jfml Journal für Medienlinguistik
Journal for Media Linguistics

Vol. 3 (2020), No 1: 7–13
DOI: 10.21248/jfml.2020.37

Rezension

Schneider, Roman (2019): Mehrfach annotierte Textkorpora. Strukturierte Speicherung und Abfrage. Tübingen: Narr Francke Attempto (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Band 8). 315 Seiten. € 98,00 ISBN 978-3-8233-8286-7

Friedrich Markewitz

Hinwendungen zur Korpus- und Computerlinguistik sind inzwischen umfassend erfolgt und diese „von einem technischen Spezialgebiet zu einer Standardmethode avanciert, die in vielen Universitätscurricula fest verankert ist“ (Hirschmann 2019: 2). Der „data-driven turn“ (Bubenhofer/Scharloth 2015: 3) erzwingt aber auch entsprechende Reflexionen des Umgangs mit der zunehmenden Datenfülle: So kann zwar von einer „tiefgreifenden und beinahe flächendeckenden Verdatung der Welt“ (Bubenhofer 2018: 209; vgl. auch Bubenhofer/Scharloth 2015: 1) gesprochen werden, deren Auswertungsmöglichkeiten und -grenzen aber ebenso thematisiert wie reflektiert werden müssen. Dem damit verbundenen Zusammenhang zwischen Korpusgröße und Untersuchungsdesign widmet sich Roman Schneider in seiner Monographie.

Ausgehend von der Herausforderung, dass die „anfallende Menge an Sprachdaten [...] immer öfter unsere Möglichkeiten der Auswertung [übersteigt]“ (15), besteht sein Ansatz darin, „komplexe linguistische Phänomenbeschreibungen in überschaubare, unabhängig voneinander abarbeitbare Aufgaben zu unterteilen“ (18). Dabei geht es dem Autor vornehmlich um Fragen der Laufzeitoptimierung von Suchanfragen. Diese Optimierung – so sein Plädoyer – soll durch „die Segmentierung linguistisch motivierter Suchkriterien als Gegenentwurf zur physischen Segmentierung des Datenbestandes“ (290) erreicht werden. Er vollzieht diese Reflexionen dabei vornehmlich anhand automatisierter bzw. maschineller Annotationen und Suchanfragen.

Kontaktperson:
Friedrich Markewitz
Universität Paderborn
Warburger Straße 100
33098 Paderborn
friedrich.markewitz@upb.de

ISSN: 2569-6491

Schneiders Ausführungen sind insgesamt methoden-reflektierend und empirisch-evaluierend ausgerichtet. Sein Buch zielt innerhalb der Linguistik einen prinzipiell offenen Adressatenkreis an, kann aber von allen Forschenden produktiv rezipiert werden, die mit großen Datenmengen und komplexen Analysekriterien arbeiten.

Das Buch ist in sich logisch und konsistent in sieben Kapitel untergliedert, innerhalb derer in den ersten drei Kapiteln die untersuchungsleitenden Themenzusammenhänge eingeführt und reflektiert werden, um in den sich anschließenden Kapiteln empirisch evaluiert und modifiziert zu werden:

Im Kapitel **Einleitung und Motivation** werden zunächst zentrale Parameter des korpus- oder computerlinguistischen Paradigmas herausgearbeitet. Die Relevanz der Arbeit mit Sprachkorpora „zur Aufdeckung bislang unbekannter bzw. primär introspektiv bewerteter Phänomene“ (30) sei allerdings nur dann gegeben, sofern sie produktiv mit geeigneten Werkzeugen und unter spezifischen Rahmenbedingungen ausgewertet werden können (vgl. 31). Diese problemorientierte Sichtweise führt zum eigentlichen Thema: Da inzwischen „Textkorpora mit Tokenzahlen im zweistelligen Milliardenbereich keine Fiktion mehr“ sind (12), stelle sich immer dringender die Frage nach geeigneten Nutzungsbedingungen dieser Korpora (vgl. 13). Ziel des Buches ist die dahingehende Modellierung und Evaluierung eines Ansatzes zur angemessenen Speicherung und Abfrage mehrfach annotierter (schriftsprachlicher) Sprachkorpora (vgl. 19). Dabei argumentiert Schneider aus einer dezidiert sprach- bzw. informationstechnischen Perspektive (vgl. 18).

Im zweiten, theoretischer ausgerichteten Kapitel zu **Linguistischen Anforderungen an Sprachkorpora** werden konzise einerseits Bedingungen des korpuslinguistischen Arbeitens sowie der korpuslinguistischen Methodik herausgearbeitet und andererseits Überblicke über Aufbau und Funktionalität existierender Korpusprojekte gegeben. Neben der Klärung grundlegender Begriffe, z. B. des Sekundärdaten-Begriffs (der sowohl im Sinne von Annotationen [vgl. 38] als auch im Sinne von Metadaten [vgl. 44] zu verstehen sei), wird konsequenterweise aufgrund der thematischen Anlage des Buches vor allem das Größen-Paradigma reflektiert. Dabei geht Schneider von einem Maximalgrößenansatz aus, kann dies aber überzeugend begründen, sowohl mit dem Hinweis darauf, dass niedrigfrequente Phänomene nur in großen Sprachdatensammlungen aufgefunden werden können (vgl. 33) als auch mit der Warnung, dass „bei der Verwendung zu kleiner Datensammlungen“ (34) die Gefahr „der Überbewertung von Zufallsfunden“ (34) bestehe. Diese klassischen Argumente einer quantitativen Korpuslinguistik sind weithin

zutreffend, doch spielt auch die forschungsfragengeleitete Korpuszusammenstellung eine wichtige Rolle und kann so Schneiders Orientierung an Frequenz- bzw. Verteilungsprobleme bis zu einem gewissen Grad relativieren. Denn es werden auch Plädoyers dahingehend formuliert, mit kleineren Korpora zu arbeiten, deren Auswertung ebenfalls zu produktiven Erkenntnissen führe (vgl. z. B. bei Scherer 2014: 6–8). Auf Prinzipien und Methoden der Generierung kleinerer Korpora geht Schneider aber nicht weiter ein.

Hinsichtlich der Korpuszusammenstellung betont er zudem die Relevanz großer Streuungen, wobei er auf eine hohe Textsortenvielfalt (vgl. 37) sowie „ein möglichst breites Autorenspektrum“ (38) Bezug nimmt. In dieser Hinsicht kommt der medialen Ausprägung der Daten eine wichtige Rolle bei der Korpuszusammenstellung zu. Dass das – mit McLuhan gesprochen – *Medium* die *Message* beeinflusst bzw. beide Größen in einem interdependenten Zusammenhang stehen, wird zwar nicht in dieser Form explizit gemacht, ist aber als Hintergrund-Argument von Schneiders Argumentation erkennbar. Weiterführende Reflexionen, z. B. des *written language bias* und dahingehende kategoriale Unterschiede oder die Notwendigkeit des Hinzuziehens anderer Korpora werden allerdings nicht weiterverfolgt. Die sich anschließenden Analysen und Reflexionen setzen das Kriterium der Korpusgröße zentral. Aspekte der Korpusmodellierung, insbesondere hinsichtlich der unterschiedlichen Medialität der Daten, werden zwar eingangs thematisiert, spielen aber für den Analysegang leider keine große Rolle mehr.

Anhand der Vorstellung verschiedener Korpusprojekte dokumentiert der Autor schließlich den Stellenwert „natürlichsprachlicher Korpora als Arbeitsgrundlage der empirisch arbeitenden Sprachwissenschaft“ (57). Eine umfassende und durch reiche Bebilderung hervorragend unterstützte Darstellung von Recherchemöglichkeiten in ausgewählten Korpussammlungen (DeReKo/Cosmas [vgl. 63–68], LCC [vgl. 68–73] und dem DWDS [vgl. 73–77]) wird dazu genutzt, zur Reflexion multidimensionaler Suchkriterien überzuleiten (vgl. 77–95). All diese in hohem Maße transparenten Zusammenhänge führen, die Ausführungen synthetisierend, in die Formulierung eines Anforderungskatalogs (vgl. 97–98), der sich aus „informationstechnologischer Perspektive in vier Sektionen gliedert“ (96):

a) Suchmuster aus diskreten Elementen mit oder ohne Platzhalterzeichen,
b) als lineare Verkettung mehrerer Einzelelemente über deren relative Position,

c) aus linearen Abfolgen sowie hierarchischen Annotationsmerkmalen und
d) mit regulären Ausdrücken (vgl. 97–98).

Der Katalog wird dann im weiteren Verlauf angewandt und kritisch überprüft.

Bevor es dazu kommt, werden im dritten Kapitel Reflexionen zu **Design und Implementierung eines Korpusabfragesystems** vorangestellt, um Bedingungen wie Möglichkeiten der Bearbeitung der Datenmengen zu thematisieren. Nach einer knappen aber konsequenten Einführung in Zusammenhänge des Designs sowie der Speichermöglichkeiten von Korpusverwaltungssystemen (vgl. 102–110) verengt Schneider seinen Fokus auf „die Praktikabilität relationaler Datenbanktechnologien für die Verwaltung von Mehrebenen-Korpora“ (111). Im ersten Schritt führt er in prototypische Aspekte relationaler Datenbanksysteme ein (wie die Behandlung von Primär- und Sekundärdaten [vgl. 111–115], die konzeptuelle Datenmodellierung [vgl. 115–119], ein physisches Datenbankschema [vgl. 119–125], die Hard- und Software [vgl. 125] sowie Datenimporte [vgl. 125–135]), um dann einzelne Designentscheidungen zu evaluieren. Hinsichtlich der Datenmodelle plädiert er für tokenorientierte Relationierungen statt für N-Gramm-Tabellen, da erstere, trotz längerer Abfragezeiten (vgl. 153), deutlich umfangreichere Abfragemöglichkeiten eröffnen (vgl. 174). Platzhalteroperatoren und reguläre Ausdrücke werden als sinnvoll angesehen, wenn Indizes angelegt sind (vgl. 175). Im Rahmen der Reflexion numerischer und textueller Schlüsselwerte sieht der Autor keine Laufzeitvorteile (vgl. 165). Die Abfrage hochfrequenter Phänomene schließlich profitiert von einer Auslagerung in „separate Tabellen“ (173). Diese Erkenntnis der Relevanz der Segmentierung von (Teil-)Abfragen zur Verbesserung der Abfrage(lauf)zeit ist für den Autor zentral und wird auch für den weiteren Verlauf leitend sein.

Nach diesen Reflexionen kommt Schneider im vierten Kapitel zur **Evaluation des Anforderungskatalogs** (aus Kapitel 2). Anhand von sechs Evaluationskorpora (mit zwischen einer Million und acht Milliarden Textwörtern) will der Autor „Zusammenhänge zwischen wachsenden Suchraum-Datenmengen, Tabellenverknüpfungen, Belegzahlen und Retrievalzeiten“ (177) genauer thematisieren. Ein weiteres Mal zeigt sich, dass Schneider vornehmlich an Korpusgrößen interessiert ist. Weiterführende Thematisierungen des Aufbaus der Korpora (z. B. nach Textsorten oder nach mündlichen oder schriftsprachlichen Daten) finden nicht statt.

Für die Evaluation greift er auf die vier Sektionen seines Anforderungskatalogs zurück (vgl. 177–202). Die einzelnen Schritte

werden – und dies zeichnet das Buch insgesamt aus – so detailliert und transparent wie möglich dargestellt. Die Vielzahl an Visualisierungen (durch Tabellen sowie Abbildungen) erleichtert das Verständnis und vermag auch unkundigeren Rezipierenden mit Erkenntnisgewinn den Analysegang zugänglich zu machen.

Insgesamt kommt Schneider zum Ergebnis, dass ein signifikanter Zusammenhang zwischen den Suchattributen sowie der „Abfragekomplexität und Abfragedauer" (214) bestehe – und zwar unabhängig von der Anzahl der Belegtreffer (vgl. 214). Dies sei bei wenigen Suchattributen noch kein großes Problem. Werden diese allerdings erhöht, „erreichen die ermittelten Laufzeiten ein für die Recherchepraxis weniger ansprechendes Niveau" (214). Aus diesem Befund leitet er die Notwendigkeit der „Modifikation des Recherchemodells" (214) ab, verweist aber zusätzlich auf die Relevanz der Beachtung von Hardwareentscheidungen: Leistungsfähigere Hardware würde sich ebenfalls positiv auf die Abfragedauer auswirken (vgl. 215), z. B. durch die „Nutzung von leistungsfähigeren [...] Mikroprozessoren" (217).

Abgeleitet aus der Evaluation des Anforderungskatalogs entwickelt der Autor im fünften Kapitel den **Versuch einer Laufzeitoptimierung durch segmentierte Abfragen**. Mit Rekurs auf den „Map-Reduce-Ansatz" (221), also der Aufteilung komplexer Anfragen in singuläre aber parallele Abfragehandlungen, die am Ende wieder zusammengeführt werden (vgl. 221–224), präsentiert er (s)einen Ansatz der segmentierten Modellierung. Dabei geht es ihm nicht nur darum, „die Eingabedaten, sondern die zu erledigenden Aufgabenstellungen" (225) aufzugliedern. Anhand der Optimierung von Abfragetypen

a) auf Wortebene,
b) unter Einbeziehung textbezogener Metadaten sowie
c) unter Einbeziehung syntaktischer Strukturen (vgl. 232)

wird der Segmentierungsansatz umfassend dargestellt. Trotz unterschiedlicher Ergebnisse hinsichtlich der Zeitersparnis und der Feststellung, dass es nicht zu einer „proportional linear[n] Verbesserung der Abfragezeiten" (298) komme, lasse sich durch Segmentierungen doch „eine signifikante Reduzierung der Suchzeiten für mittlere und große Korpora feststellen" (261). Damit scheint die Aufteilung komplexer Abfrageschritte ein sinnvoller Umgang mit großen Korpora zu sein.

Im sechsten Kapitel werden Möglichkeiten der **Integration in ein Online-Framework** diskutiert. Nach Hinweisen auf charakt-

eristische Architekturzusammenhänge (vgl. 275) geschieht dies anhand von vier Elementen (Suchformularen [vgl. 275–278], der Speicherung von Beleglisten [vgl. 278–280], Schnittstellen zu Statistikwerkzeugen [vgl. 280–282] sowie Übersichtslisten [vgl. 282–284]). Eine konzise **Zusammenfassung** schließt das Buch konstruktiv ab.

Roman Schneiders Arbeit zum Umgang mit großen Sprachdatensammlungen überzeugt sowohl hinsichtlich der nachvollziehbaren Entfaltung seines Analysegegenstandes als auch bezogen auf die so ausführliche wie transparente Darstellung der Schritte seines methodischen Designs sowie der Konstruktion, Evaluation und Modifikation seines Modellierungsansatzes. Dieser setzt an einer korpuslinguistisch wichtigen Schnittstelle zwischen Korpus und Analysedesign an und zeigt die Relevanz, nicht nur der Passung von Korpus und Forschungsfrage, sondern insbesondere zwischen Korpusgröße und methodischem Vorgehen der Modellierung des Forschungs- bzw. Auswertungsprozesses. Dabei werden oft wenig thematisierte Aspekte (wie z. B. die Abfragelaufzeit und die ‚Überforderung' bisheriger korpuslinguistischer Tools) angesprochen und konstruktiv Lösungen aufgezeigt, dem Missverhältnis zwischen wachsenden Korpusgrößen und fehlender Angepasstheit sowie Leistung der verwendbaren Tools zu begegnen.

Die „korpuslinguistische Brille" (285) ist dabei ein grundlegend gesetztes Paradigma. Kritische Reflexionen finden sich nicht, seine Produktivität gilt als umfassend bewiesen und wird – insbesondere in den ersten Kapiteln – auch hervorgehoben. Dabei verweist Schneider zwar zu Beginn auf das Zusammenspiel zwischen quantitativen und qualitativen Analyseschritten. In seinen weiteren Ausführungen wird aber nicht weiterführend darauf eingegangen. Dabei ist das Buch sehr offen angelegt, geht es doch um die Thematisierung und Reflexion der Modellierung eines Analyse- bzw. Abfrageansatzes, der aber sowohl qualitativ, quantitativ als auch beide Herangehensweisen integrierend ausfallen kann.

Dabei schafft es der Autor durch seine transparente Themenführung und seinen zugleich präzisen aber nachvollziehbaren Stil ein komplexes Thema darzustellen und mit Erkenntnisgewinn Lösungen zu skizzieren. An mittleren sowie größeren Korpora arbeitenden Forschenden, die sich insbesondere der eigenständigen Modellierung von Suchdesigns sowie Abfragemöglichkeiten zuwenden, sei das Buch daher umfassend empfohlen.

Literatur

Bubenhofer, Noah (2018): Diskurslinguistik und Korpora. In: Warnke, Ingo H. (Hg.): *Handbuch Diskurs*. Berlin/Boston: de Gruyter, 208–241.

Bubenhofer, Noah/Scharloth, Joachim (2015): Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. In: *Zeitschrift für germanistische Linguistik* (43), 1–26.

Hirschmann, Hagen (2019): *Korpuslinguistik. Eine Einführung*. Berlin: J.B. Metzler.

Scherer, Carmen (2014): *Korpuslinguistik*. Zweite, aktualisierte Ausgabe. Heidelberg: Universitätsverlag Winter.