Eszter Mózes:
Rainer Perkuhn, Holger Keiberl & Marc Kupietz: Korpuslinguistik
Argumentum 10 (2014), 102-105
Debreceni Egyetemi Kiadó

Recenzió

Eszter Mózes

Rainer Perkuhn, Holger Keiberl & Marc Kupietz: Korpuslinguistik

Paderborn: Wilhelm Fink Verlag, 2012, 144 Seiten

Das vorliegende Buch wurde im Rahmen der Reihe LIBAC – Linguistik für Bachelor 2012 veröffentlicht. Das Buch beabsichtigt nicht, eine ausführliche Beschreibung des Wissenschaftsbereichs zu geben, sondern will vor allem einige ausgewählte, bestehende Problemgebiete darstellen, mögliche Lösungen, Methoden anbieten und Gedanken für weitere Diskussionen wecken. Als Band der Reihe LIBAC wurde *Korpuslinguistik* für Bachelor-Studenten der Sprachwissenschaft konzipiert und sollte in einem Semester bearbeitet werden. Da der Umfang des Buches zur Knappheit der Ausführungen zwingt, bieten die Autoren auf ihrer Homepage begleitende Materialien an, wo vertiefende Erklärungen, interaktive Lehrangebote zur Exploration der im Buch vorgestellten Konzepte, Empfehlungen für korpuslinguistische Hausarbeiten, Hinweise auf weiterführende Literatur und Hilfsmittel sowie eine Errata-Liste zum Buch zu finden sind.

Das Buch stellt Korpuslinguistik auf 144 Seiten und in 9 Kapiteln vor. Im Folgenden möchte ich die einzelnen Kapitel in Grundzügen vorstellen.

Das erste Kapitel beschäftigt sich mit der Grundfrage, was Korpuslinguistik überhaupt ist. Die Autoren wollen keine exakte Definition angeben, da sie meinen, dass dafür schon genügend andere Bücher zur Verfügung stehen wie zum Beispiel Scherer (2006) oder Lemnitzer & Zinsmeister (2006). Stattdessen steht eher Sprache als Gegenstand der Korpuslinguistik im Mittelpunkt. Die Frage ist, wie dieser Gegenstand am besten untersucht werden kann. Die Autoren stellen kurz verschiedene Ansätze und Methoden vor wie zum Beispiel Neuro- und Psycholinguistik, Introspektion, Befragung oder Experiment, die bei der Beantwortung der Frage helfen könnten, aber sie sind der Meinung, dass Performanz, also der Sprachgebrauch die Sprache am besten erfassen kann. Emergenz als Schlüsselbegriff für das ganze Buch wird ebenfalls hier eingeführt: alles Wichtige über die Sprache lässt sich aus ihrem Gebrauch erschließen. Die Autoren wollen die Spuren der Emergenz im Sprachgebrauch finden und untersuchen.

Es wird betont, dass Korpuslinguistik die Sprache selbst untersucht und mehr über die Sprache erfahren will. Nach der Auffassung der Autoren bedeutet Korpuslinguistik als Methodologie, dass die Strukturen der Sprachen mithilfe eines Korpus aufgedeckt werden, das Korpus dient also als Ausgangspunkt der Untersuchung und setzt sich nicht das Ziel, frühere Thesen oder Theorien zu bestätigen.

Im zweiten Kapitel bekommen wir einen kurzen Einblick in die elementaren Einheiten der Korpuslinguistik und in die einfache Recherche. Die Einheiten wie Text, Absatz, Wort und

Eszter Mózes:
Rainer Perkuhn, Holger Keiberl & Marc Kupietz: Korpuslinguistik
Argumentum 10 (2014), 102-105
Debreceni Egyetemi Kiadó

Zeichen werden einer nach dem anderen diskutiert, wobei unsere Aufmerksamkeit wieder darauf gelenkt wird, dass es zahlreiche Definitionen für diese Begriffe gibt und wir klären sollten, welche Definition wir benutzen oder welche in der Software kodiert wird bevor wir mit der Recherche selbst beginnen.

Die kleinste Einheit, die im Korpus meistens gesucht wird, ist das Wort. Um ein Wort zu finden, wird das Korpus in Worttokens aufgeteilt (tokenisiert), und ein Index wird hergestellt, in dem diese Wortformen und ihre Vorkommen aufgelistet werden. Wenn ein Wort gesucht wird, wird zuerst das Index durchsucht, in dem verzeichnet ist, wo das Wort genau vorkommt. Die Autoren empfehlen dem Leser, reguläre Ausdrücke zu benutzen. Die dazu gehörenden Grundbegriffe und Grundkonzepte werden eingeführt und kurz beschrieben. An einigen Beispielen wird gezeigt, wie und wofür sie bei einer Suche im Korpus verwendet werden können. Weiter Quellen werden empfohlen, wo man sein Wissen über reguläre Ausdrücke vertiefen kann. Die Aufmerksamkeit des Lesers wird auf das Problem gerichtet, dass die Treffer einer Suche nicht nur solche enthalten, die man sich ursprünglich als Ziel gesetzt hat. Für die qualitative Bewertung der Treffermenge werden die Begriffe precision und recall eingeführt. Wir sollten die Recherche, also die Suchformel verfeinern, bis wir beinahe nur die erwünschten Treffer bekommen. Dabei spielt es eine Rolle, ob wir die Menge der zu Recht Gefundenen (precision) oder die Menge der intendierten Treffer (recall) optimaliseren möchten.

Kapitel 3 beschäftigt sich mit dem Thema, wie man Sprache sammeln kann. Die wichtigsten Eigenschaften eines Korpus werden aufgezählt: z.B. Repräsentativität und Ausgewogenheit. Die Autoren weisen darauf hin, dass wir bei einer Recherche zuerst versuchen sollten, ein existierendes Korpus zu finden, das unseren Zielen entspricht; die meisten Korpora verfügen über Teilkorpora, was uns ermöglicht, eigene Korpora aus verschiedenen Teilkorpora zu erstellen, um spezifische Fragestellungen zu untersuchen, etwa Unterschiede zwischen den geographischen Sprachgebräuchen (z.B. Deutschland vs. Österreich). Wenn wir doch ein eigenes Korpus zusammenstellen müssen, sollten wir viele Kriterien im Auge behalten, z.B., je größer das Korpus ist, desto besser kann es die Sprache repräsentieren, besonders die seltenen Formen kommen mit größerer Wahrscheinlichkeit vor. In unserer digitalen Welt ist es ganz einfach geworden, genügend Texte herunterzuladen und zu speichern. Ein wichtiger Rat ist, dass immer darauf geachtet werden muss, dass Urheberrechte geklärt werden, weil sonst die gesammelten Texte nicht genutzt werden können. Außerdem werden die zwei Hauptarten von Korpora kurz beschrieben: dynamisches Korpus (wo das Korpus immer erweitert wird, aber die Balance erhalten bleibt) und Monitorkorpus (wo das Ziel es ist, immer eine momentane Aufnahme der Sprache zu geben).

Im Kapitel 4 werden verschiedenen Lösungen behandelt, Korpora mit zusätzlichen Informationen anzureichern. Diese Informationen können unsere Arbeit mit dem Korpus erleichtern und verbessern, sowohl bei der Suche als auch bei der Interpretation. Je nach Ebene und Gegenstand können verschiedene Typen von Anreicherung unterschieden werden. Ein wichtiger Typ ist die morphosyntaktische Annotation, die angibt, zu welcher Wortart ein Wort gehört. Diese Annotationen werden teilweise automatisch, teilweise von Menschen durchgeführt. Wie gut eine Annotation ist, kann wiederum mit den schon vorgestellten Indikatoren (Präzision und Recall) gemessen werden. Die Autoren betonen, dass wir diese Annotationen kritisch betrachten sollten, da Subjektivität nicht ausgeschlossen werden kann, wenn Menschen an der Annotation beteiligt sind. Am besten ist es, mehrere Arten von Annotationen anzuschauen und zu vergleichen. Zum Schluss wird an einem Beispiel gezeigt, wie diese zusätzlichen Informationen kodiert und standardisiert werden können.

Eszter Mózes:
Rainer Perkuhn, Holger Keiberl & Marc Kupietz: Korpuslinguistik
Argumentum 10 (2014), 102-105
Debreceni Egyetemi Kiadó

Im Kapitel 5 erfahren wir, was wir außer der Existenz eines Wortes von den gefundenen Treffern ablesen können oder worauf wir schließen können, wenn wir keinen Treffer haben. Die Autoren betonen, dass wir wegen der korpusorientierten Herangehensweise alle Treffer gleich ernst nehmen sollten, aber es kann trotzdem vorkommen, dass bei der Untersuchung verschiedener sprachlicher Phänomene einige Treffer als relevant, andere als irrelevant bewertet werden können. Was unter Relevanz verstanden wird, hängt immer von der Fragestellung ab. Die Autoren machen darauf aufmerksam, dass wir keine negativen Schlussfolgerungen machen dürfen, wenn ein Wort im Korpus nicht gefunden wird; da die Größe des Korpus eingeschränkt ist, kann passieren, dass eigentlich existierende Wörter einfach nicht in ihm vorkommen. Die Autoren sind der Meinung, dass Frequenzlisten die meisten Informationen über ein Korpus enthalten. Weiterhin stellen Sie Probleme vor, die auftauchen, wenn die Wörter bei der Erstellung einer Frequenzliste verschiedenen Einheiten – z.B. Lemmas – zugeordnet werden können.

Kapitel 6 diskutiert, wie Korpushäufigkeiten gemessen, repräsentiert und interpretiert werden können bzw. wie sie miteinander verglichen werden können. Die Autoren zeigen uns drei Häufigkeitsmaße (absolute Häufigkeit, relative Häufigkeit, Häufigkeitsklassen), deren Wahl von dem Ziel der Untersuchung, unserer Vorliebe und unseren Annahmen über den Forschungsgegenstand abhängt. Sie zeigen uns an einfachen, ohne besondere statistische Vorkenntnisse verständlichen Beispielen, wo und wie welcher Maß am besten gebraucht wird. Die Aufmerksamkeit des Lesers wird auf die Tatsache gerichtet, dass man sich in Fällen, wenn man nach verschiedenen Häufigkeitsmaßen über die ganze Sprache etwas aussagen möchte, zuerst unbedingt vergewissern muss, ob diese Frequenzen zuverlässig sind, d.h. ob sie die ganze Gesamtheit abbilden. Für die Kontrolle dieser Tatsache führen die Autoren zwei Grundbegriffe ein: Konfidenzinterwall und Konfidenzniveau. Diese Begriffe werden anspruchsvoll erklärt, und deren Berechnung wird an einfachen, nachvollziehbaren Beispielen gezeigt. Weiterhin wird noch der Begriff Dispersion eingeführt, wobei die Autoren darstellen, wie zwei Wörter mit der gleichen Frequenz in einem Korpus verglichen werden können.

Im Kapitel 7 werden Assoziationsbeziehungen unter die Lupe genommen und die Frage gestellt, wie sie analysiert und in der Korpuslinguistik benutzt werden können. Als Definition für Assoziation wird Folgendes angegeben: „Eine (positive) Assoziation zwischen einem Wort und einem (außer-)sprachlichen Objekt besteht dann, wenn das Wort überzufällig häufig mit diesem anderen Objekt gemeinsam vorkommt d.h. je nach Objekttyp: in denselben Texten bzw. an denselben Textstellen." (S. 101) Die Autoren erklären, dass man mit Hilfe der Assoziationen viel über den Gebrauch eines bestimmten Wortes erfahren kann. Sie sagen, dass diese Analysen behilflich sein können, wenn z.B. in einem Wörterbuch ein Wortartikel geschrieben wird, wo neben den häufigsten Verwendungen auch die Nebenbedeutungen oder Neubedeutungen aufgelistet werden. An zahlreichen Beispielen wird gezeigt, wie und in welchen Dimensionen diese Untersuchungen ausgeführt werden können. Mit einer zeitlichen Analyse kann z.B. geprüft werden, wann und welche neue Bedeutungen ein Wort bekommen hat.

Kapitel 8 beschäftigt sich mit einem Phänomen, das nach der Meinung der Autoren in der traditionellen Sprachwissenschaft ziemlich vernachlässigt ist: Muttersprachler bevorzugen bestimmte Formulierungen im Gegensatz zu anderen, die grammatisch auch möglich wären. Die Autoren sind davon überzeugt, dass dieses Phänomen einen wichtigen Teil des Verstehens der Sprache als Untersuchungsgegenstand ausmacht. Künstlich kreierte Sprachen können die Annahme untermauern, dass für das Verstehen eines Satzes zwei Voraussetzungen erfüllt sein müssen: einerseits muss die Struktur verstanden werden, andererseits brauchen wir

Eszter Mózes:
Rainer Perkuhn, Holger Keiberl & Marc Kupietz: Korpuslinguistik
Argumentum 10 (2014), 102-105
Debreceni Egyetemi Kiadó

unser Wissen über die einzelnen Wörter. Dabei spielt die Struktur eine größere Rolle; z.B. sind Sätze in der Schlumpfsprache gut zu verstehen trotz noch nie gehörter Wörter, dank der bekannten Strukturen. Kontext spielt eine sehr wichtige Rolle, weil die Nachbarwörter eines Wortes viel über die Wortklasse und die Bedeutung verraten können. Es wird im Weiteren ein Instrument vorgestellt, mit dessen Hilfe Kookkurrenzen analysiert werden können. Es kann angegeben werden, nach welchen Kriterien die Vorkommen eines Wortes sortiert werden, mit ihrer Hilfe können sprachliche Strukturen untersucht werden: zum Beispiel können syntagmatische Muster entdeckt werden.

Im letzten Kapitel werden ähnliche Wörter und ihr Gebrauch diskutiert. Nach der Auffassung der Autoren sind Wörter ähnlich, wenn sie in den gleichen Kontexten gebraucht werden können. Diese Ähnlichkeit existiert auf verschiedenen Ebenen. Die Wörter, deren Bedeutung einem bestimmten Wort ähnlich ist, können in Form von selbstorganisierenden Karten dargestellt werden, wo der Grad der Ähnlichkeit geometrisch visualisiert wird. Mit Hilfe dieser Methode können Profile der Wörter dargestellt werden, wo man erfährt, welche unterschiedlichen Bedeutungen ein Wort hat.

Zusammenfassend kann man sagen, dass dieses Buch die am Anfang gesetzten Ziele erreicht hat. Es ist leicht zu verstehen und ist logisch aufgebaut. Die Autoren wollen sicher gehen, dass die vorgestellten Begriffe von dem Leser verstanden werden, deswegen versuchen sie die linguistischen Phänomene auch mit alltäglichen Beispielen zu illustrieren. Am Ende jedes Kapitels gibt es Aufgaben sowohl zu theoretischen als auch zu praktischen Aspekten des jeweiligen Themas. Wegen diesen Eigenschaften eignet sich das Werk als Lehrbuch für Bachelor-Studierende der Sprachwissenschaft.

Eszter Mózes
Universität Debrecen
Graduiertenkolleg Sprachwissenschaft
Pf. 47
H-4010 Debrecen
dtotheszter@gmail.com