Durrant, Philip: **Corpus Linguistics for Writing Development. A Guide for Research.** Abingdon: Routledge, 2023 (Routledge Corpus Linguistics Guides). – ISBN 978-0-367-71578-6. 194 Seiten, € 49,99.

Besprochen von **Daniel Jach:** Chengdu / VR China

https://doi.org/10.1515/infodaf-2024-0021

Mit seinem Buch *Corpus Linguistics for Writing Development*, erschienen 2023 bei Routledge, will Durrant fortgeschrittenen Anfängern in den Forschungsgebieten Erziehung, Sprache und Linguistik einige ausgewählte korpuslinguistische Techniken für die Untersuchung der Schreibentwicklung von Sprachlernenden und die entsprechenden Programmierkenntnisse in R vermitteln (Seite i). Vorkenntnisse in R seien nicht erforderlich (12). Er selbst sei, so Durrant, Associate Professor für Spracherziehung an der südwestenglischen University of Exeter, ein Autodidakt und „far from being a professional computer programmer" (18). Das Buch wolle lediglich zeigen, „how some analysis I take to be central to writing development research can be carried out in R" (18), sei aber weder eine allgemeine Einführung in die Korpuslinguistik noch eine Statistikeinführung. Dieser offene und unprätentiöse Einstieg ist sympathisch, wirft aber doch gleich zu Beginn die irritierende Frage auf, warum sich angehende KorpuslinguistInnen eigentlich für Durrants Buch entscheiden sollten und nicht etwa für eine der umfangreicheren Einführungen von Gries (2016, 2021) oder für einen der anderen, ganz ähnlichen Bände aus derselben Reihe. Im Unterkapitel *Why learn R* berichtet Durrant dann von seinen ersten eigenen Programmiererfahrungen, der steilen Lernkurve und der großen Zeitersparnis, bezieht sich dabei aber zum Erstaunen des Lesers nicht auf R, sondern auf die Programmiersprache Python (17), die allgemein als leichter erlernbar und besser für die Textanalyse geeignet gilt. Warum also nicht stattdessen *Python for Linguists* von Hammond (2020) lesen? Diese anfänglichen Zweifel bleiben auch nach der Lektüre bestehen: Das Buch hat seine Stärken, wirkt aber insgesamt unausgewogen, ist in den praktischen Teilen nicht zu Ende gedacht und als Lehrbuch für Einsteiger daher nur bedingt geeignet.

Nach einigen Sätzen zur gesellschaftlichen Relevanz von linguistischer Forschung führen die ersten beiden Kapitel in die Grundprinzipien der Korpuslinguistik, ihre forschungsmethodischen Stärken und Schwächen (Kapitel 1), Korpora und die Programmiersprache R (Kapitel 2) ein. Vor allem das zweite Kapitel will viel erreichen: Der Leser wird mit R, der Verwendung von Terminal und Skripten, der Entwicklungsumgebung RStudio und einigen grundlegenden Befehlen bekannt gemacht, zwei englischsprachige Lernerkorpora (GiG, BAWE) und

ein Parser (ein Computerprogramm zur automatischen grammatischen Analyse von Texten) werden vorgestellt und für spätere Analysen heruntergeladen, und wer möchte, der kann versuchen, sich ein eigenes Korpus aus Internettexten zusammenzustellen, und wird anschließend beim automatischen Parsen angeleitet. Das Kapitel hat die entscheidende Aufgabe, die technischen Grundlagen für den Rest des Buches zu legen, tut dies aber nur halbherzig. Die Vermittlung der Funktionsweise von R beschränkt sich auf die Beschreibung einiger Grundbefehle zum Abtippen und Ausführen, und kommt ansonsten ohne Übungen aus. Obwohl Durrant richtigerweise anmerkt, dass „[n]ovice researchers often create their first corpus by downloading texts from websites“ (23), wird die entsprechende Technik (Webscraping) nicht vermittelt, sondern auf eine umfangreiche externe Ressource verwiesen. Der Download des GiG-Korpus scheitert nach Anfrage und Wartezeit an der notwendigen Online-Registrierung. Auf dieser etwas unsicheren Grundlage baut der folgende Hauptteil des Buches auf.

Der Hauptteil besteht aus drei Abschnitten, die jeweils eine andere Perspektive auf die Entwicklung und Erforschung von Schreibkompetenz einnehmen: Wortschatz, Grammatik und sprachliche Muster *(formulaic language)*. Jeder Abschnitt besteht wiederum aus zwei Kapiteln, von denen das erste den Stand der Fachdiskussion grob umreißt, zentrale Forschungsgegenstände diskutiert und ihre korpuslinguistischen Operationalisierungen erläutert, während das zweite einige dieser Operationalisierungen auswählt und ihre praktische Umsetzung mit Korpusdaten und R-Code demonstriert. Im dritten Kapitel werden beispielsweise verschiedene Maße zur korpuslinguistischen Bestimmung der lexikalischen Entwicklung von Lernenden ausführlich erläutert. So kann z. B. mit der Type-Token-Ratio die lexikalische Vielfalt *(lexical diversity)* eines Textes bestimmt werden, während die Wortlänge, die Worthäufigkeit und die Verteilung der Wörter auf Kontexte und Register die lexikalische Differenziertheit *(lexical sophistication)* eines Lernertextes abbilden. Das Kapitel schließt mit einigen Vorschlägen für die weiterführende Lektüre zum Thema, den Endnoten und einer umfangreichen Literaturliste. An diese inhaltliche Einführung schließt sich das anwendungsorientierte Kapitel 4 an. Hier wird anhand von Daten aus dem GiG-Korpus exemplarisch ermittelt, wie sich die Type-Token-Ratio und der Anteil des wissenschaftssprachlichen Vokabulars in Lernertexten mit zunehmender Schreibkompetenz der Lernenden verändern. Die Ergebnisse werden in Grafiken visualisiert. Der für die Analyse und die Visualisierung benötigte R-Code ist im Buch in Kästen zum Abschreiben abgedruckt oder kann online als gebrauchsfertiges Skript heruntergeladen werden. Eine Liste der verwendeten R-Funktionen schließt das Kapitel ab.

Die folgenden Abschnitte und Kapitel sind analog aufgebaut. Der dritte Abschnitt geht in Kapitel 5 auf drei Ansätze der Grammatikforschung ein (Unter-

schiede in der Komplexität grammatischer Strukturen, die Verteilung grammatischer Merkmale über Register und Textsorten in Clustern und die Entstehung grammatischer Strukturen aus dem Sprachgebrauch), um dann in Kapitel 6 das geparste GiG-Korpus nach attributiven Adjektiven (einem Maß für grammatische Komplexität) zu durchsuchen. Der vierte Abschnitt behandelt sprachliche Muster, erläutert ihre allgemeine Bedeutung für Spracherwerb und Sprachgebrauch, unterscheidet lexikalische Sequenzen mit Diskursfunktion (*lexical bundles*, z.B. *im Großen und Ganzen*) von Kollokationen (miteinander assoziierte Wörter, z.B. *einen Fehler begehen*) (Kapitel 7) und führt anschließend durch den Code zur Identifizierung wissenschaftssprachlicher Kollokationen im BAWE-Korpus (Kapitel 8). Das Buch endet abrupt, ohne Schlusskapitel, Aufgaben zur Vertiefung oder Inspirationen für eigene Projekte.

Durrant kennt sich aus, daran besteht kein Zweifel. Die Inhaltskapitel (Kapitel 1, 3, 5 und 7) sind dicht, aber verständlich und trotz ihrer Kürze differenziert geschrieben. Die langen Literaturlisten am Ende jedes Kapitels sind auf dem neuesten Stand und bieten genügend Lesestoff, um sich weit über die Grundlagen hinaus in die jeweiligen Themen zu vertiefen. Auch theoretisch und methodisch schwierige Fragen werden aufgegriffen und reflektiert. Was ist ein Wort und wie wird es in einem Korpus gezählt? Wie sollen feste Mehrworteinheiten (z.B. *Grüß Gott, Vielen Dank, Mund-zu-Mund-Propaganda*) bei der Korpusrecherche behandelt werden? Warum treten Wörter überzufällig häufig zusammen auf und ab welcher Anziehungskraft ist von einer Kollokation auszugehen? Welche Länge sollte für lexikalische Sequenzen angenommen werden? Warum korrelieren bestimmte sprachliche Dimensionen miteinander und wie ist damit umzugehen? Diese und ähnliche Fragen heben das Niveau der kurzen, aber fundierten Inhaltskapitel immer wieder in unerwartete Höhen.

Die Anwendungskapitel (Kapitel 2, 4, 6 und 8) können dieses Niveau nicht halten. So folgt auf eine differenzierte Diskussion verschiedener Verfahren zur Bestimmung der lexikalischen Vielfalt in Kapitel 3 die vergleichsweise einfache Ermittlung der korrigierten Type-Token-Ratio *(corrected type-token ratio)* in Kapitel 4. Nach der sachkundigen und anregenden Einführung in verschiedene Grammatiktheorien in Kapitel 5 werden in Kapitel 6 bloß Tags in den Annotationen des Parsers gezählt und in einem langwierigen Prozess werden die Genauigkeit *(precision)* und die Relevanzrate *(recall)* der Annotationen ermittelt. Dies ist in einem Lehrbuch für Anfänger verständlich, hinterlässt aber dennoch einen schiefen Eindruck. Die Bestimmung von wissenschaftssprachlichen Wörtern und Kollokationen in den Kapiteln 4 und 8 erscheint dagegen anspruchsvoller, geht aber programmiertechnisch kaum über das vorher bereits Eingeführte hinaus. Die Vermittlung von R bleibt auch im Hauptteil des Buches für Anfänger eine Herausforderung. Anstatt gemeinsam mit den Lesern den Code Zeile für Zeile zu ent-

wickeln, wird häufig ein fertiges Skript abgedruckt, das der Leser abschreiben oder herunterladen und ausführen soll, um aus den nachfolgenden Erklärungen zu erfahren, was eigentlich gerade geschehen ist. Die Skripte folgen immer dem gleichen Format und bleiben (wohl in didaktischer Absicht, vgl. 18) weitgehend auf das R-Basispaket *(R base)* beschränkt, was sie teilweise umständlich und nicht ohne weiteres verallgemeinerbar macht. Jedenfalls erscheint es unwahrscheinlich, dass Anfänger nach der Lektüre in der Lage sein werden, ihre eigenen Textdaten selbstständig zu erheben, aufzubereiten und auszuwerten. Ein gängiges Erweiterungspaket wie *dplyr* (Wickham u.a. 2023) hätte einige der Skripte intuitiver und kompakter gestaltet, prinzipiell nützliche Funktionen eingeführt und Gelegenheit gegeben, das Verständnis von Datenmanipulation im Allgemeinen zu verbessern.

So entsteht ein insgesamt unausgewogener Eindruck. Dichten und kenntnisreichen Einführungen in die korpuslinguistische Erforschung der Schreibentwicklung von Sprachlernern stehen anwendungsorientierte Kapitel gegenüber, deren Zielsetzung oft vergleichsweise schlicht ist, deren R-Skripte eher umständlich sind und die keine Gelegenheit zur vertieften Aneignung des Codes bieten. Für Einsteiger ohne Programmierkenntnisse ist Durrants Buch daher nur eingeschränkt zu empfehlen.

Literatur

Gries, Stefan Th. (2016): *Quantitative Corpus Linguistics with R. A Practical Introduction.* 2. Auflage. Abington: Routledge.

Gries, Stefan Th. (2021): *Statistics for Linguistics with R. A Practical Introduction.* 3., überarbeitete Auflage. Berlin: De Gruyter.

Hammond, Michael (2020): *Python for Linguists*. Cambridge: Cambridge University Press.

Wickham, Hadley; François, Romain; Henry, Lionel; Müller, Kirill; Vaughan, Davis (2023): *dplyr: A Grammar of Data Manipulation*. Online: https://dplyr.tidyverse.org (18.12.2023).