeiten keine Verbindung zu außersprachlichen Bedingungen zeigen oder wenn der datengeleitete Ansatz lediglich oberflächliche Erklärungen für die Auffälligkeit liefert (vgl. S. 200 f.).

Kap. 9 diskutiert die Ergebnisse: Die datengeleiteten Methoden haben es ermöglicht, vieles über die Linguistik und die Literaturwissenschaft sowie deren Unterschiede zu erkennen, doch zeigten sich auch einige Grenzen der angewendeten Methoden. Die Anwendung der syntaktischen Annotationen ermöglicht zusätzliche Erkenntnisse (vgl. S. 207), wobei die Qualität automatischer Annotationen nicht optimal ist, und daher eine Fehlerrate einbezogen werden muss. Andresen zufolge wäre ein stärker regelgeleiteter und automatischer Ansatz bei Annotation für weitere Studien vielversprechend und eine zusätzliche semantische Analyse könnte es erlauben, über die Oberfläche der Texte hinauszugehen und die Methode entsprechend zu erweitern (vgl. S. 208 f.). Der hypothesengenerierende Charakter der n-Gramme ermittelnden Methode wird von Andresen hervorgehoben (vgl. S. 209). Ebenso wäre ein multimodaler Ansatz für künftige Studien sinnvoll, damit die in dieser Analyse ausgeschlossenen Daten wie Abbildungen und Tabellen in die Untersuchung einbezogen werden können (vgl. S. 210).

Im Fazit (**Kap. 10**) hebt Andresen (vgl. S. 212) hervor, dass datengeleitete Forschung und theoriegeleitete Forschung nicht als konkurrierende Ansätze gesehen werden sollten, sondern dass die Kombination beider Methodologien eine Bereicherung für Analysen darstellt. Obwohl Forscher*innen in datengeleiteten Analysen eigentlich mit nicht theoretisch eingebetteten Ergebnissen umgehen sollen, erweisen sich solche aber immer als erklärungsbedürftig und können nicht aus sich selbst heraus interpretiert werden. Die Monografie endet mit einem Anhang (**Kap. 11**), der Tabellen zu dem SST-Label (Schiller et al. 1999) und dem TIGER-Dependenzlabel (Albert et al. 2003) enthält.

Andresens Monografie stellt einen reproduzierbaren, datengeleiteten Ansatz vor. Der Schwerpunkt der Arbeit liegt auf den automatischen syntaktischen Annotationen. Die Autorin stellt fest, dass die automatische syntaktische Annotation einen Ansatz bietet, der linguistische Analysen bereichern kann. Die Methode, die Andresen anwendet, verbindet syntaktische Annotationen mit Token- und Wortartenbeschreibung. Die Kombination dieser Ebenen ermöglicht eine strukturierte und vielschichtige Analyse, die sich als vielversprechend auch für weitere Disziplinen und Textgattungen erweist. Die von Andresen angebotene Perspektive ist breit und verweist auf zahlreiche Studien, die auch für parallele Forschungsvorhaben hilfreich sein können. Die Interdisziplinarität der Methoden und der herangezogenen Literaturquellen ist zweifelsohne eine Stärke dieser Arbeit. Der Eindruck ist stets, dass Andresen nicht nur alle Studien zum Thema, sondern auch solche aus unterschiedlichen Bereichen und Ländern genau kennt. Das trägt dazu bei, dass ihre Monografie für Forscher*innen aus diversen Disziplinen interessant sein kann.

Ein kritischer Punkt betrifft die Anwendung bzw. die Rechtfertigung eines *datengeleiteten* Verfahrens. Dieses wird in der Monografie in mehreren Passagen betont, aber auch mit theoretischen Annahmen kombiniert. Diese Kombination wird von der Autorin ausführlich gezeigt und kommentiert, was die Methode fundiert aufbaut. Allerdings stellt sich die Frage, ob in diesem Fall tatsächlich von einem *datengeleiteten* Verfahren gesprochen werden kann, wenn es in einer hybriden bzw. partiellen Form angewendet wird. Es könnte sinnvoll sein, einen alternativen Begriff zu verwenden, um die methodische Kombination auch terminologisch zu erfassen. So könnte der Ansatz ggf. als *partiell datengeleitet* bezeichnet werden. Weiterhin bleibt die qualitative Perspektive in der Arbeit sowohl methodisch als auch argumentativ unterrepräsentiert. Sie wird teilweise in die Analyse integriert, wird aber nie systematisch vertieft, da die quantitative Untersuchung zweifelsohne im Vordergrund steht. Eine ausführliche qualitative Analyse könnte die Vergleichselemente zwischen den untersuchten Disziplinen ggf. erweitern sowie ein spezifischeres Verständnis der disziplinspezifischen Unterschiede zwischen Linguistik und Literaturwissenschaft in der Germanistik ermöglichen.

Abschließend schreibt Andresen, dass eine Erweiterung auf die mündliche Wissenschaftssprache vorstellbar sei, wobei sich die Register zwischen gesprochener und geschriebener Sprache unterscheiden. Das von Andresen vorgeschlagene Verfahren könnte bei der Anwendung auf gesprochene Sprache zusätzliche Herausforderungen enthalten, insbesondere die Notwendigkeit, die gesprochene Sprache zuerst transkribieren zu müssen. Daher könnte das Verfahren mit weiteren Tools und datengeleiteten Instrumenten kombiniert werden.

Literatur

Albert, Stefanie/Anderssen, Jan/Bader, Regine/Becker, Stephanie/Bracht, Tobias/Brants, Sabine/Brants, Thorsten/Demberg, Vera/Dipper, Stefanie/Eisenberg, Peter/Hansen, Silvia/Hirschmann, Hagen/Janitzek, Juliane/Kirstein, Carolin/Langner, Robert/Michelbacher, Lukas/Plaehn, Oliver/Preis, Cordula/Pußel, Marcus/Rower, Marco/Schrader, Bettins/Schwartz, Anne/Smith, George/Uszkoreit, Hans (2003): *TIGER Annotationsschema*.
URL: www.linguistics.ruhr-uni-bochum.de/~dippper/pub/tiger_annot.pdf

Biber, Douglas (2006): Register: overview. In: Brown, Keith (Hg.): *Encyclopedia of language and linguistics*. 2. Auflage. Amsterdam: Elsevier, 476–482.

Kübler, Sandra/McDonald, Ryan/Nivre, Joakim (2009): *Dependency parsing*. San Rafael: Morgan & Claypool (= Synthesis Lectures on Human Language Technologies 2).

Schiller, Anne/Teufel, Simone/Thielen, Christine/Stöckert, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Tübingen: Universität Tübingen.
URL: https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf