Stefaniya Ptashnyk, Erla Hallsteinsdóttir, Noah Bubenhofer, Hrsg. 2010. Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie/Corpora, Web and Datab ...

Pon, Leonard

Leonard Pon
Filozofski fakultet
Sveučilišta Josipa Jurja Strossmayera
Osijek

Stefaniya Ptashnyk, Erla Hallsteinsdóttir, Noah Bubenhofer, Hrsg. 2010. *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie/Corpora, Web and Databases. Computer-Based Methods in Modern Phraseology and Lexicography*. (Bd. 25. Phraseologie und Parömiologie). Nürnberg: Schneider Verlag Hohengehren. 267 Seiten. € 26. ISBN: 978-3-8340-0733-9.

Dieses im April 2010 in Deutschland veröffentlichte Buch ist ein neuer Band aus der Reihe *Phraseologie und Parömiologie*. Die herkömmliche Vorgehensweise in phraseologischen und lexikographischen Untersuchungen und die moderne Forschung, die grundsätzlich von digitalen Textkorpora und von den Recherchemöglichkeiten des Webs ausgeht, werden aufeinander bezogen. In einem längeren Vorwort wird unterstrichen, welche Ziele dieser Band hat. Es folgt ein dreiteiliger Hauptabschnitt, der acht Beiträge auf Deutsch und acht Beiträge auf Englisch enthält. Am Ende findet sich auch ein benutzerfreundliches, vierseitiges Stichwortverzeichnis.

Im Vorwort des Bandes gehen Noah Bubenhofer und Stefaniya Ptashnyk von den US-Präsidentschaftswahlen 2008 aus und heben Unterschiede im Sprachgebrauch Barack Obamas und John McCains hervor. Im Zusammenhang damit wird deutlich gemacht, dass die Analyse der Kollokationen äußerst hilfreich ist, wenn der für bestimmte Bereiche, Themen oder Textsorten typische Sprachgebrauch zu beschreiben ist. Großer Wert wird dabei auf verschiedene Aspekte der computergestützten linguistischen Arbeit im Bereich der Phraseologie und Lexikographie gelegt. Es wird auf die Methoden hingewiesen, die die Arbeit mit großen Textmengen nicht nur möglich, sondern auch einfach(er) machen. Das Web bietet sich als Quelle der elektronischen Texte aller Textsorten an. Der Forscher kann von den vorhandenen Suchmaschinen Gebrauch machen, um elektronische Texte nach gewünschten Kriterien zu durchsuchen, oder er kann aufgrund der frei verfügbaren Web-Daten ein eigenes Korpus zusammenstellen. Heutzutage liegt also ein so schneller und leichter Zugang zu den empirischen Daten vor, wie wir ihn früher nie gekannt haben, und das wird bei der Untersuchung verschiedener Arten der phraseologischen Einheiten auch genutzt. Die moderne Forschung wirft ein neues Licht auf die Definition des Begriffs Kollo-

kation. So wird hier vorgeschlagen, dass beim Definieren der Kollokation auch die statistischen Signifikanzmaßen berücksichtigt werden.

Der Band gliedert sich wie gesagt in drei Teile. Im Folgenden wenden wir uns den Beiträgen aus einzelnen Teilen zu.

Der erste Teil trägt den Namen *Theoretische Erkenntnisse der korpus- und webbasierten Phraseologie-Untersuchungen* und enthält sechs Beiträge. Im ersten Beitrag (*The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology*) stellt uns Jean-Pierre Colson seine Gedanken über eine allgemeine Theorie der Phraseologie vor. Er unterstreicht die Unzulänglichkeit der semantischen und kognitiven Kriterien, die für die bisherigen Untersuchungen der festen Wendungen kennzeichnend sind, und setzt sich dafür ein, dass beim Definieren von festen Wendungen auch statistische Daten berücksichtigt werden. Diese Daten ließen sich durch die Analyse von webbasierten Korpora gewinnen. Die Stärke des statistischen Kriteriums besteht darin, dass es wiederholbare Experimente ermöglicht, was für eine wissenschaftliche Beschäftigung mit Phraseologismen von fundamentaler Bedeutung ist. Außerdem können webbasierte Korpora das Fremdsprachenlernen fördern. Den Fremdsprachenlernern wären die Informationen über das neue, in Wörterbüchern noch nicht verzeichnete phraseologische Gut einer Sprache in einem webbasierten Korpus jederzeit zugänglich.

Uwe Quasthoff, Fabian Schmidt und Erla Hallsteindóttir (*Häufigkeit und Struktur von Phraseologismen am Beispiel verschiedener Web-Korpora*) untersuchen die Frequenz von Wörtern innerhalb der Mehrworteinheiten und die Frequenz von Mehrworteinheiten in unterschiedlichen Textsorten, wobei sie das Zipfsche Gesetz testen. In dieser Untersuchung zeigte sich beispielsweise, dass Phraseologismen mit veraltenden oder selten verwendeten Wörtern auch selten(er) gebraucht werden, und umgekehrt. Die Autoren dieses Aufsatzes gehen auch der Frage nach, ob und unter welchen Umständen neue Phraseologismen automatisch auffindbar wären.

Hrisztalina Hrisztova-Gotthardt stellt in ihrem Beitrag (*Methoden und Ergebnisse einer korpusbasierten Untersuchung zur Vorkommenshäufigkeit bulgarischer Sprichwörter in zeitgenössischen Zeitungstexten*) fest, dass zurzeit keinerlei Daten über die aktuelle parömiologische Situation im Bulgarischen vorliegen. Die Autorin möchte Ermittlungen vornehmen, die einen Einblick in die tatsächliche passive Kenntnis sowie in die aktive Verwendung bulgarischer Sprichwörter ermöglichen. Dafür eignet sich das Teiltextpräsentation-Verfahren und die Korpusanalyse. Die erste Methode ermittelt die bekanntesten Sprichwörter einer Sprachgemeinschaft und erstellt ein Sprichwort-Minimum, während die

zweite die Informationen über die aktive Verwendung und die Vorkommenshäufigkeit der Sprichwörter in einer Sprache liefert.

Jussi Niemi, Juha Mulli, Marja Nenonen, Sinikka Niemi, Alexandre Nikolaev und Esa Penttilä (*Body-Part Idioms across Languages: Lexical Analyses of VP Body-Part Idioms in English, German, Swedish, Russian and Finnish*) analysieren Idiome mit Körperteilbezeichnungen in fünf Sprachen. Bisherige Untersuchungen zu diesem Thema seien methodologisch nicht genügend herausgearbeitet. Folglich versucht diese Autorengruppe, strukturell ähnliche Idiome in diesen fünf Sprachen miteinander zu vergleichen. In Frage kommen nur Idiome mit der Struktur „Verb + Nomen". Es wurde u. a. festgestellt, dass als Nomen vor allem Konkreta erscheinen und dass die Idiome vor allem diejenigen Nomina enthalten, die auch im nicht-idiomatischen Gebrauch hochfrequent sind.

Nach einer kurzen, aber gut durchdachten Auseinandersetzung mit unterschiedlichen Auffassungen des Begriffs Kollokation widmet sich Christine Konecny (*Lexikalische Kollokationen und der Beitrag der Internet-Suchmaschine ‚Google' zu ihrer Erschließung und Beschreibung*) der Frage, ob und wie mittels der Suchmaschine Google eine Frequenzanalyse lexikalischer Kollokationen durchgeführt werden kann. Analysiert werden fünf ausgewählte italienische Kollokationen. Es wird von einem engen Verständnis des Phänomens Kollokation ausgegangen, bei dem unter Kollokation eine binäre Konstruktion aus Basis und Kollokator verstanden wird. Übrigens wird Kollokation nicht als ein starres Konstrukt, sondern eher als ein dynamisches Phänomen angesehen. Wie die Autorin in ihren Ausführungen zeigt, lässt sich mithilfe der Suchmaschine Google der Kollokationsstatus einer Wortgruppe nicht bestätigen, aber die Suche kann helfen, wenn wir die zentralen und peripheren Basen für den gegebenen Kollokator herausfinden wollen.

Ausgehend vom kroatischen Idiom *loviti u mutnom*, stellt Jelena Parizoska in ihrem Aufsatz (*The Canonical Form in Murky Waters: Idiom Variation and the Croatian National Corpus*) die Frage nach der kanonischen Form eines Idioms bzw. nach der Variabilität des gegebenen Idioms. Sie versucht zu beweisen, dass das angeführte Idiom ein schematisches Idiom ist, dessen konzeptueller Kern das idealisierte kognitive Modell MUTNA VODA ist, und zieht dabei eine Reihe der Belege aus dem kroatischen Nationalkorpus (HNK) heran. Alle Variationen dieses (wie auch jedes anderen) Idioms seien vorhersagbar, weil sie unabdingbar mit dem ihnen zugrunde liegenden konzeptuellen Kern übereinstimmten. Anders ausgedrückt: Es herrscht ein enger Zusammenhang zwischen konzeptueller Motivation der Idiome und ihrer lexikalischen und syntaktischen Veränderlichkeit.

Der zweite Teil des Bandes (*Methodische Probleme und Tools in der computergestützten Phraseologie-Forschung*) setzt sich aus vier Aufsätzen zusammen. Im ersten Beitrag (*Semi-Automatic Retrieval of Phraseological Units in a Corpus of Modern Norwegian*) stellen Ruth Vatvedt Fjeld, Lars Nygaard und Eckhard Bick das Tool namens DeepDict Lexifier dar. Dieses Tool ermöglicht automatisches Auffinden von Kollokationen, indem es reale syntaktische Relationen wie Subjekt-Verb oder Verb-Objekt und nicht die typischen N-Grame analysiert. Da für das Norwegische gegenwärtig kein Wörterbuch für Kollokationen vorliegt, ist das erwähnte Tool als ein geeignetes Mittel anzusehen, die genannte *lexikographische Lücke* zu beseitigen.

Peter Ďurčo (*Einsatz von Sketch Engine im Korpus – Vorteile und Mängel*) zeigt, wie das statistische Tool Sketch Engine beim Kompilieren von Kollokationsprofilen für das erste slowakische Kollokationswörterbuch angewendet werden kann. Die Stärke der statistischen Methoden liegt in der Möglichkeit der Vorstrukturierung von Massendaten sowie im Erkennen von Usus-Phänomenen. Als Nachteil erweist sich dagegen die Tatsache, dass individuelle Eingriffe in den vorprogrammierten Prozess der Datenerhebung in verschiedenen Stadien nur eingeschränkt oder überhaupt nicht möglich sind.

Oksana Petrova (*Computer-Mediated Discourse vs. Traditional Text Corpora as a Data Source for Idiom Variation Research in Finnish*) thematisiert zwei mögliche Herangehensweisen bei der Untersuchung von phraseologischer Variation. Im ersten Fall werden die benötigten Daten der finnischen Sprachdatenbank entnommen, im zweiten Fall verwendet man als Datenquelle Usenet Newsgroups. Während die traditionellen Textkorpora Tools mit *linguistisch interessanten* Suchoptionen enthalten, die beispielsweise bei der Suchmaschine Google nicht vorhanden sind, ist für Google groups und Usenet die Menge der Daten kennzeichnend. Da phraseologische Einheiten eine relativ niedrige Frequenz aufweisen, müssen bei phraseologischen Untersuchungen große Textmengen herangezogen werden. Dementsprechend stehe vor dem Forscher nun die Aufgabe, die Suchoptionen der Sprachdatenbanken auch im Bereich des Webs anwendbar zu machen.

Melita Aleksa Varga (*Methoden und Tools zur Erstellung eines korpusbasierten Kollokationswörterbuchs – am Beispiel des Kroatischen*) beschreibt die Vorarbeiten, die bei der Erstellung eines korpusbasierten kroatischen Kollokationswörterbuchs zu erledigen wären. Für das Fremdsprachenlernen ist ein kroatisches Kollokationswörterbuch unentbehrlich, aber als Schwierigkeit bei dessen Zusammenstellung erweist sich der Umstand, dass derzeit ein großer Mangel an korpuslinguistischen Daten zum tatsächlichen Sprachgebrauch im Kroatischen herrscht. Für die Erstellung des Kollokationswörterbuchs wählt die Autorin die

Methode der Recherche in einem Korpus und die automatische Extrahierung der Kollokationen. Als geeignet hierfür erscheint das Programm „Ngram Statistics Package“. Der genannte Vorschlag setzt übrigens auch eine weitere manuelle Bearbeitung der Ergebnisse bei der Auswahl repräsentativer Kollokationen voraus.

Im dritten Teil des Bandes begegnen wir sechs Beiträgen, die unter dem gemeinsamen Titel *Korpora und Web in der phraseographischen Praxis* erscheinen. Marcel Dräger und Britta Juska-Bacher (*Online-Datenerhebungen im Dienste der Phraseographie*) setzen sich mit den Methoden der digitalen Datenerhebung in der einsprachig ausgerichteten Phraseographie auseinander. Dargestellt wird die Methode der Online-Befragung und das interaktive Online-Lexikon. Bei der Online-Befragung wird ein auf dem Server abgelegter Fragebogen von Probanden online ausgefüllt. Da die Probanden aus unterschiedlichen Regionen stammen, können auf diese Weise auch areale Aspekte der Phraseologismen untersucht werden. So können die Lücken der traditionell genutzten Quellen bei der Behandlung von Phraseologismen beseitigt werden.[1] Im anschließenden Teil wird am Beispiel des interaktiven Online-Lexikons gezeigt, wie der Benutzer nicht nur als Informationskonsument, sondern auch als Datenlieferant auftreten kann.

Der Aufsatz von Maria Toporowska Gronostaj und Emma Sköldberg (*Swedish Medical Collocations: A Lexicographic Approach*) ist vor dem Hintergrund eines Projekts entstanden, dessen Ziel die Zusammenstellung eines elektronischen Lexikons mit medizinischer Terminologie ist. Dieser Aufsatz verschafft uns einen Einblick in die Analyse der medizinischen Kollokationen mit einem Substantiv und einem Verb. Es wird auf den Unterschied zwischen zwei Typen der Kollokationen hingewiesen, was sich auch im Format des betreffenden Eintrags widerspiegelt. Zum ersten Typ gehören die Kollokationen, für die eine feste lexikalische Bindung zwischen Verb und Substantiv charakteristisch ist. Den zweiten Typ machen die Kollokationen mit bedeutungsverwandten Konstituenten aus. Da hier eine schwache Bindung zwischen Verb und Substantiv herrscht, sind im Unterschied zum ersten Typ mehrere Varianten ein und derselben Kollokation möglich.

[1] Welche Vorteile auch immer dieses Verfahren mit sich bringt, wir müssen auch Folgendes betonen: Obwohl die Online-Datenerhebung bei der Analyse ausgewählter Phraseologismen wichtige Einsichten in die aktuelle parömiologische Situation gewähren kann, ist es immer noch fraglich, inwieweit diese Methode bei der Herstellung eines (phraseologischen) Wörterbuchs anwendbar ist. Zum Problem der Rekrutierung von Probanden äußern sich auch die Autoren dieses Beitrags.

Franziska Wallner (*Kollokationen in Wissenschaftssprachen: Zur lernerlexikographischen Relevanz der Textarten- und Diskursspezifik von Kollokationen*) widmet ihren Beitrag den Kollokationen in Wissenssprachen, denn allem Anschein nach können Kollokationen wissenschaftsspezifische Züge tragen. Anhand eines Korpus der deutschen Wissenschaftssprache und eines pressesprachlichen Korpus konnte Wallner feststellen, dass bisweilen hochsignifikante Unterschiede im Gebrauch der Kollokationen zwischen Wissenssprachen und anderen Kontexten vorliegen. Solche Informationen zum Gebrauch der Kollokationen, die übrigens in den herkömmlichen Wörterbüchern fehlen, sind nach Wallner insbesondere für die Lernerlexikographie höchst relevant.

Ausgehend von drei Zeilen aus Shakespeares Hamlet, analysieren Sixta Quassdorf und Annelies Häcki Buhofer (*„... you are quoting Shakespeare': Quotations in Practice*) den sprachlichen Einfluss Shakespeares auf einzelne sprachliche Bereiche und Perioden. Berücksichtigt wird die Form, Funktion und Bedeutung, die diese Zeilen in unterschiedlichen Texten besitzen. Der vorliegende Aufsatz versteht sich als Beitrag zu einem umfassenderen Projekt, in dem das Phänomen Shakespeare in unterschiedlichen Kontexten erforscht wird.

Natalia Filatkina, Ane Kleine und Birgit Ulrike Münch (*Verbale und visuelle Formelhaftigkeit: Zwischen Tradition und Innovation*) interessieren sich dafür, wie die korpus- und computergestützten Verfahren bei einer interdisziplinären Erforschung der historischen verbalen und visuellen Formelhaftigkeit eingesetzt werden können. Der Begriff *formelhafte Wendungen* umfasst dabei neben linguistischen auch kunsthistorische Phänomene. Ausgehend vom gegenwartssprachlichen Idiom *nach jmds. Pfeife tanzen*, wird das Vorhaben der Forschungsgruppe gezeigt. Hier sei angemerkt, dass diese Forscher im Rahmen eines Projekts arbeiten, dessen Hauptanliegen die Entwicklung einer multimedialen und internetbasierten Forschungsressource ist.

Frank Richter, Manfred Sailer und Beata Trawiński (*The Collection of Distributionally Idiosyncratic Items: An Interface between Data and Theory*) präsentieren in ihrem Aufsatz eine elektronische Sammlung verschiedener Untergruppen lexikalischer Elemente, deren besonderes Merkmal idiosynkratische Distribution ist. Diese Elemente verhalten sich also distributionell nicht so, wie aufgrund ihrer syntaktischen Kategorienzugehörigkeit zu erwarten wäre. Zudem beurteilen die Muttersprachler bestimmte Strukturen, in denen solche Elemente fälschlicherweise fehlen, als nicht grammatisch. Die erwähnte Sammlung versteht sich als eine Plattform, die zur linguistischen Theorienbildung viel beitragen kann. Die Autoren sind überzeugt, dass unter Berücksichtigung sowohl konstruktionsgrammatischer als auch kollokationsorientierter Aspekte implizites linguistisches Wissen des Muttersprachlers modelliert werden kann.

Im vorgestellten Buch werden Korpora, Web und Datenbanken nicht nur aus linguistisch-theoretischer, sondern auch aus sprachdidaktischer Sicht behandelt. Wie mit den neuen Datenquellen umzugehen ist, wird an verschiedenen europäischen Sprachen gezeigt – in den sechzehn Beiträgen werden folgende Sprachen genauer untersucht: Bulgarisch, Englisch, Finnisch, Deutsch, Italienisch, Kroatisch, Norwegisch, Russisch, Schwedisch, Slowakisch. Der potentielle Leser des Buchs wird demnach nicht nur mit den theoretischen Grundlagen der aktuellen Strömungen auf dem Gebiet der Phraseologie vertraut gemacht, sondern ihm wird auch gezeigt, inwiefern er linguistisch profitiert, wenn er sich bei der Untersuchung einer Einzelsprache an die besprochenen Prinzipien hält.