*Eszter Mózes:*
*Tony McEnery & Andrew Hardie: Corpus Linguistics. Method, Theory and Practice*
*Argumentum 8 (2012), 92-96*
*Debreceni Egyetemi Kiadó*

# *Recenzió*

Eszter Mózes

## Tony McEnery & Andrew Hardie: Corpus Linguistics. Method, Theory and Practice*

Cambridge: Cambridge University Press, 2012, 294 Seiten

Das vorliegende Buch wurde im Rahmen der Serie *Cambridge Textbooks in Linguistics* im Jahr 2012 veröffentlicht. Die Autoren haben sich zum Ziel gesetzt, einen Überblick darüber zu geben, wie sich die Korpuslinguistik methodologisch entwickelt hat und bislang angewandt worden ist, welche theoretischen Fragen heutzutage bei der Anwendung auftauchen und welche Probleme gelöst werden müssen, wenn innerhalb der Linguistik oder in anderen Disziplinen mit Korpora gearbeitet wird. Die Autoren weisen darauf hin, dass schon zahlreiche einführende Fachbücher über Korpuslinguistik vorhanden sind. Aus diesem Grund wollen sie die Grundkenntnisse – wo das nicht unbedingt nötig ist – nicht ausführlich beschreiben; zu diesem Zweck geben sie verschiedene Fachbücher an, d.h. sie setzen bestimmte linguistische Vorkenntnisse voraus. Es sollen vielmehr die im Hintergrund liegenden Probleme, Prozesse und die neuesten Richtungen der Korpuslinguistik unter die Lupe genommen und diskutiert werden. Durch dieses Ziel unterscheidet sich dieses Buch von den anderen Fachbüchern wie zum Beispiel Biber, Conrad & Reppen (1998) oder McEnery, Xiao & Tono (2006).

Das Buch stellt die Korpuslinguistik auf 294 Seiten und in 9 Kapiteln vor. Im Weiteren möchte ich den Gedankengang der einzelnen Kapitel in Grundzügen darstellen.

Kapitel 1 beschreibt ausführlich, was ein Korpus ist. Es wird darauf hingewiesen, dass sich Korpora wesentlich voneinander unterscheiden können. Als eine ganz allgemeine Definition wird vorgeschlagen, dass ein Korpus eine Sammlung von Texten ist, wobei unter Texten maschinenlesbare Texte zu verstehen sind. Außer dieser gemeinsamen Charakteristik werden verschiedene Merkmale angegeben, nach denen sich Korpora unterscheiden können. Je nach der Art der Kommunikation werden verschiedene Korpora unterschieden wie zum Beispiel Sammlungen von mündlichen oder schriftlichen Texten. Die Autoren sind der Meinung, dass diese Unterscheidung eine wichtige Rolle spielt, denn die gesprochene und die geschriebene Sprache verfügen über unterschiedliche Eigenschaften. Der nächste Punkt, der behandelt wird, ist, ob ein Korpus als Quelle zur Aufstellung linguistischer Theorien oder als Quelle zum Nachweis existierender oder hypothetischer Theorien dienen sollte. Die Autoren lehnen jene Richtung der Korpuslinguistik ab, welche den theoretischen Status dieser linguistischen

* Die vorliegende Publikation entstand mit Unterstützung des Projekts TÁMOP 4.2.2/B-10/1-2010-0024. Das Projekt wurde im Rahmen des Entwicklungsplans Neues Ungarn verwirklicht und teilweise durch den Europäischen Sozialfonds (ESF) sowie den Europäischen Fonds für regionale Entwicklung (EFRE) finanziert.

*Eszter Mózes:*
*Tony McEnery & Andrew Hardie: Corpus Linguistics. Method, Theory and Practice*
*Argumentum 8 (2012), 92-96*
*Debreceni Egyetemi Kiadó*

Disziplin anerkennt und sie nicht nur als Methode betrachtet. Je nachdem, ob die Texte des Korpus von Zeit zu Zeit erweitert werden, oder nur einmal zusammengestellt werden und dann unverändert bleiben, unterscheiden wir zwischen Monitor- und Referenzkorpus. Ob Korpora zusätzliche Informationen über die Texte enthalten sollten, ist wieder ein Punkt, wo unterschiedliche Meinungen vertreten werden. Die Autoren sind der Meinung, dass im Falle der Korpora linguistische Informationen immer hinzugegeben, aber sie werden nicht immer dokumentiert. Als letztes Merkmal wird die Zahl der Sprachen erwähnt. Abhängig von der Zahl der Sprachen, aus denen die Texte gesammelt wurden, können wir von einsprachigen und mehrsprachigen Korpora sprechen. Viele dieser Punkte dienen als Kernfragen zu weiteren Diskussionen, die in den folgenden Kapiteln erläutert werden.

Im zweiten Kapitel werden die Methoden, mit denen Korpuslinguistik arbeitet, behandelt. Arbeiten mit Sammlungen von Texten wurden und werden immer noch verschiedenerweise beurteilt. Als ein Pol dieser Meinungen werden Chomsky und sein Standpunkt erwähnt, der schon seit Mitte des zwanzigsten Jahrhunderts sehr stark vertreten ist. Chomsky hat die Untersuchung der Korpora wegen seiner Beschränktheit völlig abgelehnt. Aber seit den achtziger Jahren steigt die Zahl jener Sprachwissenschaftler ständig, die daran fest glauben, dass Korpora von großem Nutzen sein können. Es wurde demonstriert, dass Korpora auch für verschiedene andere Teildisziplinen wie zum Beispiel kontrastive Linguistik, Konversationsanalyse, Semantik usw. nützlich sein können. Und gerade wegen dieser Vielfalt von Resultaten vertreten die Autoren die Meinung, dass die Eistellung von Chomsky falsch ist: mithilfe der Korpuslinguistik können wir die Sprache besser verstehen. Allerdings warnen uns die Autoren zugleich, dass wir die aus einem Korpus entnommenen Daten nicht mit der Sprache selbst verwechseln oder identifizieren dürfen.

Korpora können zahlreiche Fragen über die Sprache beantworten, dazu werden aber verschiedene Mittel gebraucht. Ein annotiertes Korpus kann verschiedene Informationen wie zum Beispiel Metadaten, Markup und Annotation enthalten. Diese sind kodierte Informationen, die mit verschiedenen Programmen, vor allem Konkordanzprogrammen entnommen werden können. Heutzutage wird schon die vierte Generation der Konkordanzprogramme benutzt. Das Problem liegt aber daran, dass sie sich inhaltlich wenig verändert haben, es wurden hauptsächlich nur technische Entwicklungen erreicht. Nicht nur Konkordanzprogramme, sondern auch die Statistik kann bei der Analyse der Daten helfen.

Kapitel 3 bietet einen Überblick über Themen, die nach der Meinung der Autoren in der Fachliteratur stark vernachlässigt werden. Da das Internet als eine immer wichtigere Quelle der Korpora dient, spielen die rechtlichen und ethischen Fragen im Zusammenhang mit Internetbenutzung eine zunehmend größere Rolle. Eine wichtige Frage ist, wie das Urheberrecht behandelt werden sollte. Die im Internet vorhandenen Texte können schnell heruntergeladen werden, aber es stellt sich die Frage, ob ihr Herunterladen und ihre Weitergabe als Teil der Korpora zulässig sind. Viele Webseiten wie zum Beispiel Wikipedia haben keine Beschränkung, deswegen können sie frei genutzt und sogar verändert werden, aber andere Seiten können nur mit einer Erlaubnis benutzt werden.

Was die ethischen Fragen betrifft, erläutern die Autoren, dass es Richtlinien gibt, die entweder in die Praxis umgesetzt oder außer Acht gelassen werden. Die Forscher tragen eine große Verantwortung. Wenn Korpora gesprochener Sprache gesammelt werden, verraten die Teilnehmer wichtige Details über ihr Privatleben, die aber nicht in den Korpora erscheinen sollten, denn das könnte ihre Rechte verletzen.

*Eszter Mózes:*
*Tony McEnery & Andrew Hardie: Corpus Linguistics. Method, Theory and Practice*
*Argumentum 8 (2012), 92-96*
*Debreceni Egyetemi Kiadó*

Nach der Meinung der Autoren sollten Korpora für viele erreichbar sein, denn so könnten sie eines ihrer Ziele erfüllen, nämlich die Benutzung in weiten Kreisen. Korpora können aber nicht nur für wissenschaftliche, sondern auch für andere Ziele gebraucht werden, wie zum Beispiel Mission, wenn religiöse Texte gesammelt, und dann in Form eines Korpus verbreitet werden.

In den vorigen Kapiteln wurden die praktischen Teile der Korpuslinguistik vorgestellt. Das vierte und die darauf folgenden Kapitel beschreiben verschiedene Traditionen über die Stellung der Korpuslinguistik in der Sprachwissenschaft.

In Kapitel 4 geht es um die englische Korpuslinguistik, die als Grundlage für die späteren Traditionen gedient hat. Das englische Korpus war ein Pionier in dem Sinne, dass verschiedene Mittel und grundlegende Begriffe wie zum Beispiel Kollokation und Annotation dieser Richtung zu verdanken sind. Hier gab es die ersten Korpora, die den Interessierten freien Zugang erlaubt haben. Die ersten Korpora haben sich mit dem britischen Englisch beschäftigt, erst nachher wurden auch Korpora aus dem amerikanischen Englisch erstellt. In der zweiten Hälfte des Kapitels werden die Zentren beschrieben, die zu dieser Entwicklung beigetragen haben: Erwähnt wird unter anderem das University College London, wo das bedeutendste Korpus der verschiedenen Varianten des Englischen entstanden ist, und dem auch eine neue umfassende Grammatik, Quirk, Greenbaum, Leech & Svartvik (1972) zu verdanken ist. Die Forscher der Lancaster University haben viele Ergebnisse auf dem Gebiet der Annotation und des Tagging erreicht. Die University of Birmingham wird ferner als das bedeutendste Zentrum vorgestellt. Deren Forscher vertreten den korpusgetrieben Ansatz (die so genannte Lexikogrammatik), nach dem das Lexikon und die Grammatik eng zusammenhängen. Die Erforschung der Kollokationen ist auch eng mit dieser Forschungsstätte verknüpft. Außerdem werden noch die Tätigkeiten an der Université Catholique de Louvain, der University of Nottingham sowie der Northern Arizona University zusammengefasst.

Kapitel 5 beschäftigt sich mit der Frage, wie sich die Sprache verändert, entwickelt und wie diese Vorgänge mithilfe der Korpora untersucht werden können. Dazu werden zwei Ansätze gewählt: der eine ist der mehrdimensionale Ansatz, wo die Grundeinheit der Text ist; der andere ist die so genannte variationistische Soziolinguistik, wo die Grundeinheit der Sprecher selbst ist. Die Zahl der historischen Korpora steigt an. Nicht nur die Veränderung der geschriebenen Sprache, sondern auch die der gesprochenen Sprache kann in ihnen untersucht werden. Als eine berühmte Sammlung der geschriebenen Sprache erwähnen die Autoren die Brown-Familie, die aus einer Reihe von Korpora besteht, die alle nach den gleichen und strikten Auswahlgrundlagen zusammengestellt wurden. Mithilfe dieser Korpora können die Variation und der Wandel im heutigen Englischen erforscht und beschrieben werden. Anhand der Brown-Korpora wurden verschiedene grammatische Phänomene untersucht wie zum Beispiel Amerikanisierung, Grammatikalisierung, Kollokationisierung und Demokratisierung. All diese Phänomene beeinflussen den diachronen Wandel der Sprache. Sie können verschiede Motivationen haben. Als ein möglicher Ansatz für die Untersuchung dieser Änderungen beschreiben die Autoren die multidimensionale Methode von Biber. Er wollte die Sprache mit verschiedenen statistischen Methoden untersuchen. Sein Ziel war, die Merkmale verschiedener Sprachverwendungen nach neuen Gesichtspunkten zu gruppieren. Er hat verschiedene Dimensionen der Merkmale aufgestellt, nach denen die Texte variieren können. Zahlreiche andere Disziplinen könnten diese Methode benutzen, wie zum Beispiel die Spracherwerbsforschung, aber die Methode ist nicht weit verbreitet. Eine Kritik bezieht sich auf die Anwendbarkeit auf andere Textsorten, die Wiederholbarkeit und die untersuchten Merkmale.

*Eszter Mózes:*
*Tony McEnery & Andrew Hardie: Corpus Linguistics. Method, Theory and Practice*
*Argumentum 8 (2012), 92-96*
*Debreceni Egyetemi Kiadó*

Im Kapitel 6 beschreiben die Autoren die Gruppe der Neo-Firthianer, deren bekanntester Vertreter Sinclair ist. Was für diese Gruppenach der Meinung der Autoren besonders charakteristisch ist, ist die Behandlung von Kollokationen und Konversation. Die Neo-Firthianer verstehen unter Kollokation, dass ein Teil der Bedeutung der Wörter nur im Zusammenhang mit anderen Wörtern auftritt. Es gibt verschiedene Methoden, um diese zusammen auftretenden Wörter zu finden. Sinclair ermittelt Kollokationen mit statistischer Signifikanz. Eine andere Methode wäre die Suche nach Kollokationen über Konkordanzen, aber da sich diese Methode meistens auf die Intuition der Verarbeiter stützt, ist sie nicht konsequent und völlig zuverlässig. Die meisten Neo-Firthianer ziehen die letztere Methode vor. Ihrer Meinung nach kann ein Wort nicht nur mit einem anderen Wort, sondern auch mit grammatischen Kategorien oder Markern zusammen erscheinen, was sie Kolligation nennen. Wie die Autoren beschreiben, spielt die Bedeutung der Wörter bei Sinclair eine zentrale Rolle. Kollokationen dienen auch als Grundlage beim Studieren der Lexikon und Semantik. Er geht so weit, dass er Grammatik durch Wörter definiert. Die Autoren des Buches teilen die extreme Position der Neo-Firthianer nicht. Sinclairs Motto ist „vertraut dem Text", wobei das Korpus die zentrale Rolle spielt und alle anderen Umstände bei den Untersuchungen fast völlig ausgeschlossen werden. Außerdem lehnt Sinclair Tagging ab, weil er meint, dass dadurch wichtige Informationen verloren gehen könnten, die Einheit des Textes gebrochen wird und außer dem Text stammende Informationen hinzugefügt werden können.

Im Kapitel 7 erläutern die Autoren kurz, was der Unterschied zwischen Funktionalisten und Formalisten ist. Im Gegensatz zu Formalisten betonen die Funktionalisten die Wichtigkeit der Sprachbenutzung. Das wird von den Autoren als eine Gemeinsamkeit mit der Korpuslinguistik erwähnt, und deswegen können diese zwei Gebiete der Linguistik sich berühren. Zum Funktionalismus rechnen die Autoren die Kognitive Linguistik und die Typologie. Diese Gebiete weisen gemeinsame Charakteristiken wie zum Beispiel Methoden, Ziele und Ergebnisse auf. Kognitive Sprachwissenschaftler und andere Funktionalisten sind der Meinung, dass Syntax und Semantik nicht zwei unabhängige Module sind, sondern eng zusammenhängen. Statt Regeln gibt es Konstruktionen, die lexikalisch bedingt sind. Die *konzeptuelle Metapherntheorie* stützt sich auf die Intuition der Sprachwissenschaftler. Aber wenn Korpuslinguistik zur Hilfe gerufen wird, kann mit hoher Wahrscheinlichkeit behauptet werden, ob etwas eine Metapher ist, oder nicht, da die hohe Kookkurrenz mit bestimmten Wörtern darauf hinweisen kann, dass etwas übertragen gemeint ist.

Im Kapitel 8 fahren die Autoren mit der Beschreibung der Beziehung der Korpuslinguistik zu anderen Zweigen der Linguistik fort. Funktional kognitive Sprachwissenschaftler und Psycholinguisten vertreten die Meinung, dass Sprache und nicht-sprachliche Kognition in engem Zusammenhang stehen. Sprachverarbeitung kann durch verschiedene Experimente untersucht werden, unter anderem durch Blickerfassungsexperimente. Mit diesen wird ermittelt, wie die Wörter und Sätze vom Gehirn bearbeitet werden. Wenn eine Kombination von Wörtern wahrscheinlicher als eine andere ist, dann braucht das Gehirn weniger Zeit für die Verarbeitung. Korpuslinguistik kann in diesem Zusammenhang auf zweierlei Weise benutzt werden: als Kontrolle der Natürlichkeit der gebrauchten Sprache und als Quelle der Frequenzdaten. Um die Kindersprache zu studieren, wurde das CHILDES-Korpus aus sprachlichen Äußerungen von Kindern zusammengestellt. Ein weiteres Gebiet, wo Korpuslinguistik ebenfalls benutzt werden kann, ist die konnektionistische Forschung, wo aus Korpora stammenden Daten als Teil eines Trainings funktionieren und dessen Ziel ist, den Verlauf des Spracherwerbs zu reproduzieren. Formelhafte Sprache ist ein weiteres Forschungsgebiet, das immer bedeutender

wird. Formelhafte Sprache und Kollokationen können gleichgesetzt werden. Psycholinguisten untersuchen sie, weil diese Phänomene mit der Sprachverarbeitung und dem Spracherwerb eng zusammenhängen. Als letztes zeigt dieses Kapitel, dass die Neo-Firthianer, die Korpuslinguisten und die Funktionalisten Ähnlichkeiten in den Resultaten aufweisen; sie untersuchen die Sprache aus verschiedenen Aspekten, aber sie kommen zu ganz ähnlichen Folgerungen. Die Ähnlichkeit zeigt sich in ihrer Auffassung über Lexikon und Grammatik. Sie sind der Meinung, dass die Sprache durch domänenübergreifende kognitive Prozesse erklärt werden kann.

Im letzten Kapitel fassen die Autoren zuerst die Geschichte der Korpuslinguistik zusammen. Anschließend stellt sich die Frage, was die Zukunft der Korpuslinguistik sein wird. Nach der Meinung der Autoren ist eine mögliche Richtung die, dass sie nicht mehr selbstständig behandelt wird, sondern von allen Sprachwissenschaftlern als Methode aufgefasst wird. Die Autoren denken aber nicht, dass die Entwicklung in diese Richtung weisen wird. Sie erwarten vielmehr eine Spaltung innerhalb des Begriffs Korpuslinguistik: der eine benutzt sie als Methode, der andere arbeitet daran, wie ihre Benutzung noch erweitert werden könnte. Die Autoren bedauern, dass die Beziehung zwischen Korpuslinguistik und Computerlinguistik schwächer wird, und hoffen, dass sie in der Zukunft wieder enger zusammenarbeiten werden. Aber die Zusammenarbeit mit anderen Disziplinen wird stärker: Korpora werden immer öfter auf geisteswissenschaftlichen Gebieten verwendet.

Am Ende eines jeden Kapitels findet der Leser einerseits eine Sammlung von praktischen Übungen, die ermöglichen, die verschiedenen vorgestellten Theorien auch in die Praxis umzusetzen und anderseits Fragen zur Diskussion, die den Leser veranlassen sollen, über die Themen nachzudenken. Die Autoren geben auch öffentlich erreichbare Korpora an, mit deren Hilfe der Gebrauch der Korpora geübt werden kann. Außerdem enthält das Buch Lektürehinweise, die der Orientierung beim vertiefenden Weiterlesen dienen. In einem Register werden die wichtigsten Fachbegriffe und Namen aufgelistet, was die Orientierung im Buch erleichtert.

Zusammenfassend kann festgestellt werden, dass das Buch den im Vorwort genannten Zielen gerecht wird. Es erfordert zwar linguistische Vorkenntnisse, es ist aber logisch aufgebaut und leicht zu verstehen. Mit zusätzlichen Erklärungen und mit der Besprechung der Aufgaben im Buch würde ich es für Studierenden der Sprachwissenschaft empfehlen.

## Literatur

Biber, D., Conrad, S. & Reppen, R. (1998): *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

McEnery, T., Xiao, R.Z. & Tono, Y. (2006): *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1972): *A Grammar of Contemporary English*. London: Longman.

Eszter Mózes
Debrecener Reformierte Theologische Universität
Lehrstuhl für Fremdsprachen
H-4026 Debrecen
Péterfia Str. 1-7
esther.mozes@gmail.com