Korpora, Datenbanken und das Web: State of the Art computergestützter Forschung in der Phraseologie und Lexikographie

Noah Bubenhofer/Stefaniya Ptashnyk

Barack Obama gewann die US-Präsidentschaftswahlen 2008, weil er in seinen Reden an die Wählerschaft die richtigen Leitvokabeln mit den passenden Kollokationen benutzte. In Verbindung mit *Iraq* nannte er häufig Lexeme und Wortkombinationen wie *critical issue*, *national issue*, *focus on al Qaeda*, *our troops*, *family*, *war*, *talk*, *Afghanistan*, *Pakistan* und andere mehr (vgl. Abbildung 1). Sein Herausforderer John McCain hingegen verwendete Kollokatoren wie *we … win*, *suceed*, *come home*, *they* und *I* (vgl. Abbildung 2). Es zeigt sich also, dass McCain andere *Irak*-Kollokationen benutzt und damit zwar ebenfalls viel über dieses Thema spricht, unterschiedlich zwischen Obama und McCain ist jedoch der Sprach*gebrauch*, die *Redeweise* (Bubenhofer 2009). Auffällig ist nämlich nicht nur die Art der Kollokatoren zu *Iraq*, sondern auch die Tatsache, dass in McCains Rhetorik überhaupt eine geringere Vielfalt an Kollokatoren als bei Obama festzustellen ist. Obama scheint also eine differenziertere (aber damit auch kompliziertere) Ausdrucksweise bevorzugt zu haben (vgl. Bubenhofer u. a. 2008a,b).

Die Behauptung, die richtigen Kollokationen führen zur US-Präsidentschaft, mag etwas überspitzt sein, aber Analysen zu typischen Sprachgebrauchsmustern in Wahlkämpfen zeigen die Relevanz von Kollokationen, um den Sprachgebrauch zu charakterisieren.[1] Um den Sprachgebrauch zu bestimmen, der *typisch* für bestimmte Diskurse, Themen, Menschen, Textsorten, Zeitpunkte etc. ist, interessieren in erster Linie Kollokationen, die im Vergleich mit Referenzkorpora statistisch signifikant für die jeweiligen Bereiche sind.

Nicht nur Mehrworteinheiten im weitesten Sinne, sondern auch enger definierte Phänomene wie etwa Idiome oder Sprichwörter sind zentral, um die Eckpunkte von Sprachgebrauch zu bestimmen. Im US-Wahlkampf gelangte z. B. die Wendung *Joe the Plumber* zu Berühmtheit, die eben nicht für den wörtlich zu verstehenden *Hans, der Klempner* stand, sondern für den exemplarischen Kleinunternehmer der unteren Mittelschicht und für ein ganzes Wirtschaftsprogramm von McCain. Es gibt zwar einen Joe, der Klempner und die Ursache für die Wendung ist; der Ausdruck hatte sich jedoch im Verlauf des Wahlkampfes verfestigt und etabliert und ist dann als Idiom zu einem frequenten Element der US-Wahlkampfrhetorik geworden.

---

1 So auch Analysen zu den deutschen Bundestagswahlen 2009: Bubenhofer u. a. (2009).

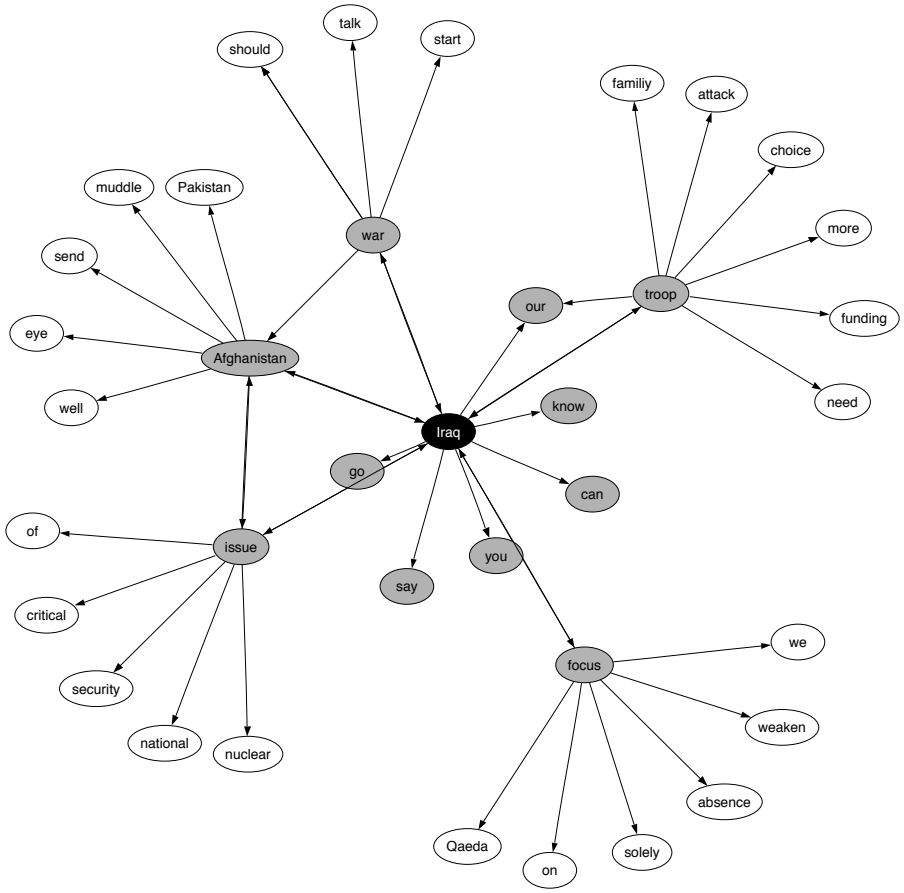

Abbildung 1: Kollokatoren zu *Iraq*, die Barack Obama typischerweise während der ersten TV-Debatte verwendete (vgl. Bubenhofer u. a. 2008a; Grafik: Forschergruppe semtracks).

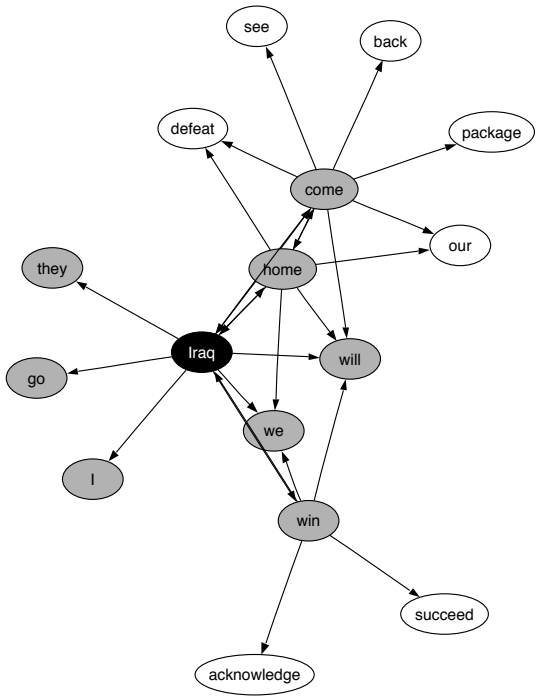

Abbildung 2: Kollokatoren zu *Iraq*, die John McCain typischerweise während der ersten TV-Debatte verwendete (vgl. Bubenhofer u. a. 2008a); Grafik: Forschergruppe semtracks.

Das kurze Analysebeispiel zeigt verschiedene Aspekte der computergestützten linguistischen Arbeit im Bereich von Phraseologie und Lexikographie auf:

- Die Arbeit mit großen Textmengen ist im Hinblick auf die einsetzbaren Methoden einfacher geworden. Die Korpuslinguistik stellt eine Reihe von Standardtools zur Verfügung, um Sprachgebrauchs- und Kollokationsprofile oder typische Mehrworteinheiten zu berechnen, Daten zu annotieren und durchsuchbar zu machen.

- Die Arbeit mit großen Textmengen ist aber auch im Hinblick auf die Datenquellen einfacher geworden. Das Web ist eine Fundgrube elektronisch verfügbarer Texte unterschiedlichster Textsorten, seien es Transkriptionen politischer Reden, Diskussionsforen, Blogs, Zeitungen und Magazine etc. Zwar gibt es teilweise urheberrechtliche Probleme zu lösen, zumindest technisch ist die automatische Beschaffung von Textdaten vergleichsweise einfach.

- Neben der vielfältigen Nutzung korpuslinguistischer Daten für phraseologische Studien, die das Sprachsystem im Fokus haben, hilft der phraseologische Blick auch in den Bereichen der Text-, Diskurs-, Kultur- oder Rhetorikanalyse. Die zuletzt genannten linguistischen Methoden nehmen Sprachgebrauchsmuster in den Blick, die typisch für einen bestimmten Teilbereich im Vergleich zum allgemeinen Sprachgebrauch sind. Gemeinsam ist jedoch allen Untersuchungen die Überzeugung, dass Mehrworteinheiten (wie eng oder weit sie auch immer definiert sein mögen) relevante Untersuchungskategorien sind.

- Die Untersuchungsergebnisse können schließlich in ganz unterschiedlichen Formen weiterverwendet werden, etwa für rein wissenschaftliche (didaktische, lexikographische oder sprachbeschreibende) Zwecke. Auch für ein politisch interessiertes Publikum können sie brauchbar gemacht werden, wie dies im Fall der US-Wahlanalysen erfolgte: Die Daten wurden so aufbereitet, dass sie die sprachlichen Besonderheiten des Wahlkampfs demonstrieren.[2]

Die in diesem Band versammelten Beiträge präsentieren Lösungen und Ansätze in allen diesen Bereichen. Die Gemeinsamkeit liegt darin, dass sie den neuesten Entwicklungen in der Phraseologie- und Lexikographie-Landschaft Rechnung tragen, in der die computergestützen und textkorpusbasierten Recherche- und Analysemethoden immer mehr an Bedeutung gewinnen.

Die Anwendung der computergestützen Methoden in der Phraseologie ist als eine logische Fortsetzung der bisherigen Dynamik in diesem linguistischen Teilbereich zu sehen. Nach traditionellen Untersuchungen, die auf die sprachsystematische Beschreibung der Phraseologismen, insbesondere ihrer Semantik, sowie auf ihre Typologisierung abzielten, folgten – als Folge der pragmatischen Wende – primär textstilistische und pragmatisch begründete Fragestellungen. Das Voranschreiten der Korpus- und Computerlinguistik der

2 Die Analysen erschienen nicht nur im wissenschaftlichen Kontext, sondern auch als Blog (*www.semtracks.com*) und waren Gegenstand verschiedener Medienberichte.

letzten zwei Jahrzehnte eröffnete der Phraseologie-Forschung neue Wege: Im Mittelpunkt zahlreicher Forschungsvorhaben steht heute die Nutzung elektronischer Ressourcen im Allgemeinen und großer Textkorpora im Besonderen.

Dabei hat man längst erkannt, dass neben den wissenschaftlich aufbereiteten und durchdacht strukturierten Textkorpora auch Web-Ressourcen als umfangreiche Textgrundlage für phraseologische Untersuchungen genutzt werden können, z. B. Nachrichtenportale, Websites, Blogs etc.

Die Beiträge zeigen deutlich, welche Probleme mithilfe der elektronischen Ressourcen überhaupt angegangen und auf welche Art und Weise Textkorpora sowie das WWW gewinnbringend für wissenschaftliche Zwecke allgemein und für phraseologische Fragestellungen im Besonderen genutzt werden können.

Im Einzelnen geht es um die Vor- und Nachteile der automatischen/formalisierten Suche nach festen Wortverbindungen bzw. der Identifizierung von Phraseologismen in umfangreichen Textkorpora, um die Anwendung der Corpus-driven-Analysen bei der Beschreibung typischer Sprachgebrauchsmuster (etwa in bestimmten Textsorten oder Diskursen), um die Nutzbarkeit statistischer Methoden für qualitative Untersuchungen sowie um die anwendbare Software für die Analyse großer Korpora.

Im Zuge des intensiven Einsatzes der Textkorpora hat der Begriff „Kollokation" in der Phraseologieforschung erneut an Bedeutung gewonnen. Der korpusbasierten Kollokationsanalyse als einem wichtigen Instrumentarium der Phraseologieforschung und der praktischen Lexikographie wird besondere Aufmerksamkeit geschenkt. Angesprochen werden dabei verschiedene Auffassungen von Kollokationen (etwa als Kontinuum zwischen Idiomen und freien Wortverbindungen, als hierarchisch organisierte Konstruktionen oder als nach dem Kriterium der Frequenz definierte Wortkombinationen) sowie die terminologische Problematik der Kollokationsforschung. Darüber hinaus wird die lexikographische, sprachdidaktische, textsortenbeschreibende und diskursanalytische Relevanz von Kollokationen aufgezeigt.

Schließlich werden aus der theoretischen Perspektive Aspekte und Fragestellungen angerissen, die das allgemeine Verständnis phraseologischer Phänomene betreffen: Führen etwa korpusbasierte Untersuchungen zum grundlegenden Umdenken traditioneller phraseologischer Grundbegriffe und -konzepte, etwa der Auffassung phraseologischer Festigkeit und Variabilität oder der phraseologischen Norm/Nennform etc.? Welche Berührungspunkte lassen sich zwischen syntaktischen Phänomenen, etwa den syntagmatischen Mustern und der Phraseologie feststellen und inwiefern lassen sich syntaktische Fragestellungen im Rahmen der Phraseologie behandeln?

1 Korpora als Datenressource und Analysegegenstand

Die moderne Korpuslinguistik, die umfangreiche Korpora in digitaler Form recherchierbar macht, ist auch für die Phraseologie zu einer unverzichtbaren Methode geworden: Korpora können einerseits als umfangreiche Zettelkästen verwendet werden, in denen Belege für bestimmte Phänomene und deren Distribution gesucht und untersucht werden. Anderer-

seits bietet die statistische Datenanalyse die Möglichkeit, musterhaften Sprachgebrauch in großen Korpora zu entdecken und anschließend zu kategorisieren.

Korpora werden mehrheitlich für die Identifikation von Phraseologismen sowie für die Untersuchung ihrer Distribution, Variation und Verwendungsweisen eingesetzt. Die im vorliegenden Band versammelten Arbeiten lassen sich zum größten Teil unter diese Aufgaben subsumieren, wobei für ihre Lösung eine Vielfalt von Methoden eingesetzt werden kann.

1.1 Identifikation von Mehrworteinheiten

Ist die maschinelle Identifikation von Phraseologismen nach wie vor eine besonders „harte Nuss" oder „pain in the neck" für die Computer- und Korpuslinguistik (Filatkina/Kleine/Münch[3])? Die Arbeiten im vorliegenden Band verwenden eine Reihe unterschiedlicher Methoden, um das Problem anzugehen, wobei die Erwartungen an die Methoden sehr unterschiedlich sind. Einerseits bestimmt das Untersuchungsinteresse, welche Art von Mehrworteinheiten überhaupt gesucht werden: lexikalische Kollokationen, Idiome, Sprichwörter oder auch nicht-idiomatische Mehrwortverbindungen? Andererseits wird mehr oder weniger Handarbeit in Kauf genommen, um die gewünschten Einheiten zu finden.

Im Hintergrund steht die Uneinigkeit darüber, wie nützlich die theoretischen Konzepte der Phraseologie für quantitative Analysen überhaupt sind. Grundsätzlich besteht bei vielen Forschern die Tendenz zur Ausweitung des Phraseologie-Begriffes sowie das Bedürfnis, für die Begriffsbestimmung von Mehrworteinheiten neue, nicht traditionelle Kriterien zu wählen. Colson argumentiert beispielsweise für eine statistisch operationalisierbare Definition von Kollokationen, um objektive und reproduzierbare Ergebnisse zu erhalten. Er kritisiert, dass Kriterien wie Idiomatizität oder Kompositionalität semantisch oder kognitiv begründet und deshalb schwer objektivierbar seien. Kollokationen müssten deshalb als Phänomene definiert werden, die mit statistischen Signifikanzmaßen erfasst werden können.

Auf der anderen Seite stehen konkrete Ziele, wie das Erstellen von phraseologischen Wörterbüchern (Aleksa; Toporowska Gronostaj/Sköldberg; Ďurčo etc.) oder Untersuchungen zur Verbreitung von Idiomen und Sprichwörtern, die eine didaktische Anwendung finden könnten (Hrisztova-Gotthardt; Petrova etc.). In solchen Fällen werden Korpora auf der Grundlage klarer Theorien nach den gewünschten Phänomenen durchsucht. Dabei besteht die Erwartung, mit maschinellen Methoden möglichst exakt die Phänomene zu finden, die die phraseologische Theoriebildung traditionellerweise als ‚Kollokation', ‚Phrasem', ‚Idiom' oder ‚Sprichwort' definiert.

Eine Mittelstellung nehmen Methoden ein, die pure Signifikanzmaße mit Wissen über syntaktische Strukturen oder Semantik kombinieren. Tools wie der ‚DeepDict Lexifier' (Fjeld/Nygaard/Bick) oder die ‚Sketch Engine' (Ďurčo) verlassen sich auf solche Verfahren.

---

3 In Kapitälchen ausgeschriebene Autorennamen verweisen auf Beiträge in diesem Band.

In den Beiträgen des vorliegenden Bandes werden sowohl einfache technische Tools als auch elaborierte Methoden, die teilweise Programmierkenntnisse erfordern, angewandt. Im Detail sind es die folgenden allgemein verfügbaren Tools oder eigens entwickelte Methoden:

1. Bestehende Software:
   a) Programme, die mit Signifikanzmaßen Kollokationen oder beliebig lange Mehrworteinheiten berechnen:
      – Kwic Concordance for Windows: *http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html* (vgl. Aleksa)
      – NSP Ngram Statistics Package: *http://ngram.sourceforge.net/* (vgl. Aleksa)
      – Collocation Extract: *http://pioneer.chula.ac.th/~awirote/colloc/* (vgl. Aleksa)
      – Manatee/Bonito: *http://nlp.fi.muni.cz/projekty/bonito/* (vgl. Ďurčo)
   b) Sketch Engine: *http://www.sketchengine.co.uk/* (vgl. Ďurčo)
   c) DeepDict Lexifier: *http://gramtrans.com/deepdict/* (vgl. Fjeld/Nygaard/Bick)
2. Eigene Methoden:
   a) Auffinden von Adjektiv-Nomen-Kollokationen über das Web unter der Nutzung der Programmierschnittstellen (APIs) von Web-Suchmaschinen. Das Verfahren folgt dem Prinzip der Suche nach Modifikatoren wie *most*, *rather*, *quite*, *too* im Kontext eines gegebenen Nomens; anschließend erfolgt die maschinelle Bearbeitung der Suchresultate, etwa Extraktion aller Adjektive, Ermittlung der Frequenzen etc. (vgl. Colson).
   b) Systematische Überprüfung aller n-Gramme in einem gegebenen Text über die Programmierschnittstellen (APIs) von Web-Suchmaschinen. Hierbei wird ihre Fixiertheit, d. h. das Verhältnis der exakten Form zu einer variablen Form getestet. Bei Trigrammen erfolgt dies nach folgender Formel: Frequenz [Wort 1 + Wort 2 + Wort 3] im Verhältnis zu [Wort 1 + beliebiges Wort + Wort 3] (vgl. Colson).
   c) Untersuchung von Gebrauchsfrequenzen möglicher Kollokatoren zu einem Ausgangslexem über Google-Abfragen. Diese Methode funktioniert nach dem Prinzip der Frequenzanalysen für verschiedene Kollokationen; anschließend werden die Frequenzen der Kollokationen untereinander verglichen, um besonders häufige Kollokatoren zum Ausgangslexem zu finden (vgl. Konecny).
   d) Kombination verschiedener Kriterien zwecks automatischen Auffindens von Phraseologismen in den Korpora (vgl. Quasthoff/Schmidt/Hallsteinsdóttir). Hierbei werden mehrere statistisch operationalisierbare Eigenschaften von Phraseologismen ausgenutzt wie geringe Variabilität, typische Wortartenkombinationen oder das Vorhandensein von mindestens zwei „wichtigen Wörtern“ (typischerweise Autosemantika) etc.

1.2 Überprüfung der Verwendungsweisen und der Variationsspielräume von Phraselogismen

Während die Identifikation von Phraseologismen besonders komplex ist, sind Untersuchungen, die von bestimmten (vordefinierten) Phraseologismen oder ihren Konstituenten ausgehen, deutlich einfacher. Es ist unbestritten, dass umfangreiche Textkorpora eine gute Basis bilden, um die Distribution und die Verwendungsweisen von Phraseologismen, aber auch ihre usuellen Varianten und okkasionellen Modifikationen zu überprüfen (vgl. Ptashnyk 2009). Eine Reihe von Beiträgen des vorliegenden Bandes untersuchen das Vorkommen von Phraseologismen in bestimmten Textsorten, Sprachen oder Themenbereichen. So interessiert sich etwa Wallner für die unterschiedlichen Verwendungen von Kollokationen in wissenschaftlichen und allgemeinsprachlichen Korpora. Mit Signifikanztests wird dabei überprüft, ob die Frequenzunterschiede von Kollokatoren in den beiden Korpora überzufällig sind. Hrisztova-Gotthardt benutzt eine Sammlung von bulgarischen Sprichwörtern als Ausgangsbasis, um deren Verbreitung in Zeitungstexten zu analysieren. Von vordefinierten Idiomen gehen auch Niemi et al. aus: Die Autoren untersuchen Verwendungsunterschiede von Körper-Idiomen in verschiedenen Sprachen.

Während die Suche nach usuellen Phraseolgismen über ihre Konstituenten in Verbindung mit Platzhaltern sich relativ einfach gestaltet, ist das Auffinden von modifizierten Einheiten deutlich schwieriger. Ein solches Vorhaben ist das Projekt HyperHamlet (vgl. Quassdorf/Häcki Buhofer), in dem eine spezielle Klasse von festen Wendungen, nämlich Zitate aus Shakespeares „Hamlet“ und ihre Modifikationen, untersucht werden.

Im Zusammenhang mit der korpusgestützten Überprüfung der Verwendungsweisen von Phraseologismen wird kritisch die Frage diskutiert, wie stark die Korrelation zwischen der Vorkommenhhäufigkeit dieser Einheiten in den Korpora und ihrer tatsächlichen Geläufigkeit bei Sprechern einer Sprache ist. Dräger/Juska-Bacher betonen, dass die Verwendungshäufigkeit nicht mit der Bekanntheit übereinstimmen muss, und fordern deshalb, Korpusstudien durch Online-Befragungen zu ergänzen. Andere Arbeiten deuten jedoch darauf hin, dass es sich bei fehlender Korrelation um ein Problem des Korpus, insbesondere einer zu kleinen Datengrundlage, handeln könnte (vgl. etwa Quasthoff/Schmidt/Hallsteinsdóttir): Teilweise sind Phraseologismen ein Phänomen gesprochener Sprache, weshalb nur bei genügend großen Korpora, die möglichst viele Textsorten enthalten, zu erwarten ist, dass geläufige Phraseologismen auch in den Korpusdaten frequent sind. Die Arbeit von Petrova zeigt dabei, dass die Nutzung von Newsgroups als Korpus eine interessante Möglichkeit ist, die Textsortenvielfalt zu erhöhen und die gesprochene Sprache stärker zu berücksichtigen.

Aus der Perspektive der Korpuslinguistik stellen bei der Recherche die Variabilitätsspielräume von Phraseologismen ein Problem dar, seien es unterschiedliche Formen von usuellen (z. B. lexikalischen, stilistischen oder orthographischen) Varianten oder okkasionelle phraseologische Modifikationen. Auch die Flexion der Phraseologismen, die bei der syntagmatischen Einbettung der Mehrworteinheiten meist unabdingbar ist, erschwert in vielen Fällen die Recherche. Dieses Problem ist aus technischer Sicht verhältnismäßig einfach zu lösen, etwa durch die von der Computerlinguistik entwickelten Wortarten-Tagger

und Lemmatisierungsverfahren (vgl. z. B. den TreeTagger, Schmid 1994), die für eine Reihe von Sprachen bereits trainiert sind. Wenn darüber hinaus ein manuell annotiertes Korpus und Lemmalisten zur Verfügung stehen, können diese Tagger auf neue Sprachen trainiert werden. Trotz dieser Tatsache scheinen diese Tools für einige Forscherinnen und Forscher im Bereich der Phraseologie noch nicht zu den verwendeten Standardtools zu gehören. Diese Tatsache offenbart das Desiderat nach einfach bedienbarer computer- und korpuslinguistischer Software, die auch von programmiertechnisch unerfahrenen Forschenden bedient werden kann.

Die orthographischen Variantionsspielräume sind ebenfalls vergleichsweise einfach in den Griff zu kriegen: Ist der Spielraum bekannt, kann dies bei der Suchabfrage berücksichtigt werden, indem z. B. sog. „reguläre Ausdrücke" verwendet werden, d.h. eine komplexe Sprache, die durch Platzhalter und durch Formulierung von Bedingungen variantentolerante Suchen zulässt.[4] Im Beitrag von Filatkina/Kleine/Münch werden darüber hinaus Algorithmen erwähnt, die die Ähnlichkeit zwischen Phraseologismen berechnen und so auch Phraseologismen finden, die minimale Varianz in der Schreibung aufweisen.

Komplexer ist aber die Aufgabe, lexikalisch modifizierte Phraseologismen zu finden, wie Hrisztova-Gotthardt, Parizoska und Petrova zeigen. Die Lösung könnte darin liegen, die Suche auf Schlüssellexeme des Phraseologismus zu beschränken und die syntaktische Struktur mit einzubeziehen, wie das Quasthoff/Schmidt/Hallsteinsdóttir skizzieren. Auch könnten semantische Datenbanken und Ontologien wie WordNet, GermaNet etc. verwendet werden, um automatisiert Synonyme, Hypero- und Hyponyme in die Suche nach modifizierten Phraseologismen zu integrieren. Auch Ressourcen wie die Sammlung von Phraseologismen mit unikalen Komponenten (vgl. Richter/Sailer/Trawiński) können für verschiedene Zwecke der maschinellen Analyse hilfreich sein.

Bestehende Korpora bieten kaum variantentolerante Recherchen, obwohl diese sehr nützlich wären (vgl. etwa die Untersuchung von Petrova anhand des Kielpankki-Korpus). Deshalb sind Kooperationen mit Computer- und Korpuslinguisten notwendig, die dafür geeignete Rechercheinstrumente programmieren können und darüber hinaus reiche Erfahrung in der maschinellen Textanalyse haben.

2 Aufbau und Nutzung von Korpora verschiedener Typen

2.1 Web und Korpus, Web als Korpus

Seit geraumer Zeit werden in der linguistischen Forschung Korpora genutzt, die nach bestimmten Kriterien aufgebaut sind und entsprechend „ausgewogene" Datenmengen darstellen. Seit einiger Zeit wird das Web als zunehmend wichtige Ressource gesehen, und in diesem Zusammenhang wird auch die Rolle des Webs als Korpus diskutiert. Dabei gibt es zwei grundsätzlich unterschiedliche Nutzungsweisen: Zum einen werden die bereits vorhandenen Suchmaschinen verwendet, um verfügbare elektronische Texte nach bestimm-

4 Vgl. für eine kurze Einführung in die Verwendung von regulären Ausdrücken Bubenhofer (2006), „Anhang" → „RegExp".

ten Phänomenen zu durchsuchen; hierbei wird das gesamte Web als ein riesiges Korpus interpretiert. Zum anderen besteht die Möglichkeit, auf der Basis der frei verfügbaren Web-Daten die eigenen, strenger am jeweiligen Forschungsinteresse orientierten Korpora zusammenzustellen.

A. Nutzung von Suchmaschinen

Das Web kann über die vorhandenen Suchmaschinen wie ‚Google' etc. nach bestimmten sprachlichen Phänomenen durchsucht werden. Auf diese Weise kommt man schnell und ohne weitere Investitionen an ein sehr großes Korpus heran. Wollte man ein Korpus der Größe, wie Suchmaschinenbetreiber es indizieren, selbst zusammenstellen und recherchierbar machen, wäre das mit sehr hohen Kosten verbunden und ist deshalb für Forschungszwecke kaum realistisch.

Nachteile der Nutzung bestehender Suchmaschinen liegen einerseits darin, dass sie nicht für linguistische Recherchen gedacht sind und der Suchalgorithmus nicht im Detail bekannt ist. Andererseits kann die Datengrundlage schlecht kontrolliert werden. Die Suchmaschinenbetreiber schweigen sich über den Umfang des indizierten Korpus aus. Es ist deshalb nicht möglich, Frequenzen in Relation zur Korpusgröße zu setzen. Zudem verändert sich die Zusammensetzung des Korpus naturgemäß laufend.

Beispiele für solche Verfahren liefert im vorliegenden Band Konecny, die den Einsatz der Suchmaschine ‚Google' für phraseologische Studien schildert. Colson umgeht die Beschränkungen von Suchmaschinen dadurch, dass er die Schnittstellen (APIs) der Suchmaschinen für maschinelle Abfragen nutzt, um die Daten im Anschluss mit eigenen Methoden zu verarbeiten.

B. Aufbau eigener Web-basierter Korpora

Die im Web publizierten und frei verfügbaren Texte können als Grundlage für den Aufbau eines eigenen Korpus verwendet werden. In solchen Fällen sind die Forscher nicht an die Möglichkeiten einer Suchmaschine gebunden, sondern sie können die Daten nach Belieben aufbereiten und haben deshalb die vollständige Kontrolle über die Zusammensetzung des Korpus. Quasthoff/Schmidt/Hallsteinsdóttir zeigen detailliert ihre Methode der Web-Nutzung, mit der sie im Rahmen des Wortschatz-Projekts der Universität Leipzig eigene Korpora aufbauen.

Die automatische Beschaffung der Dokumente ist vergleichsweise einfach, wie zahlreiche Arbeiten zeigen (vgl. Baroni/Bernardini 2006, Fletcher 2007, Kilgarriff/Grefenstette 2003, Sharoff 2006). Primär gibt es urheberrechtliche Bedenken, wobei Ansätze eines „Open-Source"-Korpus eine Lösung versprechen, bei dem die Daten für die Analyse zwar gespeichert, jedoch nur in Form von URL-Listen weitergegeben werden (Sharoff 2006).

2.2 Datenbanken zur Recherche und Verwaltung sprachlicher Daten

Mit der zunehmenden Professionalisierung der Korpusanalysen werden immer komplexere Systeme zur Verwaltung von Analyseresultaten eingesetzt. In der Phraseologie und Lexiko-

graphie ist die Korpusanalyse oft nur der erste Schritt zu einer bestimmten Untersuchung; im Anschluss folgt die Sichtung und die Kategorisierung von Belegen oder von Zwischenresultaten maschineller Analysen, beispielsweise der Phraseologismen-Kandidaten. So verwundert es nicht, dass diese Analyseresultate in Datenbanken abgelegt und dort mit weiteren Informationen versehen werden. Die Datenbanken bieten dabei den Vorteil, dass die hier gesammelten und repräsentierten Daten problemlos für unterschiedliche Endprodukte verwendet werden können. Aufgrund einer Online-Datenbank, die während des Forschungsprozesses in der vollen Komplexität benutzt wird, kann z. B. ein klassisches gedrucktes Wörterbuch oder ein für die Laiennutzer aufbereitetes Portal produziert werden. Auf diese Mehrfachnutzung von Datenbanken verweisen Dräger/Juska-Bacher.

Während es in der Pionierphase der Datenbanken reichte, auf dem Arbeitscomputer lokal installierte Systeme zu verwenden, verlangt das Zeitalter des Web netzwerkfähige Datenbanken, die von mehreren Benutzern gleichzeitig über Webschnittstellen benutzt werden können. Solche Datenbanken ermöglichen eine neue Form von Kooperation für beliebig viele Forschende unabhängig von ihren Arbeitsorten. Selbst der Einbezug von (informierten) Laien nach Vorbild von kollaborativen Web-Anwendungen wie der Wikipedia wird möglich, wie Dräger/Juska-Bacher skizzieren: Es handelt sich dabei um eine Art ‚Crowdsourcing', das Auslagern von Aufgaben an eine große Masse von Freiwilligen, die über das Web zu einem gemeinsamen Projekt beitragen. In der Summe machen diese Beiträge die Projekte erst möglich. Die Wikipedia ist ein erfolgreiches Beispiel für Crowdsourcing. Dies lässt sich auf die Linguistik übertragen: Dräger/Juska-Bacher beschreiben in ihrem Beitrag das Online-Phraseologismenwörterbuch, das dank Kommentaren und Beiträgen der Nutzer an Qualität gewinnt. Darüber hinaus kann das Nutzungsverhalten analysiert werden: Welche Einträge werden besonders häufig nachgeschlagen? Bei welchen existiert ein großer Diskussionsbedarf? Darüber lassen sich Hinweise über die Bekanntheit, Gebräuchlichkeit oder Strittigkeit von Phraseologismen gewinnen. Ein weiteres Beispiel für eine solche offene Datenbank ist „HyperHamlet" (vgl. Quassdorf/Häcki Buhofer).

Eine besondere Herausforderung stellt die Verwaltung historischer Daten dar. Von elaborierten Datenbanksystemen für historische Phraseologie berichten Filatkina/Kleine/Münch: Belege für formelhafte Sprache aus dem Mittelalter und der Frühen Neuzeit werden im Projekt „Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)" in einer webfähigen Datenbank (MySQL mit entsprechenden Webschnittstellen) gesammelt und kategorisiert. Dabei erweisen sich auch neue Möglichkeiten der Informatik, die als Details erscheinen mögen, als große Hilfen: So ist es erst mit der Definition des Unicode-Standards für die Zeichencodierung (UTF-8) möglich geworden, historische (aber auch nicht-westeuropäische) Schriftsysteme mit nicht speziell dafür eingerichteten Systemen zu verarbeiten. Filatkina/Kleine/Münch machen deutlich, dass Datenbanken auch multimedial eingesetzt werden können, wie das Projekt „Gnomisches Wissen im Raum der Bilder", wo Text- mit Bilddaten kombiniert werden, zeigt.

Eine zunehmend wichtige Rolle nimmt die Auszeichnungssprache XML ein, mit deren Hilfe zu beliebigen Daten Metainformationen oder Annotationen hinzugefügt werden können. Richter/Sailer/Trawiński verwenden für die Datenbanken der unikalen Wörter und positiven und negativen Polaritätselemente XML als Auszeichnungssprache. Mit ei-

ner Datenbank wie ‚eXist' (*exist.sourceforge.net*) lassen sich beliebige XML-Dokumente darüber hinaus einfach verwalten. Damit werden die Vorteile von XML mit den Vorteilen einer Datenbank kombiniert: So ist es möglich, die Metadaten zu einem Text oder Textausschnitt im Datenbanksystem zu verwalten, gleichzeitig aber am Text mit XML Annotationen vorzunehmen, die wiederum automatisiert ausgewertet werden können.

Neben der Verwaltung von Belegen bieten Datenbanken einen weiteren Vorteil: Die Menge der Datensätze kann leicht nach unterschiedlichen Kriterien ausgewertet werden. Auf Knopfdruck können statistische Angaben zu den Daten (entsprechende Kategorisierungen vorausgesetzt) zusammengetragen werden. Besonders interessant sind dabei jedoch Verfahren, die Ähnlichkeiten von unterschiedlichen Datensätzen algorithmisch entdecken. Hilfreich ist dies beispielsweise bei Belegdatenbanken von historischer formelhafter Sprache, wo orthographische und syntaktische Varianten wegen fehlender Standardisierung häufiger und computerlinguistische Verfahren der Lemmatisierung schwieriger sind. Wie die Varianten einer Mehrworteinheit in variationsreichen historischen Texten automatisch ermittelt werden, dies zeigen beispielsweise Filatkina/Kleine/Münch auf.

3 Ausblick/Fazit

Die Beiträge, die in diesen Sammelband eingegangen sind, demonstrieren teils sehr erschöpfend und teils exemplarisch die breite Palette der Einsatzmöglichkeiten korpuslinguistischer und computergestützter Verfahren, welche für theoretische und angewandte Lexikologie- und Phraseologie-Untersuchungen sowie für die praktische Lexikographie und Sprachdidaktik eingesetzt werden. Neue schnelle Zugänge zu den empirischen Daten bilden den wichtigsten Nutzen sowohl für den Forscher, als auch für den Nutzer von Forschungsergebnissen, seien das Laien oder Experten.

Im Zuge dieser Entwicklung zeigt sich zunehmend und mit prägnanter Deutlichkeit die eigentlich schon längst in der traditionellen Phraseologie angesprochene und nicht gelöste Frage nach den Grenzen des Phraseologischen. Wenn der Computer genutzt werden soll, um Phraseologismen maschinell zu entdecken, muss das Phänomen ‚Phraseologismus' so operationalisiert werden, dass es auf der sprachlichen Oberfläche durch klare Regeln gefunden werden kann. Die verschiedenen Algorithmen, die entwickelt wurden, erfassen manchmal mehr, manchmal aber auch weniger als das, was klassischerweise als Phraseologismus bezeichnet wird. Es gibt gute Gründe, das Phänomen breiter zu fassen und durch statistische Verfahren eine breite Palette von musterhaftem Sprachgebrauch abzudecken. Zahlreiche Beiträge haben bereits gezeigt, dass etwa für lexikographische, didaktische und textlinguistische Fragestellungen solche Phänomene des musterhaften Sprachgebrauchs von großer Bedeutung sind, aber eben nicht mehr immer im Rahmen der traditionellen Phraseologie anzusiedeln sind.

Welches Ergebnis vermag die Öffnung der primär phraseologisch fokussierten Forschungsinteressen gegenüber neueren Methoden zu erzielen, die allenfalls die Grenzen der Phraseologie hinterfragt? Ob diese Entwicklung dazu führt, dass sich die Phraseologie einen engeren, deutlich abgesteckten Untersuchungsgegenstand reserviert, oder dass an der

Schnittstelle traditioneller Phraseologie und Syntax sich ein neues linguistisches Teilgebiet behauptet, wird die künftige Forschung zeigen.

Literaturverzeichnis

Baroni, Marco/Bernardini, Silvia (Hgg.) (2006): *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.

Bubenhofer, Noah (2006): *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge*. Elektronische Ressource (*http://www.bubenhofer.com/korpuslinguistik/*).

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen; 4).

Bubenhofer, Noah/Klimke, Martin/Scharloth, Joachim (2008a): *political tracker – U.S. Presidential Campaign '08: A Semantic Matrix Analysis*. Elektronische Ressource (*http://semtracks.com/politicaltracker/*).

Bubenhofer, Noah/Klimke, Martin/Scharloth, Joachim (2008b): The Word War: „Yes, He Did". How Obama won the (rhetorical) battle for the White House. In: *International Relations and Security Network, ISN ETH Zurich* (*http://www.isn.ethz.ch/Current-Affairs/Special-Reports/The-Word-War-Yes-He-Did/Analysis*).

Bubenhofer, Noah/Klimke, Martin/Scharloth, Joachim (2009): *political tracker – Bundestagswahl '09. Eine Semantische Matrixanalyse*. Elektronische Ressource (*http://semtracks.com/politicaltracker/*).

Fletcher, William H. (2007): Implementing a BNC-Compare-able Web Corpus. In: *Building and Exploring Web Corpora – Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval (WAC3-2007, September 2007), UCL*, hg. v. C. Fairon, H Naets, A. Kilgarriff u. G-M de Schrijver, Louvain: Presses Universitaires de Louvain.

Kilgarriff, Adam/Grefenstette, Gregory (2003): Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics* 29, H. 3, S. 333–347.

Ptashnyk, Stefaniya (2009): *Phraseologische Modifikationen und ihre Funktionen im Text. Eine Studie am Beispiel der deutschsprachigen Presse*. Baltmannsweiler: Schneider.

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees* (*http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf*).

Sharoff, Serge (2006): Open-source Corpora: Using the Net to Fish for Linguistic Data. In: *International Journal of Corpus Linguistics* 11, S. 435–462.

Noah Bubenhofer
Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim
Deutschland
bubenhofer@ids-mannheim.de

Stefaniya Ptashnyk
Heidelberger Akademie der Wissenschaften
Deutsches Rechtswörterbuch
Karlstr. 4
69117 Heidelberg
Deutschland
stefaniya.ptashnyk@adw.uni-heidelberg.de