Phraseologie & Parömiologie

Korpora, Web und Datenbanken
Computergestützte Methoden in der modernen Phraseologie und Lexikographie

Corpora, Web and Databases
Computer-Based Methods in Modern Phraseology and Lexicography

Stefaniya Ptashnyk, Erla Hallsteinsdóttir, Noah Bubenhofer

Schneider Verlag Hohengehren

Inhalt

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).

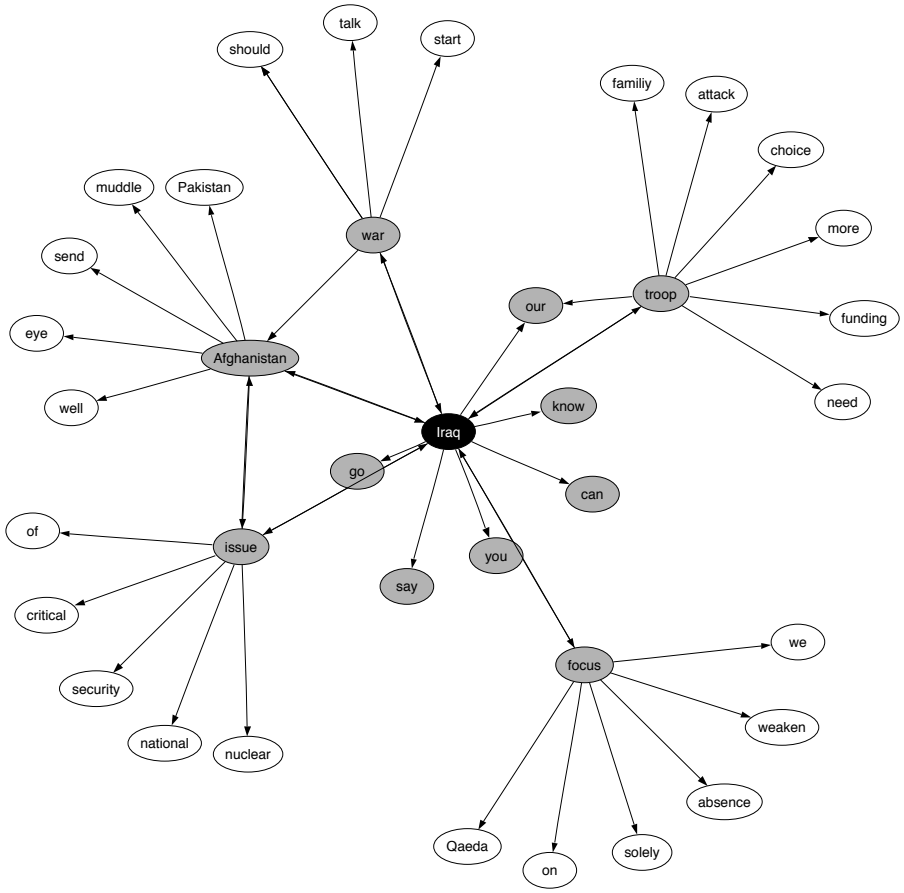

Abbildung 1: Kollokatoren zu *Iraq*, die Barack Obama typischerweise während der ersten TV-Debatte verwendete (vgl. Bubenhofer u. a. 2008a; Grafik: Forschergruppe semtracks).

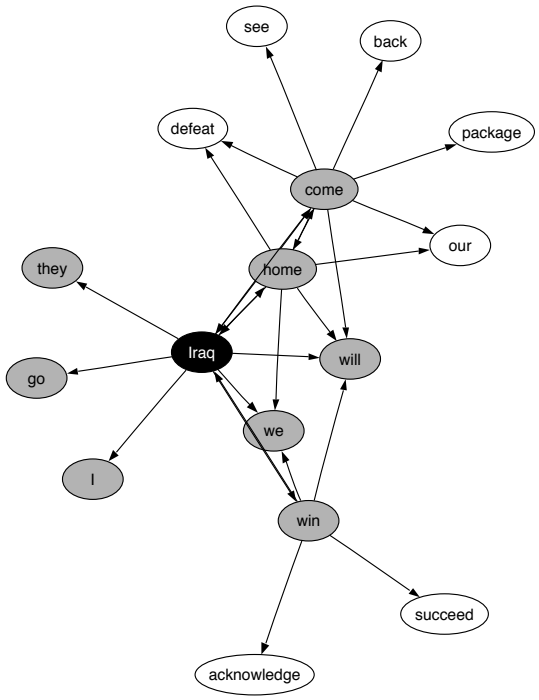

Abbildung 2: Kollokatoren zu *Iraq*, die John McCain typischerweise während der ersten TV-Debatte verwendete (vgl. Bubenhofer u. a. 2008a); Grafik: Forschergruppe semtracks.

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

I

Theoretische Erkenntnisse der korpus- und webbasierten Phraseologie-Untersuchungen

The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology

Jean-Pierre Colson

In diesem Beitrag wird mit Hilfe von drei Experimenten untersucht, welche Rolle strukturelle, statistische und informationstheoretische Faktoren für die Phraseologie und ihre theoretischen Grundlagen spielen. Im ersten Experiment wird versucht, adjektivische Kollokationen automatisch aus dem Internet zu extrahieren, indem Modifikatoren (z. B. *sehr, wenig, zu…*) eingesetzt werden. Die ersten Ergebnisse sind ermutigend, weil die Kollokationen, die auf diese Weise extrahiert werden können, in den meisten Fällen zahlreicher und vielfältiger sind, als in den traditionellen Wörterbüchern. Die Interdependenz zwischen strukturellen Faktoren und Phraseologie ist ein spannendes Forschungsgebiet, das für Sprachstudenten auch praktische Aspekte enthält.

Das zweite Experiment fragt nach den Frequenzen von phraseologischen N-Grammen. Studien zeigen, dass die Verteilung von N-Grammen in sehr großen Korpora und im Internet in der gleichen Weise wie die der Einwortlexeme dem Gesetz von Zipf-Mandelbrot folgt. Die Frequenzen von phraseologischen N-Grammen (in diesem Beitrag von 100 adjektivischen Kollokationen) bringen jedoch im Vergleich zu den Frequenzen von zufällig gewählten N-Grammen keine Unterschiede hervor. Für den theoretischen Rahmen der Phraseologie hat diese Entdeckung Konsequenzen, die noch untersucht werden müssen. Im dritten Experiment wird versucht, einen neuen Algorithmus zu entwerfen, um phraseologische Trigramme automatisch zu extrahieren.

Insgesamt gibt es viele Hinweise, dass strukturelle, statistische und informationstheoretische Elemente bei der phraseologischen Theoriebildung eine bedeutende Rolle spielen.

1 Introduction

Phraseology, the study of set phrases or fixed expressions in the broad sense, has become a major linguistic movement in Europe in the second half of the 20th century, and this is partly due to the many studies that were carried out within the framework of the European Association for Phraseology (Europhras; *http://www.europhras.org*). This does not mean, however, that there is a general theory of phraseology available. Indeed, too many issues

covered by the research on phraseology remain controversial or require more evidence. Most studies have so far concentrated on semantic and cognitive aspects of set phrases (for an overview see Burger 1998; 2007; Burger et al. 2007), but other issues ought to be explored and combined with that approach if a general theory is to be established. A major contribution in that direction has come from Dobrovol'skij and Piirainen (2005). Their comprehensive survey of figurative language and its interaction with idioms really lays the basis for broader theoretical considerations.

It is still unclear, however, what exact relations prevail between the main categories of set phrases: semantic set phrases (idioms, quasi idioms, collocations, etc.), communicative set phrases (routine formulae, clichés), and grammatical set phrases. If the phraseological approach is right, all of them share a number of common features. But is there hard evidence for this, apart from semantic and cognitive criteria?

There is on the other hand still a wide variety of linguistic theories, and the researchers from one country or one language group are not always aware of discoveries made by their colleagues from another part of the scientific world. This appears to be particularly true in the case of phraseology, whose basic principle, the existence of many set phrases in all languages, is regularly taken over by researchers who are not even aware of the very notion of phraseology. In recent years, the most striking example comes from construction grammar (Goldberg 1995; Croft 2001), and usage-based language acquisition (Tomasello 2003). As pointed out by Hüning (2007), those theories show a striking similarity with phraseology, although the term is never mentioned. The same holds true of Biber's notion of *lexical bundles* (Biber 1999).

A thorough discussion of all those theoretical issues falls beyond the scope of the present paper, in which just one complementary viewpoint on phraseology will be covered: the conbribution of Web-based corpus linguistics.

2 The structural nature of phraseology: an experiment

If phraseology is to become a global theory of language, all its semantic and cognitive aspects but also its structural nature should be the object of accurate analysis. Hard evidence should also be found for the importance of set phrases and their relationship with syntax and lexicon. With this objective in mind, we wished to test the structural nature of phraseology by means of an experiment with a Web corpus.

It should first be reminded that previous studies have underlined the frequent co-occurrence of set phrases with particles in German and French (Gréciano 1997) and of set phrases with introducers in Czech and English (Čermák 2002); in French, English and Dutch (Colson 2004). A well-known example is that of *proverbial*, a very common introducer of idioms in English, as in *to spill the proverbial beans*. In our experiment, we wanted to investigate further this structural aspect of set phrases by taking a closer look at noun / adjective collocations in a few European languages. If structural principles are closely linked with phraseology and in this case with collocations, we would then expect

that the recourse to structural patterns in the world's largest corpus, the World Wide Web, results in the extraction of a great many collocations.

The combination of computational linguistics and Web corpora makes it possible to test hypotheses such as this one in a number of minutes. From a technical point of view, we had recourse to the Application Programming Interface (API) of the search engine Yahoo (*http://developer.yahoo.com/search/*), a simple interface allowing automated requests to be passed to the search engine in a legal way. We based our search request in English on the modifiers *most, rather, quite, too, very* in the immediate context of a given noun, because they yielded the highest number of results. Thus, for *career*, the search engine rendered all examples containing for instance *a most brilliant career*. Another program then extracted all the relevant adjectives and aligned them in a table. All this took about 20 seconds per noun on an average personal computer. Table 1 shows the results obtained in this way for the English noun *career*. The frequency of the combination on the Yahoo API is also mentioned.

Obviously, one might argue about the collocational character of many of those noun + adjective combinations, which poses again the theoretical question of the relationship between frequency, co-occurrence and phraseology, a matter that is still far from being settled. Our point is precisely that this debate can only lead to fruitful and objective results if massive examples are gathered from huge corpora such as the Web. Our automatic program also extracted a few combinations that included no adjective (e.g. *recession job*), but the overwhelming majority of them were indeed relevant collocations. It should also be stressed that a simple such as the one described here provides far more results than any traditional dictionary or even dictionaries of collocations. This technique not only sheds light on a few theoretical issues underlying phraseology, but it is also of practical use for language students and translators, who often feel uncertain about the right adjective to be combined with a noun in a given context.

On the theoretical side, our experiment confirms the interdependence between structural patterns and phraseology in the case of collocations. This matter should of course be further investigated, but it shows a first application of Web-based corpus linguistics to the theory of phraseology.

3 Phraseology and word distributions

Although phraseology as a whole may represent as much as 50 percent of any written text (Sinclair 1991 and his *idiom principle*), most set phrases will correspond to very low frequencies in linguistic corpora. Moon (1998) and Colson (2003, 2004, 2007, 2008) have indeed shown that most set phrases display a frequency of less than one occurrence per million words. This means that even large corpora of 100 million words such as the commercial version of the British National Corpus will not yield so many instances of a given set phrase. Moon (1998) even pointed out that many common English set phrases were hardly represented in the British National Corpus.

long career	3 100 000	consistent career	4370
rewarding career	1 640 000	easy career	3750
good career	840 000	stressful career	2910
current career	694 000	inspiring career	2890
different career	643 000	large career	2710
short career	522 000	gratifying career	2690
possible career	480 000	logical career	2520
remarkable career	450 000	bizarre career	2500
brief career	430 000	smooth career	2140
young career	427 000	risky career	2060
extensive career	318 000	dear career	1880
extraordinary career	275 000	rare career	1750
impressive career	274 000	surprising career	1740
amazing career	272 000	uncertain career	1560
diverse career	252 000	crucial career	1230
unique career	245 000	intriguing career	1180
interesting career	238 000	narrow career	1090
late career	232 000	stagnant career	831
important career	215 000	marketable career	830
clear career	194 000	comparable career	655
strong career	143 000	classy career	604
attractive career	137 000	flat career	586
effective career	130 000	bumpy career	575
strange career	130 000	arduous career	515
appropriate career	118 000	obscure career	495
popular career	118 000	shaky career	455
typical career	108 000	apparent career	447
serious career	105 000	weak career	360
similar career	101 000	straightforward career	359
demanding career	88 900	unclear career	329
controversial career	84 300	slim career	298
useful career	74 500	paradoxical career	235
colorful career	71 500	fleeting career	201
unusual career	70 400	dependent career	190
relevant career	68 900	thankless career	179
common career	62 500	murky career	173
high career	50 500	opportunistic career	163
suitable career	45 800	inexplicable career	145
broad career	42 600	pointless career	137
essential career	40 700	ironic career	130
difficult career	33 800	remote career	112
hard career	17 000	pleasing career	109
small career	6640	ungrateful career	6

Table 1: Automatically extracted adjective collocations for *career* (Web corpus based on the Yahoo API)

In view of the rather sparse evidence for phraseology in large traditional corpora, many researchers, foreign language teachers, students, translators and terminologists have turned to the Web as a linguistic corpus. There are, however, some objections to using a WAC (Web as Corpus) methodology. We will limit ourselves to just a few:

1. To some extent, the language used on the Web may be seen as artificial and may not reflect ordinary language use.
2. Many topics are overrepresented on the Web: sex, music, sport, etc.
3. The Web has not been assembled by linguists (and it is therefore not a corpus in the strict sense), and it contains all kinds of language errors. *Absorbtion* (instead of the correct form *absorption*), for instance, displays a frequency of 918 000 on Google.

It should however be pointed out that most objections against the use of WAC techniques have been refuted by Kilgariff and Grefenstette (2003). In short, there appear to be far more advantages than drawbacks in using the Web as a linguistic corpus. Besides, Baroni (2008) has shown that the lexical distributions of the British National Corpus on the one hand and of the Web on the other are quite similar, which confirms the hypothesis that the huge size of the Web largely compensates for its weaknesses.

From a technical point of view, researchers deriving linguistic information from the Web and particularly from search engines should be very cautious about two real shortcomings of the Web: frequency results and *spamdexing*.

Frequency results fluctuate a lot from one search engine to another, or even from one month to the other. Those results are yielded by a complex algorithm and actually represent an approximation that is computed each time, so that changing news and all kinds of parameters may interfere with the results. There is besides a normalisation process in every search engine, involving capitalisation (same frequency for *poles* and *Poles*), as well as the suppression of punctuation marks. As pointed out by Lüdeling et al. (2007), *Google frequencies* should therefore not be trusted unconditionally.

Spamdexing refers to the practice of deliberately manipulating search engines in order to receive a higher ranking for a web page. As a result, many pages sent by Google or other search engines are endless repetitions of the same quotation, paragraph or address. To mention just one example: in December 2008 the phrase *jets amid intense criticism* corresponded to an impressive frequency of 19 600 on Google, which seemed to give it the status of a set phrase, but all results were actually extracted from the same repeated quotation.

With this word of caution about the pitfalls of Web-based corpus research, we can now turn to the heart of the matter: what do we exactly know about the distribution of words and word combinations in huge corpora? Although this matter is still far from being settled, Baroni (2008) confirms that word distributions in huge corpora roughly follow the Zipf-Mandebrot law. According to this principle (Zipf 1949, with an extension by Mandelbrot 1953), the general distribution of words displays a very limited number of high frequency items, a fair amount of average frequency words, and a *long* tail of words with extremely low frequency (an *army of dwarves*), according to the following formula:

$$f(w) = \frac{C}{(r(w)+b)^a} \tag{3.1}$$

where $f(w)$ represents the frequency of a word, and $r(w)$ its frequency rank.

For mathematical reasons, we also have:

$$\log f(w) = \log C - a \log(r(w) - b)) \tag{3.2}$$

We follow here the same method as in Baroni (2008): b is first set to 0; $\log C$ and a are calculated according to the data, using the least squares method. If necessary, b is then increased in small steps until the goodness of fit of equation 3.2 for the highest ranks no longer improves.

Baroni (2008) suggests that this Zipf-Mandelbrot distribution may also hold true of *word combinations* (n-grams) as well. This is a key issue to be addressed by a theory of phraseology. If phraseology is to be seen as one of the major driving forces underlying natural language, this should somehow be reflected in the language distributions. The complex intertwining between frequency, meaning, syntax and phraseology remains largely obscure, but the available huge corpora and the Web in the first place now make it possible to shed some gleams of light on this thorny issue.

This is where our second experiment comes in. The starting point is trivial: an English text containing about 100 bigrams (combinations of two words):

> I normally drink between 18 and 24 units per week. I consider the week worryingly quiet if I haven't had cause to drink one bottle of wine by Wednesday. Tuesday feels far enough into the week to be a heavy night, then the next night is usually a day of relative rest. If I don't have another big night on Thursday, I will let loose on Friday or Saturday. That'll be a good bottle. Sunday is generally a recovery night. I pondered the dry week ahead. I realised I had not had one in four years. I felt like a voyager setting off to a featureless land. At 23, did I really need alcohol to enjoy my evenings? (The Times Online, August 03, 2005)

As most texts, this fragment is a good combination of free and set phrases. The frequencies of all bigrams contained in this text were checked on the Web (Google search engine). Figure 1 below displays the results in a log-log table. This means that x stands for the $\log_{10}$ of the rank of the items (shown in decreasing order of frequency) and y for the log of the frequency for each item. If the Zipf-Mandelbrot principle applies here, such a log-log table should (for mathematical reasons) display a straight line, with an abrupt fall at the right end of the table and some minor irregularities along the line (Mandelbrot's corrections).

Figure 1 presents the log-log results for the bigrams (*BigramsDrinking*), as well as two projections according to the law of Zipf-Mandelbrot. We follow here the same method as Baroni (2008) for the bigram frequencies on the Brown Corpus (1 000 000 words): b is set to 0, and a is estimated according to the least squares method, which in this case yields: $a = 0.76$ (in the legend of Figure 1: *ZipfMan076*). By way of comparison, Figure 1 also presents a projection with $a = 1$ (*ZipfMan1*).

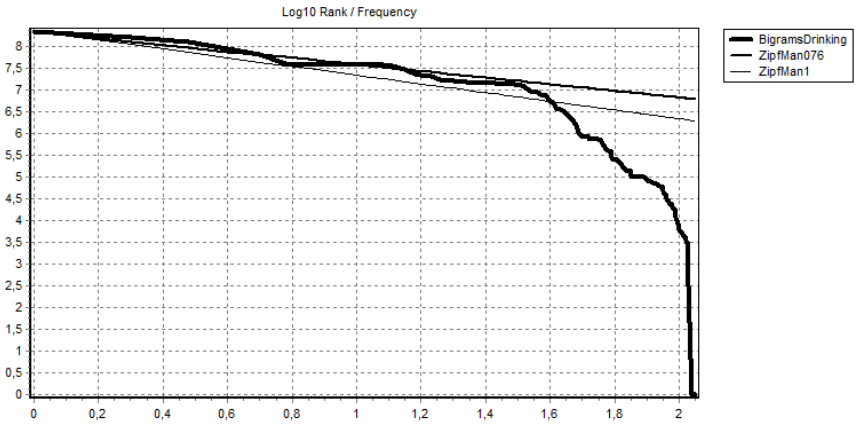

Figure 1: Bigram frequencies in an English text, presented in a log-log table (web frequencies with Google API)

As shown in Figure 1, the distribution of the bigrams in this text comes closer to a Zipf-Mandelbrot distribution with $a = 0.76$ than with $a = 1$. The matching is not quite perfect, and there is besides an abrupt fall at the right end of the table. A possible explanation might be that the results are based on a text containing only 100 words. As mentioned above, Baroni (2008) has checked all bigrams from the Brown Corpus (1 000 000 words) and has demonstrated that the distribution of the English bigrams follow the Zipf-Mandelbrot law if we set a to 0.76.

Among the 100 bigrams from this text, all kinds of word combinations occur. Some of them make no real communicative sense (*phrases The*, *pondered the*, *alcohol to*), other ones are grammatical combinations (*as in*, *cause to*, *did I*), and some are set phrases in the broad sense (*heavy night*, *let loose*, *felt like*). But what will now happen if we start not from mixed types of bigrams, but from 100 bigrams of which we know that they are collocations?

This time, 100 adjective collocations were selected from general dictionaries, collocation dictionaries, and the Internet in combination with the English nouns *criticism*, *job*, *proposal*, *reaction* and *remark*. For lack of space, only the frequencies of those adjectival collocations with *job* (e. g. *decent job*) are displayed in Figure 2, following the same method as in Figure 1.

It is quite striking to notice that the picture is roughly the same as in Figure 1: the 100 adjectival collocations display a tendency toward a Zipf-Mandelbrot distribution, at least if $a = 0.76$, and an abrupt fall at the right end of the table, probably due to the limited number of bigrams (100).

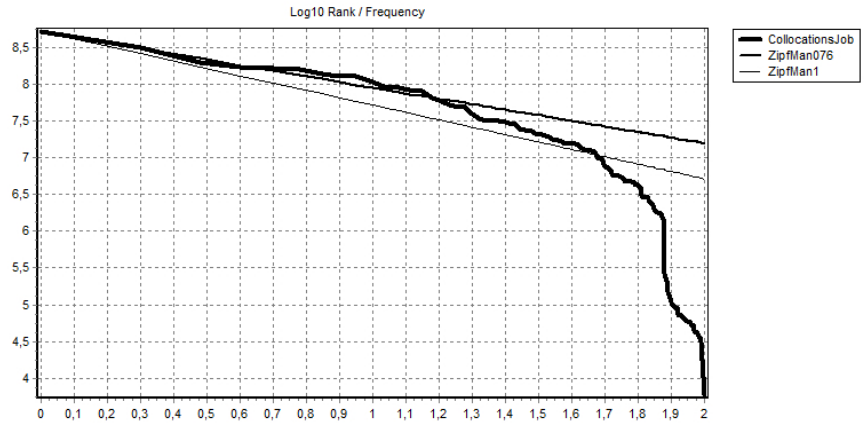

Figure 2: Bigram frequencies for adjective collocations, presented in a log-log table (web frequencies with Google API)

This is most unexpected, because collocations have often been defined by *frequency criteria*. Native speakers, teachers, translators will for instance point out that "this collocation is not common" or "not frequent". Not only language users, but also theoreticians have used raw frequency or modified frequency as a criterion for defining and extracting collocations: Justeson/Katz (1995); Frantzi/Anadiou/Mima (2000); Maynard/Anadiou (2000).

Now, if frequency is a decisive criterion for recognising collocations (as opposed to ordinary bigrams), we would expect in our example (composed of attested adjectival collocations with *job*) a LOT of very frequent combinations, at roughly the same figures for many of them, without sharp decreases. The picture we would then expect from Figure 2 would, in other words, be VERY different from that of Figure 1, if frequency were the main feature of collocations.

This is obviously not the case: no matter what the explanation for the distribution is (tendency toward Zipf-Mandelbrot or not), the picture in Figure 2 shows no clear difference with Figure 1. We have to be very careful about the interpretation of this phenomenon, but it should certainly be addressed by a global theory of phraseology. For one thing, it seems to confirm that pure frequency cannot be used as a reliable criterion for defining or extracting collocations.

4 Fresh insights into co-occurrence and fixedness

The majority of studies on phraseology have been based on semantic and cognitive principles. Obviously, these two aspects are of crucial importance for the production and interpretation of set phrases in context. Thus, the *hard shoulder* (the reserved area beside a motorway) is immediately recognised by a native speaker as being one element of meaning, whereas foreign language learners are often at a loss when confronted with such phrases. There is apparently nothing enabling us to distinguish between a policeman that is really asleep and a *sleeping policeman* meaning a speed hump or road hump (in American English: a speed bump). We have to look it up in a dictionary or ask a native speaker to solve the riddle.

Basing the definition of phraseology on meaning is somehow circular. If notions such as *idiomaticity*, *frozenness* or *fixedness* are used, these actually only make sense with reference to phraseology itself. *Idiomaticity*, for instance, is not an observable phenomenon and relies largely on the intuition of a native speaker. *Non-compositionality* is in the same way an interesting notion, but language is not mathematics and it is not always easy to determine whether a phrase is idiomatic, non-compositional or opaque. As there is no agreement on a general semantic theory, basing the theory of phraseology on semantics is not an easy task.

While meaning certainly plays a crucial role in the very essence of set phrases, a global theory of phraseology should also explore other ways of defining its founding concepts. In the case of collocations, considered by most researchers as a major category within phraseology and sometimes seen as a generic term for set phrases, the Anglo-Saxon tradition has now largely abandoned the semantic definition in favour of a statistical one: "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991: 6-7). The great advantage of this statistical criterion is that it yields reproducible experiments, one fundamental point in a fully scientific approach. Moreover, it is no longer necessary to be a native speaker of the language in order to identify collocations in any text.

If a fully automatic extraction of all types of set phrases was possible, this would bring about a linguistic change comparable to the Copernican Revolution in astrophysics. Detecting all set phrases in a text is indeed a crucial step in computer-aided and automatic translation, translation quality assessment, automated text correction, computational terminology and lexicography, to mention just a few domains. As Heid (2007:1041) puts it, "there is a growing need for semi-automatic or automatic tools to extract phrasemes from text". Unfortunately, these tools have so far mainly shown their limitations. Although the quest for the perfect algorithm has been attracting researchers for a few years now (Church/Hanks 1990, Biber 1999, Evert 2004, Deane 2005, Heid 2007), the diversity of statistical parameters used (more than 30 in total: t-score, log-likelihood, dice, mutual information etc.) has proved to be largely insufficient and can only be improved by adding grammatical criteria (Pazos Bretaña/Pamies Bertrán 2008). Moreover, most studies are restricted to bigrams and cannot easily be extended to higher grams: "If the mathematics of word pair statistics is based on the well-understood 2x2 contingency table, an extension to higher dimensions (three, four, ... words) is not yet fully understood even theoretically"

(Heid 2007:1042). It is a pity that no automatic extraction is possible for trigrams, because trigrams would already be much more reliable detectors of phraseology than bigrams. As pointed out by Heid (2007:1039), they may indeed be idiomatic as a whole (e.g. *spill the beans*), idiomatic because they are instances of regular patterns (e.g. *middle class background*), or idiomatic because they include an idiomatic bigram, typically a collocation (e.g. *writing sharp criticism*).

The beneficial contribution of Web-based linguistics to this debate may come again from the search engine APIs. By sending automated requests to the Application Programming Interface of search engines, an impressive amount of instantaneous frequencies and linguistic examples is now within hand's reach.

In our research project, several new techniques for extracting trigrams and higher grams are tested on the APIs of search engines. Further experimentation is however necessary before this method may be considered as reliable. The ultimate goal is to reproduce somehow the natural association that native speakers will immediately recognise between constituent elements of set phrases. Thus, a native speaker of English confronted with *down and hatch* knows that those two words are often associated in the phrase *down the hatch*. His competence contains this piece of information, probably related to the frequency of occurrence in his linguistic experience. It is not excluded, therefore, that a skilful manipulation of the search engine API would simulate this natural association in some way. If you type, for instance, "down * hatch" on Google, most results that are sent back (not all of them) are indeed instances of the phrase *down the hatch*. This is not the case when typing "with * man", hoping to get *with the man*. Building on this, several statistical parameters may be computed, taking into account frequency figures with and without the asterisk. Equation 4.1 shows such a formula that we are presently testing.

$$\begin{aligned} \mathrm{T} &= \mathrm{Trig}(\mathrm{G1}, \mathrm{G2}, \mathrm{G3}) \\ \mathrm{T\#} &= \mathrm{Trig}(\mathrm{G1}, *, \mathrm{G3}) \\ \mathit{WLR} &= 100\ln(\mathrm{fT})/\ln(\mathrm{fT\#}) \end{aligned} \tag{4.1}$$

The algorithm in equation 4.1 reads as follows: T stands for a trigram composed of three grams G1, G2, G3. In T#, we replace the central gram by an asterisk (wild card replacing any word). We then check the web frequency on Google (Web Log Ratio) by dividing the natural logarithm of the respective frequencies and we multiply by 100 in order to get an easily readable score ranging from 0 to 100. The logarithm is used in order to limit somehow the huge differences between extreme frequencies.

This is just a tentative formula, but it already produces interesting results. Consider the following English sentence:

"We understand that nothing can be done to fix the electorate's grave mistake; there are no hanging chads to save America this time." (The Harvard Crimson, November 15, 2006)

The WLR score for all trigrams from this sentence is presented in table 2.

It is noteworthy that the high scores in this passage indeed correspond to set phrases at the level of trigrams (*we understand that*, *can be done*, *there are no*) or to embedded

We understand that	66	grave mistake there	18
understand that nothing	45	mistake there are	37
that nothing can	49	there are no	90
nothing can be	66	are no hanging	35
can be done	96	no hanging chads	97
be done to	80	hanging chads to	34
done to fix	84	chads to save	8
to fix the	78	to save America	66
fix the electorate's	50	save America this	18
the electorate's grave	4	America this time	32
electorate's grave mistake	50		

Table 2: Automatic extraction of idiomatic trigrams from an English text

set phrases at the level of bigrams (*nothing can be*, *be done to*, *done to fix*, *no hanging chads*). The American phrase *hanging chads* is a particularly interesting example. It refers to the controversial 2000 presidential election between George Bush and Al Gore (some votes were considered as being invalid because of some incomplete punched holes, leaving some pieces of paper hanging: the hanging chads). Phraseology is very hard to grasp for foreign language learners, especially when it contains references to recent history or politics. In the theory of phraseology, this aspect may be taken into account by improving the automatic extraction of phrases. If applied to newspaper articles, this method is really useful in advanced foreign language learning, because it enables students to discover new phrases that are often absent from all dictionaries (at the moment of this writing, the phrase *hanging chads* is only explained on Wikipedia).

5 By way of conclusion

There is nowadays overwhelming evidence that structural, statistical and information-theoretical aspects play a crucial role in phraseology. This evidence comes not only from phraseological studies, but also from corpus linguistics, computational linguistics and construction grammar. Understanding the semantic and cognitive underpinnings of phraseology is of paramount importance, but language is complex and should be studied from several perspectives. Bridging the gap between those complementary approaches may in the future constitute a global theory of phraseology.

Bibliography

Baroni, M. (2008): "Distributions in text". In: Lüdeling/Kytö (eds.) (2008); 803–821.
Biber, D. (1999): "Lexical bundles in conversation and academic prose". In: Hasselgard/Oksefjell (eds.) (1999); 181–189.

Bouillon, H. (éd.) (2004): *Langues à niveaux multiples. Hommage au Professeur Jacques Lerot à l'occasion de son émériat.* Louvain-la-Neuve.
Burger H. (2007): *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin.
Burger, H./Häcki Buhofer, A./G. Gréciano (Hrsg.) (2003): *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zu Methodologie und Kulturspezifik der Phraseologie.* Baltmannsweiler.
Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N.R. (eds.) (2007): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research* Berlin/New York.
Church, K. W./Hanks, P. (1990): "Word association norms, mutual information and lexicography", in: *Computational Linguistics, 16*; 22–29.
Colson, J.-P. (2003): "Corpus Linguistics and Phraseological Statistics: a Few Hypotheses and Examples". In: Burger et al. (eds.) (2003); 47–59.
Colson, J.-P. (2004): "Phraseology and computational corpus linguistics: from theory to a practical example". In: Bouillon (éd.) (2004); 35–45.
Colson, J.-P. (2007): "The World Wide Web as a corpus for set phrases". In: Burger et al. (eds.) (2007); 1071–1077.
Colson, J.-P. (2008): "Cross-linguistic phraseological studies: an overview". In: Granger/Meunier (eds.) (2008); 191–206.
Croft, W. (2001): *Radical construction grammar. Syntactic theory in typological perspective.* Oxford.
Čermák, F. (2002): "Text Introducers of Proverbs and Other Idioms". In: Földes/Wirrer (Hrsg.) (2002); 27–46.
Deane, P. (2005): "A nonparametric method for extraction of candidate phrasal terms", in: *43d Annual Meeting of the Association for Computational Linguistics*; 605–613.
Dobrovol'skij, D./ Piirainen, E. (2005): *Figurative Language. Cross-Cultural and Cross-Linguistic Perspectives.* Amsterdam.
Evert, S. (2004): *The statistics of word cooccurrences, word pairs and collocations.* Stuttgart.
Földes, C./Wirrer, J. (Hrsg.) (2002): *Phraseologismen als Gegenstand sprach- und kulturwissenschaftlicher Forschung.* Baltmannsweiler.
Frantzi, K./Ananiadou, S./Mima, H. (2000): "Automatic recognition of multi-word terms: the CValue and NC-Value Method", in: *International Journal on Digital Libraries*, 3; 115–130.
Goldberg, A.E. (1995): *Constructions: a construction grammar approach to argument structure.* Chicago.
Granger, S./Meunier, F. (eds.) (2008): *Phraseology; an interdisciplinary perspective.* Amsterdam, Philadelphia.
Gréciano, G. (1997): "Qui se ressemble s'assemble: locutions, particules et compères", in: *Nouveaux Cahiers d'Allemand, 17*; 451–460.
Ha, L.Q./Sicilia-Garcia, E./Ming, J./Smith, F. (2002): "Extension of Zipf's law to words and phrases". In: Proceedings of COLING (2002); 315–320.
Hasselgard, H./Oksefjell, S. (eds.) (1999): *Out of corpora: studies in honor of Stig Johansson.* Amsterdam.
Heid, U. (2007): "Computational linguistc aspects of phraseology II". In: Burger et al. (eds.) (2007); 1036–1044.
Hoey, M. (1991): *Patterns of Lexis in Text.* Oxford.
Hundt, M./Nesselhauf, N./Biewer, C. (eds.) (2007): *Corpus linguistics and the Web.* Amsterdam.
Hüning, M. (2007): "Constructiegrammatica als verlengstuk van de fraseologie". In: Moerdijk et al. (eds.) (2007); 377–384.
Jackson, W. (ed) (1953). Communication theory. London.

Justeson, J. S./Katz, S. M. (1995): "Technical terminology: some linguistic properties and an algorithm for identification in text", in: *Natural Language Engineering*, 1; 9–27.
Kilgariff, A./Grefenstette, G. (2003): "Web as corpus. Introduction to the special issue", in: *Computational Linguistics, 29*; 1–15.
Lüdeling, A./Evert, S./Baroni, M. (2007): "Using Web data for linguistic purposes". In: Hundt et al. (eds.) (2007); 7–24.
Lüdeling, A./Kytö, M. (eds.) (2008): *Corpus linguistics. An international handbook.* Berlin, New York.
Mandelbrot, B. (1953): "An informational theory of the statistical structure of languages". In: Jackson (ed.) (1953); 486–502.
Maynard, D./Ananiadou, S. (2000): "Identifying Terms by their Family and Friends". In: Proceedings of COLING (2002); 530–536.
Moerdijk, F./Van Santen, A./Tempelaars, R. (eds.) (2007): *Leven met woorden. Opstellen aangeboden aan Piet Sterkenburg bij zijn afscheid als directeur van het Instituut voor Nederlandse lexicologie en als hoogleraar Lexicologie aan de Universiteit Leiden.* Leiden.
Moon, R. (1998): *Fixed Expressions and Idioms in English.* Oxford.
Sinclair, J. (1991): *Corpus, Concordance, Collocation.* Oxford.
Pazos Bretaña, J.-M./Pamies Bertrán, A. (2008): "Combined statistical and grammatical criteria for the retrieval of phraseological units in an electronic corpus". In: Granger/Meunier (eds.) (2008); 391–406.
Tomasello, M. (2003): *Constructing a language. A usage-based theory of language acquisition.* Cambridge (Mass.).
Zipf, G.K. (1949): *Human Behavior and the Principle of Least Effort.* Cambridge (Mass.).

Prof. Dr. Jean-Pierre Colson
Institut Marie Haps
rue d'Arlon 11
1050 Bruxelles
Belgique

Université catholique de Louvain (Louvain-la-Neuve)
place B. Pascal 1
1348 Louvain-la-Neuve
Belgique
jean-pierre.colson@uclouvain.be

Häufigkeit und Struktur von Phraseologismen am Beispiel verschiedener Web-Korpora

Uwe Quasthoff/Fabian Schmidt/Erla Hallsteinsdóttir

Large German corpora (news, Web text, Wikipedia and Projekt Gutenberg) and a list of 5058 German phrases are used to analyze phraseology in different text genres and over time. Frequency statistics for words within multi word units and for multi word units within different genres are explored by describing some differences and testing Zipf's law. POS Tagging and the number of non-stopwords are used to classify German phrases quantitatively.

Einführung

Unser Ausgangspunkt ist die Beobachtung, dass die meisten Phraseologismen relativ selten sind, weshalb für statistische Aussagen über ihr Vorkommen große Textmengen benötigt werden. Hier werden größere Korpora verschiedener Textsorten untersucht, um die Unterschiede bei der Verwendung von Phraseologismen zwischen Zeitungstexten, zufällig ausgewählten Texten aus dem Web, Texten aus der Wikipedia sowie älteren literarischen Texten aus dem Projekt Gutenberg zu vergleichen. Neben den Häufigkeiten der einzelnen Phraseologismen und der Anwendbarkeit des Zipfschen Gesetzes werden auch Gruppen von Phraseologismen gleicher syntaktischer Struktur untersucht.

Geschriebene und gesprochene Sprache weisen grundlegende Unterschiede auf (vgl. ausführlich in DUDEN 4, 2006: 1180f.), sie „übernehmen in der gesellschaftlichen Kommunikation unterschiedliche Aufgaben und sind jeweils für bestimmte kommunikative Aufgaben besonders geeignet" (Stein 2007: 220). In der Phraseologie wird davon ausgegangen, dass die Mündlichkeit dominiert, d.h. dass Phraseologismen in der gesprochenen Alltagssprache wesentlich häufiger auftreten als in geschriebenen Texten (vgl. Stein 2007). Dass es Unterschiede in der gesprochenen und der geschriebenen Sprache gibt, hat ein Vergleich von Frequenzdaten aus Korpora und Sprecherdaten zur Geläufigkeit von Phraseologismen ergeben. Die Resultate zeigten, dass es eine Korrelation zwischen Geläufigkeit und Frequenz gibt in dem Sinne, dass Phraseologismen mit einer hohen oder mittleren Frequenz im Korpus auch von vielen Probanden verwendet werden. Allerdings weisen nicht alle geläufigen Phraseologismen eine entsprechend hohe Korpusfrequenz auf, was möglicherweise damit erklärt werden kann, dass diese vorwiegend zum mündlichen Sprachgebrauch gehören (vgl. Hallsteinsdóttir et al. 2006).

Leider gibt es bisher keine genügend großen Korpora gesprochener Sprache, um korpusbasierte Vergleiche der mündlichen und schriftlichen Phraseologie durchzuführen. In der Annahme, dass sich „*Vermündlichungstendenzen*“ (Stein 2007: 234) durch Unterschiede im Gebrauch von Phraseologismen in Korpora mit verschiedenen Textsorten widerspiegeln, wird hier auf Korpora geschriebener Sprache zurückgegriffen.

1 Datengrundlage

1.1 Korpusauswahl

Im Rahmen des Projekts Deutscher Wortschatz – *www.wortschatz.uni-leipzig.de* – stehen verschiedene deutschsprachige Korpora zur Verfügung, die hier vergleichend untersucht werden sollen. Diese Korpora wurden einheitlich aufbereitet: Zunächst wurden die Texte in Sätze zerlegt und fremdsprachiges Material wurde entfernt. Ebenso wurden mit musterbasierten Verfahren offensichtlich nicht wohlgeformte Sätze entfernt. Listen und Tabellen finden bei der folgenden Auswertung also keine Berücksichtigung. Außerdem wurden bei mehrfach auftretenden Sätzen die Dubletten entfernt. Die verbliebenen Korpora haben die folgenden Größen:

- Zeitungstexte: Texte großer Tageszeitungen und anderer Nachrichtenportale aus dem Jahr 2008, Umfang 27,8 Mio. Sätze;
- Web-Texte: Zufällig gesammelte deutschsprachige Texte aus dem Web, gesammelt im Jahr 2002, Umfang 100 Mio. Sätze;
- Projekt Gutenberg: Meist literarische Texte mit abgelaufenem Urheberrecht, Entstehungszeit also i.d.R. vor 1900; gesammelt 2003, Umfang 2,75 Mio. Sätze;
- Wikipedia-Texte: Material der deutschsprachigen Wikipedia aus dem Jahr 2007, Umfang 8,7 Mio. Sätze.

Auf den ersten Blick lassen die unterschiedlichen Textsorten ein stark voneinander abweichendes Vorkommen von Phraseologismen erwarten. Unsere Hypothese ist, dass die Texte aus dem Web durch die vermutete große Anzahl persönlicher Beiträge die größte Ähnlichkeit zu gesprochener Sprache haben, und deshalb einen höheren Anteil Phraseologismen enthalten. In den enzyklopädischen Texten der Wikipedia sind weniger Phraseologismen zu erwarten. Die Texte aus dem Projekt Gutenberg sind deutlich älter, was sich auch auf die Verwendung von Phraseologismen auswirken sollte. Inwieweit sich diese Vermutungen bestätigen, wird in Abschnitt 2.3 deutlich.

1.2 Auswahl der Phraseologismen

Bei der Untersuchung wird eine vorgefertigte Liste mit deutschen Phraseologismen eingesetzt. Der Großteil der Phraseologismen wurde zwei Wörterbüchern für Deutsch als Fremdsprache entnommen. Die im *Wörterbuch Deutsch als Fremdsprache* von de Gruyter

(*GDaF*) (Kempcke 2000) mit einem Stern (*) markierten Phraseologismen bildeten das Ausgangsmaterial. Zwar werden im *GDaF* keine genauen Kriterien für die Auswahl von Phraseologismen angegeben. „Der Weg der Materialselektion bleibt also opak" (Wotjak 2001: 269). Die Auswahl wird trotzdem – allerdings auch ohne genauere Angabe der Kriterien dafür – als gelungen und treffend angesehen: „Es wird also deutlich, dass das GDaF dem Lernenden ein sehr reichhaltiges, dabei aber durchaus geläufiges und nicht antiquiertes phraseologisches Material zur Verfügung stellt" (Wotjak 2001: 270). Somit ist davon auszugehen, dass sich die Phraseologismen aus diesem Wörterbuch gut für eine Untersuchung zur Frequenz deutscher Phraseologismen eignen. Die Phraseologismen wurden ergänzt mit den mit ID markierten Phraseologismen aus dem *Langenscheidt Wörterbuch für Deutsch als Fremdsprache* (Götz et al. 1997), den Phraseologismen aus dem Wörter- und Übungsbuch von Hessky/Ettinger (1997) und aus dem phraseologischen Wörterbuch von Langenscheidt (Griesbach/Schulz 2000) sowie durch die Liste intersubjektiv geläufiger deutscher Idiome von Dobrovol'skij (1997: 265-288). Insgesamt ergaben die Wörterbücher etwas über 6000 unterschiedliche Phraseologismen. Zu 5058 Phraseologismen konnten insgesamt 8397 weitgehend eindeutige Suchformen (vgl. Abschnitt 2.2) konstruiert werden. Bei den übrigen handelt es sich hauptsächlich um Phraseologismen mit einer zu variablen Form, wie. z. B. *einen Bart haben* oder *etwas für sich behalten* (vgl. ausführlich in Hallsteinsdóttir 2005 und Hallsteinsdóttir et al. 2006).

2 Häufigkeitsanalyse

2.1 Wortstatistik der Phraseologismen

Welche Wörter treten in Phraseologismen auf? Welche semantischen Merkmale sind besonders häufig?

Um diese Fragen zu klären, betrachten wir ein kleines Korpus, bestehend aus den 5058 Phraseologismen in ihrer Wörterbuchform. Die Worthäufigkeiten für Substantive in diesem Korpus werden mit den entsprechenden Worthäufigkeiten aus dem Webkorpus verglichen. Die Beschränkung auf Substantive erfolgt deshalb, da der momentane Schwerpunkt auf semantischen Eigenschaften der Wörter liegt.

Die folgenden Tabellen zeigen extrem auffällige Wörter in Phraseologismen sowie die häufigsten Wörter aus dem Zeitungskorpus, die (auch in einer anderen flektierten Form) in überhaupt keinem der betrachteten Phraseologismen auftreten.

In Tabelle 1 auf der nächsten Seite sind zunächst die 50 häufigsten Substantive aus dem betrachteten Zeitungskorpus aufgeführt. Die Wörter zeigen deutlich Schwerpunkte auf den Themen Politik und Wirtschaft, diese sind häufig mit Zahlenangaben verbunden (meist geht es um Geld), aber auch mit Eigennamen und Zeitangaben.

Im Vergleich dazu enthält Tabelle 2 auf Seite 41 die häufigsten Substantive aus den betrachteten Phraseologismen. Hier bietet sich erwartungsgemäß ein völlig anderes Bild, wir beobachten als führende Gruppe die Körperteile, aber auch religiöse Begriffe sowie weitere Wörter aus der Alltagswelt.

1. Prozent	14. Berliner	27. Leben	40. Wochen
2. Jahren	15. SPD	28. Polizei	41. Frauen
3. Jahr	16. Unternehmen	29. Kinder	42. Montag
4. Euro	17. Stadt	30. Angaben	43. DM
5. Millionen	18. Welt	31. Woche	44. Beispiel
6. Jahre	19. USA	32. Frage	45. Sonntag
7. Berlin	20. Mann	33. CDU	46. Dollar
8. Mark	21. Milliarden	34. Tag	47. Teil
9. Deutschland	22. Land	35. Dabei	48. Weg
10. Uhr	23. Frau	36. Platz	49. München
11. Zeit	24. Deutschen	37. Arbeit	50. Donnerstag
12. Menschen	25. Regierung	38. Fall	
13. Ende	26. Geld	39. Freitag	

Tabelle 1: Die häufigsten Substantive aus dem Zeitungskorpus

Beim Versuch einer Erklärung hierfür kann der bei Lerchner (2005: 60) „als der sprachzeichenhaft konstituierte Bezug von Signifikaten aus kulturspezifisch konzeptualisierten Segmenten kommunikations-gemeinschaftsbedingter Lebenswelten auf muttersprachspezifische Signifikanten“ definierte Nominationsprozess herangezogen werden. Phraseologismen entstehen durch sprachlich gezogene Analogien zu Alltagserfahrungen oder die Bezugnahme auf Dinge und Begebenheiten des Alltags. In diesen Bereichen existieren viele kurze Erbwörter. Die bevorzugte Verwendung kurzer Wörter in Phraseologismen könnte zusätzlich durch das Streben nach sprachlicher Ökonomie erklärt werden. Phraseologismen sind an sich komplexe Zeichen, die aus anderen Zeichen bestehen. Sprachökonomisch gesehen ist es daher sinnvoll, bevorzugt einfache Wörter als Komponenten in Phraseologismen zu verwenden. Natürlich finden wir auch in Phraseologismen längere und moderne Wörter aus dem technischen Bereich wie *Bahnhof* oder *Scheinwerferlicht*, aber diese tauchen in den Tabellen nicht auf.

In Tabelle 3 auf der nächsten Seite finden wir ähnliche Wörter wie in Tabelle 2, allerdings sind diesmal die Wörter nicht nach Häufigkeit sortiert, sondern ihre Häufigkeit in Phraseologismen wurde mit der Häufigkeit in Zeitungstexten verglichen und die Wörter wurden mittels Differenzanalyse (Heyer et al. 2006: 95) entsprechend der Abweichung sortiert. Dadurch werden die in Zeitungstexten selteneren Wörter aus Tabelle 2 auf der nächsten Seite deutlicher hervorgehoben.

Von Interesse ist natürlich auch das andere Extrem: Wörter, die zwar in Phraseologismen, aber sonst kaum auftreten. Dazu sind in Tabelle 4 auf Seite 42 einige Wörter aufgeführt, die im Korpus aus Zeitungstexten nicht unter den 500.000 häufigsten Wortformen sind, aber in Phraseologismen vorkommen. Bei einigen dieser Wörter handelt es sich um die

1. Kopf	14. Licht	27. Luft	40. Tür
2. Hand	15. Mann	28. Himmel	41. Gesicht
3. Welt	16. Auge	29. Kind	42. Spiel
4. Augen	17. Gott	30. Blut	43. Stein
5. Ohren	18. Hals	31. Ende	44. Haut
6. Zeit	19. Wort	32. Rücken	45. Herzen
7. Nase	20. Geld	33. Ohr	46. Händen
8. Sache	21. Teufel	34. Wind	47. Schritt
9. Mund	22. Zunge	35. Beine	48. Seite
10. Weg	23. Finger	36. Fuß	49. Tisch
11. Herz	24. Hände	37. Füßen	50. Brot
12. Leben	25. Wasser	38. Hund	
13. Tag	26. Boden	39. Tod	

Tabelle 2: Die häufigsten Substantive aus dem Korpus der Phraseologismen

1. Kopf	14. Ohr	27. Fuß	40. Sack
2. Hand	15. Hände	28. Leib	41. Strich
3. Ohren	16. Füßen	29. Luft	42. Leibe
4. Nase	17. Licht	30. Brot	43. Zahn
5. Zunge	18. Himmel	31. Händen	44. Haut
6. Hals	19. Maul	32. Herzen	45. Lachen
7. Mund	20. Blut	33. Hund	46. Kuckuck
8. Augen	21. Sache	34. Wort	47. Tür
9. Teufel	22. Pfifferling	35. Schlag	48. Fährte
10. Finger	23. Dreck	36. Stein	49. Welt
11. Auge	24. Rücken	37. Tasche	50. Trümpfe
12. Herz	25. Beine	38. Wind	
13. Gott	26. Boden	39. Hut	

Tabelle 3: Auffälligste Substantive aus dem Korpus der Phraseologismen

affenartiger	gehupft	Hammelbeine	Marschallstab
Bärenhaut	geölter	Hasenpanier	Maulaffen
Bärenhunger	gesengte	Herzdrücken	Schanden
Bohnenstroh	gespornt	Honiglecken	Schießhund
Dummsdorf	gestiefelt	Kanthaken	Spießruten
Espenlaub	Glacehandschuhen	Katzenwäsche	Strandhaubitze
Flitzbogen	gordischer	Kloßbrühe	Zuckerlecken

Tabelle 4: Wörter, die außerhalb von Phraseologismen extrem selten vorkommen

sog. Unikalia[1], d.h. phraseologisch gebundene Wörter wie *Bohnenstroh* oder *Maulaffen*. Bei anderen handelt es sich vermutlich um einzelne Flexionsformen die überwiegend in einem Phraseologismus vorkommen, bzw. um Wörter, die generell, zumindest in der geschriebenen Sprache, selten gebraucht werden, vgl. *mit **affenartiger** Geschwindigkeit* vs. *ein **affenartiger**, verkniffen blickender Gnom* bzw. *wie ein **geölter** Blitz* vs. *Landhaustisch mit **geölter** Massivholzplatte* (Bsp. aus Deutscher Wortschatz, gesehen am 9. April 2009).

Im Folgenden soll uns noch interessieren, wie sinntragende Wörter in einzelnen Phraseologismen vorkommen. Als sinntragend bezeichnen wir Wörter mit überwiegend semantischer Funktion (Autosemantika), also insbesondere keine Stoppwörter (Synsemantika). Als näherungsweise Antwort untersuchen wir die Frage, wie viele Wörter unterhalb einer Häufigkeitsschwelle in den einzelnen Phraseologismen auftreten. Diese Wörter werden hier auch als ‚wichtige Wörter' bezeichnet. Die Schwelle wird folgendermaßen gewählt: Bei der Zählung berücksichtigen wir nur die Wörter, die nicht unter den 500 häufigsten Wörtern im Zeitungskorpus und nicht unter den 50 häufigsten Wörtern im Phraseologismenkorpus sind. Hinzugefügt werden wieder alle Wörter mit großem Anfangsbuchstaben und mindestens vier Buchstaben, um auch die häufigen Substantive aus Tabelle 2 auf der vorherigen Seite zu berücksichtigen.

Tabelle 5 auf der nächsten Seite gibt erste Hinweise auf die Struktur von Phraseologismen: Mehr als 70% der untersuchten Phraseologismen enthalten zwei oder mehr wichtige Wörter. Im Abschnitt 4 wird zusätzlich die syntaktische Struktur einbezogen.

2.2 Suchformen zur Häufigkeitsbestimmung von Phraseologismen

Abhängig von ihrer syntaktischen Form können einige Phraseologismen in einer erheblichen Formenvielfalt auftreten. Dies wird erreicht durch Flexion von einzelnen Komponenten, Variabilität in der Wortstellung und eine mögliche Trennbarkeit. In vielen Fällen gibt es jedoch einen aus einem oder mehreren Wörtern bestehenden Teil des Phraseologismus, der praktisch unveränderlich ist und in der überwiegenden Mehrzahl der Formen des Phraseologismus vorkommt. Ist diese Suchform zusätzlich nicht oder nur selten außerhalb des Phraseologismus zu beobachten, dann lässt sich die Häufigkeit des Phraseologismus sehr gut durch die einfacher zu bestimmende Häufigkeit der Suchform approximieren. Die

1 Vgl. die „List of German Bound Words": *http://www.sfb441.uni-tuebingen.de/a5/codii/*

Anzahl wichtiger Wörter	Anzahl Phraseologismen	Beispiele
0	127	alles in allem so weit, so gut
1	1469	wie eh und je es wird schon klappen viel um die Ohren haben
2	2520	wie er leibt und lebt etwas stinkt zum Himmel ein Vorschlag zur Güte
3	1111	sich recht und schlecht durchschlagen jdm. den Himmel auf Erden versprechen die Trennung von Tisch und Bett
4 oder mehr	414	wissen, wo Barthel den Most holt den bitteren Kelch bis zur Neige leeren eine kesse Sohle aufs Parkett legen

Tabelle 5: Phraseologismen mit unterschiedlicher Anzahl wichtiger Wörter

bisherigen Analysen lassen die Schlussfolgerung zu, dass in der Regel weniger als 10% der Belege eine nicht-phraseologische Realisierung haben, ja dass sogar bei den meisten Phraseologismen eine nicht-phraseologische Bedeutung gar nicht vorkommt. Dies gilt auch für Phraseologismen, die durchaus eine wörtliche Lesart in freier Wortverbindung ergeben könnten, wie *grünes Licht geben*, *jmd. unter die Arme greifen*, *etw. auf Eis legen*, *gegen den Strom schwimmen* (vgl. Hallsteinsdóttir 2007).

In einigen Fällen können auch mehrere Suchformen für einen Phraseologismus angebracht sein. Während ein unveränderlicher Phraseologismus natürlich selbst als Suchform benutzt werden kann und typische Wörter wie die aus Tabelle 4 auf der vorherigen Seite sich für Suchformen anbieten, sind in ungünstigen Fällen mehrere Suchformen nötig.

Die Suchformen wurden manuell konstruiert, d. h. sie basieren auf der Intuition eines Sprechers und den Angaben der Wörterbücher, und in einzelnen Fällen könnten daher Suchformen fehlen bzw. Suchformen verwendet worden sein, die nicht nur den Phraseologismus, sondern auch andere Phraseologismen oder freie Wortverbindungen erfassen. Eine zumindest stichprobenartige manuelle Kontrolle der Daten ist daher notwendig.

Tabelle 6 auf der nächsten Seite zeigt Suchformen für einige Phraseologismen und gibt auch Formen an, die nicht durch die Suchform abgedeckt werden.

Versucht man, die Suchformen zu charakterisieren, so fällt Folgendes auf: Für Variabilität sorgen hauptsächlich die Flektierbarkeit des Verbs und die Variabilität beim Pronomen. Wenn ein Phraseologismus eine Wortgruppe mit mindestens zwei Substantiven enthält (wie *Wolf im Schafspelz* oder *Zunge im Zaum* in Tabelle 6), so sind diese gute Kandidaten für

Phraseologismus	**Suchform(en)**	**Nicht abgedeckte bzw. zusätzliche Formen/Phraseologismen**
Wolf im Schafspelz	Wolf im Schafspelz	
die/seine Zunge im Zaum halten	Zunge im Zaum	
jdn. in seinen Bann schlagen	in seinen Bann in ihren Bann in meinen Bann in deinen Bann in unseren Bann in euren Bann	jdn. in seinen Bann ziehen (anderer Phraseologismus) „Wenigstens vorübergehend schlägt es uns in Bann.“ (fehlende Form)
(nur noch) der/ein Schatten seiner selbst sein	Schatten seiner selbst Schatten ihrer selbst	ein Schatten von sich selbst (fehlend, aber selten)
Tu (doch) nicht so!	Tu doch nicht so Tu nicht so	Tun sie doch nicht so (fast 50% der Vorkommen!)
am Ende sein am Ende der Welt	am Ende	am Ende stehen sich am Ende befinden mit seinem Latein am Ende sein am Ende der Fahnenstange
jdm. in den Arsch kriechen	in den Arsch	Speziell in pornographischen Webseiten häufig in nicht-phraseologischer Lesart der Suchform

Tabelle 6: Beispiele für Suchformen

Suchformen. Für andere Phraseologismen, wie z.B. bei *jdn. in seinen Bann schlagen*, muss möglicherweise auf mehrere Suchformen zurückgegriffen werden.

2.3 Häufigkeitsvergleich für verschiedene Textsorten

In diesem Abschnitt wird die Häufigkeit einiger ausgewählter Phraseologismen für die vier Korpora untersucht. Es werden Phraseologismen aus verschiedenen Häufigkeitsbereichen ausgewählt. Außerdem wurden neben in der Alltagssprache üblichen Phraseologismen zusätzlich veraltende und vulgäre Phraseologismen untersucht, die mögliche Unterschiede zwischen den Textsorten stärker hervortreten lassen.

Die Häufigkeiten wurden auf Auftreten pro Million Sätze normiert, um die Zahlen vergleichbar zu machen.

Für Tabelle 7 auf der nächsten Seite wurden solche Phraseologismen ausgewählt, deren Häufigkeiten sich in verschiedenen Textsorten mindestens um den Faktor 10 unterscheiden und die außerdem für mehrere Textsorten mehr als dreimal pro Million Sätze vorkommen. Zwar zeigt sich ein offensichtlicher Unterschied zwischen den älteren, literarischen Texten aus dem Projekt Gutenberg einerseits und den eher modernen Texten andererseits, aber es lassen sich auch noch feinere Unterscheidungen treffen. Die Daten geben Anlass zu folgenden Beobachtungen bzw. Vermutungen:

- Veraltende oder aus anderen Gründen seltener genutzte Wörter gehen Hand in Hand mit einem selteneren Gebrauch der Phraseologismen (Tabelle 7 auf der nächsten Seite: *Galopp*, *Gott*);
- Umgekehrt: Wörter oder Wortgruppen, die sich auf die heutige Welt beziehen, finden sich in häufiger verwendeten Phraseologismen wieder (*grünes Licht*, *in Kraft*);
- Wikipedia enthält nur halb so viele Phraseologismen wie das Projekt Gutenberg. Die Texte aus dem Web und die Zeitungstexte enthalten fast gleichviel Phraseologismen und liegen zwischen Wikipedia und Projekt Gutenberg;
- Vulgäre Phraseologismen treten im Vergleich zu Zeitung und Web selten in den älteren Texten auf, in der Wikipedia jedoch ganz besonders selten. Der Grund dafür ist möglicherweise das Streben nach Formulierung von einem neutralen bzw. wissenschaftlichen Standpunkt aus.

Es konnte kein deutlicher Unterschied zwischen den Texten aus dem Web und den Zeitungstexten festgestellt werden. Die Zahl der vorkommenden Phraseologismen ist fast gleich und es wird auch nicht deutlich, dass die eher im mündlichen Gebrauch vermuteten Phraseologismen im Web häufiger anzutreffen wären.

Um dieses Ergebnis nochmals zu überprüfen, haben wir einige der Phraseologismen aus der Untersuchung in Hallsteinsdóttir et al. (2006) untersucht, die eine hohe Geläufigkeit bei Muttersprachlern und eine eher niedrige Frequenz in Texten (Geläufigkeit 76-100%, aber weniger als 100 Belege) aufweisen. Die Hypothese ist, dass diese Phraseologismen zur gesprochenen Sprache gehören. Demzufolge müssten sie in den Webtexten häufiger

Phraseologismus	**Suchform**	**Projekt Gutenberg**	**Wikipedia**	**Web**	**Zeitung**
von sich hören lassen	von sich/mir/dir/uns/euch hören	82,47	3,00	22,56	6,83
ganz und gar nicht(s)	ganz und gar	915,86	24,82	123,95	107,37
um so mehr, als	um so mehr	373,46	24,59	107,28	26,99
nicht zuletzt	nicht zuletzt	17,44	474,48	658,22	548,44
jdm. nicht (mehr) unter die Augen kommen/treten (dürfen)	unter die Augen	69,03	1,04	3,60	1,40
zu Buche schlagen	zu Buche	0,36	21,70	57,74	225,51
in der/aller Frühe	in der/aller Frühe	149,68	4,04	15,25	9,99
im Galopp	im Galopp	101,36	4,27	6,49	3,09
um Gottes willen	um Gottes willen	186,73	1,15	14,72	6,47
etwas/jdn. im Griff haben	im Griff haben	3,27	14,89	137,04	180,20
in Kraft sein/bleiben/treten	in Kraft	28,34	341,49	358,80	402,16
in Kürze	in Kürze	37,06	21,59	488,24	313,37
jdm. auf den Sack gehen	auf den Sack	3,27	0,35	4,49	3,67
im wahrsten Sinn(e) des Wortes	im wahrsten Sinn(e)	8,36	21,47	191,24	134,17
zum Kotzen	zum Kotzen	0,36	0,58	9,32	8,91
jdm. in den Arsch kriechen	in den Arsch	1,45	0,69	13,73 (s. Tab. 6)	4,56
in den Hintern kriechen/treten/beißen	in den Hintern	2,91	0,81	9,02	9,67
grünes Licht geben/haben/erhalten	grünes Licht	1,09	14,43	39,43	203,99
Summe über alle 5058 Phraseologismen		126 108	65 698	97 534	99 220

Tabelle 7: Häufigkeit von Phraseologismen in verschiedenen Textsorten

häufiger im Web	**gleich häufig**	**häufiger in der Zeitung**
reden wie ein Wasserfall	es ist höchste Eisenbahn	in den sauren Apfel beißen
die Ohren steifhalten	bis über beide Ohren verliebt	aus dem Gröbsten heraus
jdm. knurrt der Magen	die Nacht um die Ohren schlagen	sein Geld zum Fenster hinauswerfen

Tabelle 8: Frequenzvergleich Web – Zeitung

vorkommen als in den Zeitungstexten. Das hat sich für 54% der Phraseologismen bestätigt. Die anderen Phraseologismen waren in Webtexten etwa gleich häufig (15%) oder seltener. Beispiele führt die Tabelle 8 auf.

Eine Betrachtung der in Webtexten häufigen Phraseologismen, ohne sie auf seltene einzuschränken, zeigt, dass im Web neben den an die mündliche Kommunikation angelehnten Texten oft schriftliche Korrespondenz festgehalten wird. Briefe werden hingegen in Zeitungstexten nur sehr gekürzt wiedergegeben. Auch im Projekt Gutenberg gibt es nur einen geringen Anteil an Briefen (aktuell 47 von mehr als 5000 Werken, die teilweise nur aus einem Brief bestehen), die sich zudem im Sprachgebrauch und Thema deutlich von Briefen im Web absetzen. Weiterhin sind die Phraseologismen aus dem Web durch juristische Wendungen geprägt. Die folgende Liste führt die Phraseologismen auf, deren relative Häufigkeit am stärksten über den Phraseologismen aus Zeitungstexten liegt.

3 Zur Gültigkeit des Zipfschen Gesetzes

3.1 Zipfsches Gesetz für Wörter

Das Zipfsche Gesetz (Zipf 1949) für Wörter sagt Folgendes aus: Sortiert man die Wörter (genauer: Wortformen) eines Korpus in eine Rangfolge entsprechend ihrer Häufigkeit, so ist das Produkt aus Rang und Häufigkeit eines Wortes näherungsweise konstant.

Zunächst soll die Gültigkeit des Zipfschen Gesetzes für das oben beschriebene Korpus aus 5058 Phraseologismen getestet werden. Dazu vergleichen wir die Resultate mit dem Zeitungskorpus. In den Abbildungen wird eine doppelt logarithmische Darstellung des Rang-Häufigkeits-Graphen gewählt, da hier die exakte Gültigkeit des Zipfschen Gesetzes einer Geraden mit Anstieg –1 entspricht.

Da das Zipfsche Gesetz stets nur näherungsweise erfüllt ist und größere Korpora in der Regel bessere Resultate liefern, wird hier das relativ kleine Korpus aus 5058 Phraseologismen mit einem Korpus aus einer vergleichbar großen Menge Zeitungstext (3000 zufällig ausgewählte Sätze) verglichen. Die stets vorhandene Abweichung bei den häufigsten Wörtern (d.h. Rang 1-20) ist bei den Phraseologismen offensichtlich etwas stärker ausge-

alle Rechte vorbehalten	um ein Wesentliches älter	blauer Montag
Mit freundlichen Grüßen verbleibe	Träume sind Schäume!	ein edler Tropfen
bis (auf) bald!	rechter Hand	Ich bitte Sie!
Grüß dich!	jmd. zu treuen Händen übergeben	Wie spät ist es?
sehr geehrte Damen und Herren	auf eigene Gefahr	Verlass dich drauf!
bis auf Widerruf gestattet	Du bist vielleicht gut!	sich in Unkosten stürzen
Was fällt dir ein!	außer Betracht bleiben	sich keinen Rat wissen
zum mindesten	gebe Gott dass	Stille Wasser sind tief
für etwas keinen Pfennig geben	jmd. die Haare vom Kopf fressen	ein blutiger Anfänger
Machs gut!	viel/kein Wesen um etwas machen	etwas auf seine Schulter nehmen
nur eine kurze Vorstellung geben	einen Bärenhunger haben	im Irrtum sein
etwas zu Schanden machen	linker Hand	Zu seiner Ehre muss ich sagen dass
zittern wie Espenlaub	sein Licht leuchten lassen	bei jdm. klickt es
jemands Tun und Treiben	übel dran sein	einen kleinen weg haben
etwas hat seine Reize	sein Kreuz auf sich nehmen	weißt du was?
beweg dich!	Stroh im Kopf haben	einen saufen
leben Sie wohl	Geld und Gut	sein Bündel schnüren
festen Fuß fassen	Ab geht die Post!	unter die Räuber fallen
meine Wenigkeit	dein Wunsch ist mir Befehl	(Ach) du liebe Zeit!
frei Haus	mit System	
von Rechts wegen	Grüß Gott!	
	sein Wesen treiben	

Tabelle 9: Für Webtexte spezifische Phraseologismen

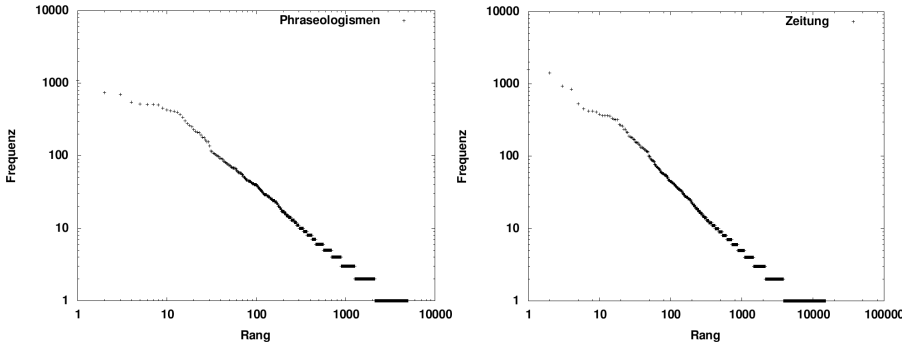

Abbildung 1: Zipfsches Gesetz für Phraseologismen (links) und vergleichbar viel Zeitungstext (rechts)

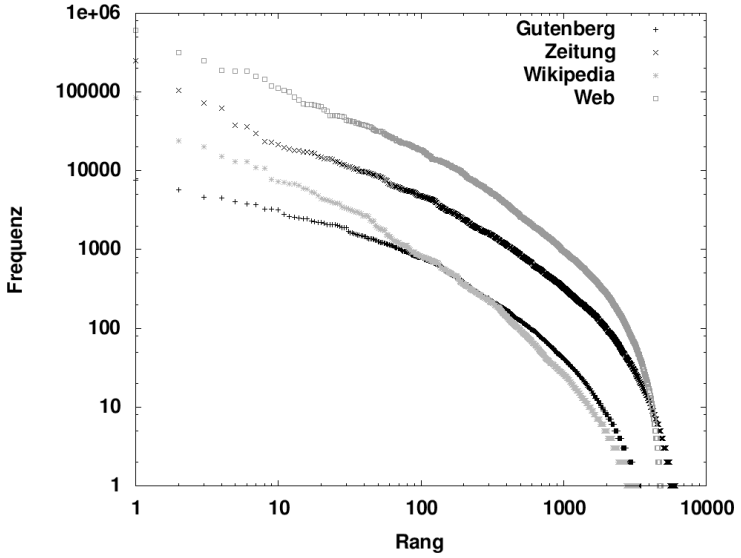

Abbildung 2: Zipfsches Gesetz für Phraseologismen für vier Korpora: Webtext, Zeitungstext, Wikipedia und Projekt Gutenberg (von oben)

prägt, im weiteren Verlauf scheint das Zipfsche Gesetz aber ähnlich gut erfüllt (vgl. dazu auch Colson in diesem Band).

Diese Abbildungen deuten auf einen sehr ähnlichen Zusammenhang zwischen Rang und Häufigkeit hin, leichte Abweichungen des Anstiegs von -1 lassen sich mit der Zipf-Mandelbrot-Verteilung (Mandelbrot 1953) erklären. Die sehr geringe Korpusgröße und die manuelle Auswahl der Phraseologismen für das Korpus setzt einer weiteren statistischen Auswertung allerdings deutliche Grenzen.

3.2 Rang und Häufigkeit für Phraseologismen

Die analoge Fragestellung für Phraseologismen vergleicht also Rang und Häufigkeit von Phraseologismen. Wie im vorigen Abschnitt werden tatsächlich die Häufigkeiten für die Suchformen verwendet, die eher den summierten Anzahlen der Formen der einzelnen Phraseologismen entsprechen. Abbildung 2 zeigt den Rang-Häufigkeits-Vergleich für die Liste von 5058 Phraseologismen für die vier Vergleichskorpora. Im Vergleich mit Abbildung 1 auf der vorherigen Seite ähneln die Linien hier viel weniger einer Geraden: Im letzten Drittel ist ein starkes Abfallen zu beobachten. Außerdem fällt folgendes auf: Während drei der vier Linien nahezu parallel verlaufen, hat die vierte Linie offensichtlich einen stärkeren Anstieg. Erklärungsversuche sollen im Folgenden gegeben werden.

POS-Tag-Muster	Anzahl	Beispiel
PIS APPR ART NN VVFIN	124	jdm. um den Hals fallen
APPR ART NN VVFIN	120	aus der Haut fahren
ART ADJA NN	100	der harte Kern
ART NN VVFIN	67	den Chef markieren
ADV APPR ART NN VVFIN	60	nur auf dem Papier stehen
ART ADJA NN VAFIN	57	eine scharfe Zunge haben
PIS APPRART NN VVFIN	48	jdn. zur Kasse bitten
APPRART NN VVFIN	47	am Drücker sitzen
ADV APPR NN VVFIN	41	irgendwie zu Werke gehen
ART ADJA NN VVFIN	40	den starken Mann spielen

Tabelle 10: Häufigste POS-Tag-Muster bei den untersuchten Phraseologismen

Zunächst zum Verhalten der Rang-Frequenz-Kurve für hohe Ränge. Eine Konsequenz der Gültigkeit des Zipfschen Gesetzes in seiner exakten Form ist, dass es sehr viele Objekte (Wörter oder Phraseologismen) mit extrem niedriger Frequenz gibt. In Abbildung 1 auf Seite 48 erkennt man, dass tatsächlich jeweils mehr als die Hälfte aller Wortformen Frequenz 1 haben. Es ist auf Grund der folgenden prinzipiellen Überlegungen zweifelhaft, dass dies auch für Phraseologismen zutreffen könnte: Ein Phraseologismus ist nach Definition eine Wortgruppe, die sich unter anderem durch relativ feste Form und Wiederholung auszeichnet. Ist das untersuchte Korpus ausreichend groß, sollte sich diese Wiederholung auch im Korpus zeigen und deshalb sollten nur vergleichsweise wenige extrem seltene Phraseologismen auftreten. Der steilere Anstieg der Rang-Frequenz-Kurve für das Wikipedia-Korpus muss sich aus einem noch zu erklärenden Unterschied der Textsorten ergeben.

4 Struktur und Häufigkeit

In diesem Abschnitt wollen wir uns für häufige syntaktische Strukturen von Phraseologismen interessieren. Mittels Part-of-Speech-Tagging (kurz POS-Tagging, siehe Brants 2000) lassen sich die Wörter eines Phraseologismus automatisch mit einem Wortart-Kürzel (POS-Tag) versehen. Wir interpretieren Phraseologismen mit gleichen POS-Tag-Folgen als strukturell gleich. Diese Interpretation ist häufig korrekt, aber nicht immer: Einmal ist POS-Tagging nicht völlig fehlerfrei (insbesondere bei den meist aus nur wenigen Wörtern bestehenden Phraseologismen), andererseits kann mittels POS-Tags beispielsweise nicht zwischen Subjekt und Objekt unterschieden werden.

Tabelle 10 zeigt die häufigsten POS-Tag-Muster mit Beispielen. Da die meisten Phraseologismen im Text nicht in ihrer Wörterbuchform auftreten, sollen auch noch die POS-Tag-Muster bei den Suchformen angegeben werden (vgl. Tabelle 11 auf der nächsten Seite).

POS-Tag-Muster	Anzahl
APPR ART NN	528
ART ADJA NN	339
APPRART NN	249
APPR PPOSAT NN	156
APPR NN	145
APPR ADJA NN	131
ADJA NN	128
ART NN APPR	118
ART NN APPR ART NN	111
APPR ART ADJA NN	107

Tabelle 11: Häufigste POS-Tag-Muster bei den Suchformen

Die komplette Liste der häufigen POS-Tag-Muster kann verwendet werden, um Antworten auf die im folgenden Abschnitt gestellte Frage nach Möglichkeiten der automatischen Auffindung neuer Phraseologismen zu finden.

5 Sind neue Phraseologismen automatisch auffindbar?

Die Aufgabenstellung, das Web oder andere Texte nach neuen, bisher nicht dokumentierten Phraseologismen zu durchsuchen, ist auf den ersten Blick faszinierend, aber alles andere als einfach. Die Entscheidung, ob es sich bei neu gefundenen Wortgruppen tatsächlich um Phraseologismen handelt bzw. wann eine Wortverbindung als lexikalisiert und somit als sprachliches Zeichen gilt, wird auch von Experten nicht einheitlich getroffen. Diese Unschärfe stellt jedoch keine zusätzliche Schwierigkeit dar, da alle momentan verfügbaren automatischen Verfahren nur Listen von Kandidaten für Phraseologismen erzeugen werden. Solche Listen enthalten auch viele abzulehnende Kandidaten, und die Qualität einer solchen Liste dürfte von verschiedenen Experten ähnlich bewertet werden.

Während in Quasthoff/Schmidt (im Druck) ausführlich die strukturelle Festigkeit von Phraseologismen untersucht und das Kookkurrenzverhalten der beteiligten Wörter beschrieben wurde, ergänzen die hier dargestellten Ergebnisse die algorithmisch nutzbare Beschreibung von Phraseologismen.

Kombiniert man die Ergebnisse, so erscheint ein Phraseologismus automatisch identifizierbar, wenn er die folgenden Eigenschaften besitzt:

- Er enthält mindestens 2 wichtige Wörter im Sinne von Abschnitt 2.1.
- Im zur Verfügung stehenden Korpus kommt der Phraseologismus mindestens zehnmal vor.

Weitere Eigenschaften erleichtern das Auffinden:

– Geringe Variabilität des Phraseologismus;

– Typisches, bekanntes POS-Tag-Muster;

– Vorkommen von Wörtern, die typischerweise auch in anderen Phraseologismen auftreten.

Mit diesen Kriterien erscheint die automatisierte Suche nach unbekannten Phraseologismen in nächster Zeit durchführbar.

Literaturverzeichnis

Brants, T. (2000): „TnT – A Statistical Part-of-Speech Tagger". In: *Proceedings of the sixth conference on Applied natural language processing.* Morristown; 224-231.

Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N. R. (Hrsg.) (2007): *Phraseologie/Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin, New York.

Dobrovol'skij, D. (1997): *Idiome im mentalen Lexikon. Ziele und Methoden der kognitivbasierten Phraseologieforschung.* Trier.

DUDEN, Bd. 4. Die Grammatik (2006). Hrsg. von der Dudenredaktion. 7. völlig neu erarbeitete und erweiterte Aufl. Mannheim.

Ďurčo, P. (Hrsg.) (im Druck): *5. Kolloquium zur Lexikographie und Wörterbuchforschung. The Fifth International Colloquium on Lexicography. Feste Wortverbindungen und Lexikographie/Fixed word combinations and Lexicography. Bratislava, 20.- 21. Oktober 2008.* Tübingen.

Fix, U./Lerchner, G./Schröder, M./Wellmann, H. (Hrsg.) (2005): *Zwischen Lexikon und Text – lexikalische, stilistische und textlinguistische Aspekte.* Leipzig.

Götz, D./Haensch, G./Wellmann, H. (Hrsg.) (1997): *Langenscheidts Großwörterbuch Deutsch als Fremdsprache.* Berlin.

Griesbach, H./Schulz, D. (2000): *1000 deutsche Redensarten. Mit Erklärungen und Anwendungsbeispielen.* Berlin

Häcki Buhofer, A./Burger, H./Gautier, L. (Hrsg.) (2001): *Phraseologiae Amor. Aspekte europäischer Phraseologie.* Baltmannsweiler.

Hallsteinsdóttir, E. (2005): „Vom Wörterbuch zum Text zum Lexikon". In: Fix et al. (Hrsg.); 325-337.

Hallsteinsdóttir, E. (2007): „Wörtliche, freie und phraseologische Bedeutung. Eine korpusbasierte Untersuchung des Vorkommens von freien und phraseologischen Lesarten bei deutschen Idiomen". In: Kržišnik/Eismann (Hrsg.); 107-121.

Hallsteinsdóttir, E./Šajánková, M./Quasthoff, U. (2006): „Phraseologisches Optimum für Deutsch als Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen", in: *Linguistik-online 27*; 119-138 (*www.linguistik-online.de/27_06/hallsteinsdottir_et_al.pdf*).

Hessky, R./Ettinger, S. (1997): *Deutsche Redewendungen. Ein Wörter- und Übungsbuch für Fortgeschrittene*. Tübingen.

Heyer, G./Quasthoff, U./Wittig, Th. (2006): *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*. Bochum.

Kempcke, G. (2000): *Wörterbuch Deutsch als Fremdsprache*. Berlin, New York.

Kržišnik, E./Eismann, W. (Hrsg.) (2007): *Phraseologie in der Sprachwissenschaft und anderen Disziplinen*. Lubljana.
Lerchner, G. (2005): „Wie werden Lexeme zu Schlag-, Mode- oder Leitwörtern? Zu lexikalischen Ergebnissen textgeleiteter semiotischer Prozesse". In: Fix et al. (Hrsg.); 57-63.
Mandelbrot, Benoit B. (1953): „An information theory of the statistical structure of language". In: Jackson, W. (Ed.): Communication Theory. New York, Academic Press; (503-512).
Projekt Gutenberg: *http://gutenberg.spiegel.de/*
Quasthoff, U./Schmidt, F. (im Druck): „Die korpusbasierte Identifikation fester Wortverbindungen". In: Ďurčo (Hrsg.) (im Druck).
Stein, S. (2007): „Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive." In: Burger et al. (Hrsg.); 220-236.
Wotjak, B. (2001): „Phraseologismen im neuen Lernerwörterbuch – Aspekte der Phraseologiedarstellung im de Gruyter-Wörterbuch Deutsch als Fremdsprache". In: Häcki Buhofer et al. (Hrsg.); 263-279.
Zipf, G. K. (1949): *Human Behavior and the Principle of Least Effort.* Cambridge, MA.

Uwe Quasthoff
Universität Leipzig
Institut für Informatik
Johannisgasse 26
04103 Leipzig
Deutschland
quasthoff@informatik.uni-leipzig.de

Fabian Schmidt
Universität Leipzig
Institut für Informatik
Johannisgasse 26
04103 Leipzig
Deutschland
fschmidt@informatik.uni-leipzig.de

Erla Hallsteinsdóttir
Syddansk Universitet Odense
Institut for Sprog og Kommunikation
Campusvej 55
5230 Odense
Danmark
erla@language.sdu.dk

Methoden und Ergebnisse einer korpusbasierten Untersuchung zur Vorkommenshäufigkeit bulgarischer Sprichwörter in zeitgenössischen Zeitungstexten

Hrisztalina Hrisztova-Gotthardt

The aim of the present paper is the presentation of the methods and the results of a pilot research dealing with the frequency of the occurrence of Bulgarian proverbs in contemporary newspaper texts. In the first part, there the need for corpus-based research is elaborated on, namely the one with the aim of evaluating the frequency of usage of Bulgarian proverbs in written language. The second part deals with the question of the means of determining the most familiar and frequently used proverbs in a particular language. The presentation of the methods and the results of the corpus-based research makes up the core of the last chapter. Aspects discussed are: the compilation of a corpus, the choice of a method used for the searching for proverbs, as well as the representativeness of the results obtained. Further steps which are due to be introduced in the scope of future research with the aim of achieving more precise and reliable results are also presented.

1 Einleitung

Der vorliegende Beitrag setzt sich zum Ziel, die Methoden und die Ergebnisse einer Pilotstudie zur Verwendung und Vorkommenshäufigkeit bulgarischer Sprichwörter in zeitgenössischen Zeitungstexten darzustellen. Die Notwendigkeit einer solchen korpusbasierten Untersuchung wird dadurch begründet, dass bis zum jetzigen Zeitpunkt keinerlei Daten über die aktuelle parömiologische Situation, d. h. über den momentanen Gebrauch bulgarischer Sprichwörter weder in der geschriebenen noch in der gesprochenen Sprache vorliegen. Das zeichnet sich besonders deutlich in den modernen parömiographischen Werken ab, die ausschließlich auf älteren Sammlungen und Wörterbüchern aufbauen. Um das Gesagte zu illustrieren, gewährt der erste Teil des Beitrages am Beispiel einer neu erschienenen Sammlung (Stojkova 2007) einen kurzen Einblick in die moderne bulgarische Parömiographie mit ihren Stärken und Schwächen.

Daran anknüpfend setzt sich der zweite Teil dieser Arbeit mit der Frage auseinander, welche Verfahren zur Ermittlung der bekanntesten und gebräuchlichsten Sprichwörter einer Sprache dem Parömiologen zur Verfügung stehen. Es wird davon ausgegangen, dass die Analyse eines – wenn auch seiner Größe und Textvielfalt nach relativ begrenzten – Korpus geschriebener Texte erste wichtige Rückschlüsse auf die aktive Kenntnis und typische Verwendung der Parömien[1] zulassen kann.

Die Präsentation der Methoden und Ergebnisse der korpusbasierten Untersuchung bildet den Gegenstand des letzten Abschnittes. Eingehend erörtert werden Fragen bezüglich der Zusammensetzung des Korpus, der Auswahl der anzuwendenden Verfahren bei der Suche nach Sprichwörtern sowie der Repräsentativität der Ergebnisse.

2 Zum aktuellen Stand der modernen bulgarischen Parömiographie

Die Behauptung, dass die einschlägigen Sammlungen oft viele veraltete und nicht mehr gebräuchliche Parömien aufzeichnen und dementsprechend als „Datenfriedhöfe" bezeichnet werden können (vgl. Baur/Chlosta 1996: 92), trifft bedauerlicherweise auch auf die bulgarischen Werke zu. Erwähnt sei an dieser Stelle die 2007 erschienene Sammlung bulgarischer Sprichwörter und Redensarten von Stefana Stojkova. Nach einem langjährigen Stillstand ist dies die erste phraseologische Sammlung, die auf den bulgarischen Buchmarkt gekommen ist und deren Umfang deutlich über jenen der phraseologischen Schulwörterbücher hinausgeht. Mit über 6000 Sprichwörtern und Redensarten stellt das Werk eine besonders reiche Dokumentation dar. Positiv zu erwähnen sind außerdem die von der Autorin angeführte thematische Klassifikation der Texte, die (wenn auch nur sporadisch angegebenen) Bedeutungserläuterungen, Verweise auf synonyme und antonyme Parömien sowie die ausführlichen Informationen über die Herkunft zahlreicher bulgarischer Sprichwörter. Leider beinhaltet diese Sammlung keinerlei Informationen über den aktuellen Gebrauch von Sprichwörtern in der bulgarischen Gegenwartssprache. Laut Quellenverzeichnis (vgl. Stojkova 2007: 19) entstammen alle kodifizierten Einheiten älteren Sammlungen, deren Erscheinung einige Jahrzehnte zurückliegt. Folglich spiegeln die ausgewählten Parömien keineswegs den aktuellen Sprachgebrauch wider.

In diesem Sinne erweist es sich als dringend nötig, erste Schritte in Richtung Ermittlung der tatsächlichen passiven Kenntnis und der aktiven Verwendung bulgarischer Sprichwörter einzuleiten. Die ermittelten Ergebnisse können künftig bei der Erstellung von Sprichwörterbüchern eine große Hilfe leisten.

---

1 Im Weiteren werden die aus dem Lateinischen (*Proverbium*) und aus dem Griechischen (*Parömie*) stammenden Bezeichnungen als Synonyme für den deutschen Begriff *Sprichwort* verwendet.

3 Verfahren zur Ermittlung des Bekanntheitsgrades und der Verwendungshäufigkeit von Sprichwörtern

In seinem Aufsatz über „Bekanntheit, Häufigkeit und lexikographische Erfassung von Sprichwörtern" weist Ďurčo darauf hin, dass man „die Ergebnisse der empirischen Untersuchungen zur Bekanntheit der Sprichwörter einer bestimmten Sprache und ihre lexikographische Erfassung auch noch mit ihrem realen Vorkommen in Texten vergleichen muss, um Informationen über die wirkliche parömiologische Situation zu bekommen" (Ďurčo 2001: 101). Als eine wichtige Methode der empirischen Untersuchung zur Bekanntheit von Sprichwörtern nennt er das „Teiltextpräsentation-Verfahren"[2], das auf die Ermittlung der bekanntesten Sprichwörter einer Sprachgemeinschaft sowie auf die Erstellung eines Sprichwort-Minimums für die jeweilige Sprache abzielt. Zwar liefern solche Untersuchungen Informationen über den Bekanntheitsgrad der Parömien sowie über ihre gebräuchliche Form, sie sagen aber eher wenig über ihre aktive Verwendung und ihre Vorkommenshäufigkeit in der Sprache aus. Zur Beantwortung letzterer Frage kann die Methode der Korpusanalyse angewendet werden, bei der umfangreiche Korpora der geschriebenen Sprache auf die Gebrauchsfrequenz von Sprichwörtern geprüft und dabei das reale Vorkommen, die konkreten Kontexteinbettungen, Verwendungssituationen und Transformationsfähigkeiten von Sprichwörtern untersucht werden (vgl. Čermák 2003; Ďurčo 2006).

Die hier präsentierte Pilotstudie verfolgt wesentlich bescheidenere Ziele: Es sollen lediglich einige erste Ergebnisse über die Vorkommenshäufigkeit und demnach über die aktive Verwendung bulgarischer Sprichwörter in Zeitungstexten der letzten achteinhalb Jahre dargestellt werden. Die Studie strebt keineswegs die Ermittlung der „gebräuchlichsten" bulgarischen Parömien an, sondern versteht sich vielmehr als Ausgangspunkt für weitere empirische und korpusbasierte Forschung.

4 Das Korpus und die Methoden der Untersuchung

4.1 Das Korpus

Die Auswahl der im Korpus vorhandenen Texte richtete sich nach allgemein verbreiteten korpuslinguistischen Kriterien (vgl. dazu Scherer 2006: 53ff.). Da im konkreten Fall der Gebrauch bulgarischer Sprichwörter in zeitgenössischen Zeitungstexten untersucht werden soll, gestalteten sich die Kriterien wie folgt:[3]

– Sprachauswahl: Beim Aufbau des Korpus wurden lediglich bulgarische Texte berücksichtigt, es handelt sich also um ein monolinguales Korpus.

---

2 Mehr zu diesem Verfahren in Permjakov (1988), Grzybek (1991), Baur/Chlosta (1996), Tóthné Litovkina (1996) etc.

3 Die hier aufgelisteten Kriterien wurden in Anlehnung an Lemnitzer/Zinsmeister (2006) und Scherer (2006) formuliert.

- Medium: Die gespeicherten Primärdaten sind in der geschriebenen Sprache entstanden, demnach ist von einem Korpus der geschriebenen Sprache die Rede.
- Größe: Mit seinen rund 58,5 Millionen Textwörtern aus 240 886 Artikeln weist das Korpus eine mittlere Größe auf.
- Persistenz: An der in einem bestimmten Zeitraum gesammelten und für die weitere Verarbeitung gespeicherten Textmenge wurden später keine Veränderungen vorgenommen, damit erfüllt das Korpus das Kriterium der Beständigkeit.
- Sprachbezug: Das erstellte Korpus erhebt nicht den Anspruch, repräsentativ für das Bulgarische in seiner Gesamtheit zu sein. Es sollte vielmehr dazu dienen, einen Teil einer bestimmten Varietät der Sprache,[4] und zwar der Zeitungssprache in einem begrenzten Zeitraum zu erforschen. Demgemäß ist es als ein Spezialkorpus zu betrachten, in dem Artikel aus der bulgarischen Tageszeitung Стандарт [Standart] gespeichert sind, die zwischen dem 4. 1. 2000 und dem 3. 8. 2008 in der elektronischen Ausgabe von Стандарт erschienen sind.[5]
- Annotation: Die Daten wurden als reine Textdateien gespeichert, es fehlt eine Annotation, d. h. es wurden keine linguistischen Informationen hinzugefügt. Das Korpus verfügt jedoch über Metadaten, die extrahiert und auf eine bestimmte Art und Weise markiert wurden, etwa Informationen zur Identifikationsnummer des jeweiligen Artikels, zur Rubrik, in der er erschienen ist, sowie zu seinem Verfasser und Erscheinungsdatum.

4.2 Methoden der Untersuchung

Die Auswahl der geeigneten Methoden zur Suche nach Sprichwörtern bereitete aus folgenden Gründen gewisse Schwierigkeiten: Zum einen stand keine auf ein brauchbares Minimum reduzierte, empirisch erprobte Liste bekannter bulgarischer Sprichwörter zur Verfügung, sondern nur eine Liste mit 2301 Sprichwörtern. Sie entstammten zwei phraseologischen Sammlungen, die in der Anfangsphase der Untersuchung immer noch als die „neuesten" galten: Grigorov/Kazarov (1986) und Vlahov (1996). Zum anderen war das Korpus, an dem die Forschung durchgeführt wurde, nur tokenisiert und nicht lemmatisiert.[6] Demzufolge konnte man lediglich nach konkreten Wortformen suchen. Damit wurde die

4 Laut Scherer ist eine *Varietät* eine bestimmte Ausprägung der Sprache, die durch außersprachliche Faktoren wie Zeit, Raum, Sprechergruppe oder Kommunikationssituation geprägt wird (Scherer 2006: 4).

5 Die elektronische Ausgabe von Стандарт ist zugänglich unter: *http://paper.standartnews.com/bg/*.

6 Der Terminus *Tokenisierung* wird in der Computer- und Korpuslinguistik verwendet und bezeichnet die Segmentierung eines Textes in kleinste Einheiten der Wortebene, auch *Token* genannt. Dabei werden nicht nur Wörter im gängigen Sinne, sondern auch Zahlen, Satzzeichen, Klammern, Anführungsstriche und andere Symbole als Tokens identifiziert, die (meist) durch Leerzeichen voneinander abgegrenzt sind (vgl. Carstensen 2004: 408f.; Lemnitzer/Zinsmeister 2006: 64f.). Unter der *Lemmatisierung* wird eine Annotation auf der

Möglichkeit ausgeschlossen, auf korpuslinguistische Methoden zuzugreifen, die im Laufe früherer Untersuchungen bereits angewendet und in sprachwissenschaftlichen Studien dargestellt wurden (vgl. etwa Čermák 2003, 2006; Ďurčo 2006 u. v. m.). Es musste also auf statistische Herangehensweisen zur Ermittlung der absoluten und relativen Häufigkeiten und Kookkurenzen von Sprichwortkonstituenten wie MI-score und T-score verzichtet werden (vgl. u.a. Čermák 2006), da sie eine enorm lange Liste von Treffern ergeben hätten, die alle manuell hätten durchgesehen werden müssen. Ebenfalls erwies sich die Option der Suche nach konkreten Sprichwortmodellen mithilfe von regulären Ausdrücken als unausführbar (zu vergleichbaren Ergebnissen siehe Ďurčo 2006 9f.), da sich nicht alle 2301 Sprichwörter einem bestimmten strukturellen Muster zuordnen ließen.

Es sei an dieser Stelle erwähnt, dass das Fehlen einer vorgegebenen Liste bulgarischer Sprichwörter sogar als vorteilhaft angesehen werden kann: Es wurde keine subjektive Selektion vorgenommen; die ausgewerteten Daten spiegelten die authentische parömiographische Situation in Bulgarien wider.

In Anbetracht der obigen Ausführungen und nach reiflicher Überlegung wurde folgende Vorgehensweise gewählt:

– In einem ersten Schritt wurden alle in Form von Textdateien heruntergeladenen Artikel in eine einzige große Datei übertragen und gespeichert. Danach wurde die Tokenisierung durchgeführt. Üblicherweise erfolgt die Tokenisierung automatisch auf Grund der vorgegebenen Standardregeln.[7] Es ist allerdings nicht möglich, bei der Formulierung der Standardregeln alle Sonderfälle in allen Sprachen zu berücksichtigen. Aus diesem Grund muss die Standard-Spezifikation je nach Sprache und Korpus ergänzt werden. Für das Bulgarische wurden zwei zusätzliche Regeln aufgestellt, die besagen, dass die durch Bindestrich verbundenen Wörter (по-добър, най-сладък) und von einem Leerzeichen getrennten Zahlen (*40 000*) nicht weiter zerlegt werden dürfen.[8] Alle Interpunktionszeichen wurden gelöscht, um der schnelleren Suche willen wurden alle Groß- in Kleinbuchstaben umgewandelt. Was im Endergebnis übrig blieb, war eine Tokenliste, in der nur die ursprüngliche Reihenfolge der Token beibehalten worden war. Ebenfalls tokenisiert und in einer anderen Textdatei gespeichert wurden die 2301 bulgarischen Sprichwörter.

Wortebene verstanden, bei der die Tokens morphologisch analysiert und auf ihre Grundform zurückgeführt werden, d. h. Worteinheiten, die sich nur in ihren Flexionsmerkmalen unterscheiden, werden unter dem Begriff *Lemma* zusammengefasst (s. Haß 2005: 75ff.; Scherer 2006: 33).

Bedauerlicherweise wurde das sich im ständigen Aufbau befindende lemmatisierte Bulgarische Nationalkorpus für wissenschaftliche Zwecke noch nicht freigegeben.

7 Eine internationalisierte Standard-Spezifikation für die automatische Zerlegung von Texten an den Wortgrenzen bietet das Unicode Consortium. Die von ihm formulierten Regeln (*Unicode Text Segmentation*) können unter *http://unicode.org/reports/tr29/* abgerufen werden.

8 Einige wenige bulgarische Sprichwörter enthalten Zahlen, die durch Ziffern dargestellt werden, wie z. B.: Ако да носеха всички луди звънци, желязото би станало 100 гроша ока̀та (Trügen alle Irren Glocken, dann kostete ein Kilo Eisen 100 Taler).

Artikel-Token ID	Artikel-Token	Frequenz in Artikeln
304 638	на	2 768 949
203 138	и	1 738 888
103 991	в	1 592 627
142 639	да	1 293 641
184 540	за	1 172 172
362 382	от	1 168 153
480 362	се	1 101 692
168 219	е	1 097 626
470 498	с	715 346
587 265	ще	559 953

Tabelle 1: Token-Frequenz in Zeitungsartikeln

Sprichwort-Token ID	Sprichwort-Token	Frequenz in Sprichwörtern
1978	не	922
2993	се	585
1075	и	399
638	да	382
1858	на	370
853	е	343
3028	си	250
2198	от	231
2915	с	201
1378	който	181

Tabelle 2: Token-Frequenz in den ausgewerteten Sprichwörtern

- In einem zweiten Schritt wurde die absolute Häufigkeit der Tokens für die beiden Dateien ermittelt, d. h. es wurde festgelegt, wie oft ein Token aus der Artikel-Datei im Korpus vorkommt, bzw. wie viele Male ein Token in der Sprichwörter-Datei als Sprichwortkonstituente erscheint. Dabei zeigte unter den insgesamt 3855 Sprichwort-Tokens die Partikel не (*nicht, nein*) mit 922 Treffern die höchste Frequenz auf, unter den Artikel-Token war dies die Präposition на (*auf, von*) mit 2 768 949 Treffern. Verallgemeinernd lässt sich sagen, dass die frequentesten Tokens sowohl in der Artikel- als auch in der Sprichwörter-Datei Synsemantika waren, wie Tabellen 1 und 2 zeigen.
- In einem dritten Schritt wurde ermittelt, mit welcher Häufigkeit die Sprichwort-Tokens im Korpus vorkommen. Da nur das konkrete Vorkommen einer bestimmten Worteinheit im (nicht lemmatisierten) Korpus berücksichtigt wurde und nicht

Sprichwort-Token ID	Sprichtwort-Token	Frequenz in Artikeln
1858	на	2 768 949
1075	и	1 738 888
245	в	1 592 627
638	да	1 293 641
908	за	1 172 172
2198	от	1 168 153
2993	се	1 101 692
853	е	1 097 626
2915	с	715 346
3805	ще	559 953

Tabelle 3: Frequenz von Sprichwort-Tokens in den ausgewerteten Zeitungsartikeln

alle Flexionsformen des jeweiligen Lexems, musste hier ein gewisser Grad an Informationsverlust in Kauf genommen werden. Es hat sich herausgestellt, dass die frequentesten Wörter wiederum vorwiegend Synsemantika waren, die nicht als bedeutungstragende Konstituenten eines Sprichwortes und daher als nicht signifikant für die Suche einzuschätzen sind (Tabelle 3).

Auf Grund der obigen Überlegungen und Ergebnisse wurden die Sprichwort-Tokens mit einem absoluten Häufigkeitswert von 50 oder mehr von der Sprichwort-Tokenliste gestrichen, d. h. aus der Sprichwort-Tokendatei gelöscht.

– Als Nächstes wurde geprüft, in welchen Artikeln eines oder mehrere der restlichen Sprichwort-Tokens vorkommen. Jene Artikel, in denen kein einziges Sprichwort-Token aufgefunden wurde, wurden aus der Untersuchung ausgeschlossen.

– Im Weiteren wurde der Frage nachgegangen, aus welchen Sprichwörtern die Tokens stammen, die in den nach der Selektion übriggebliebenen Artikeln zu finden waren. Der Algorithmus sollte also folgende Fragen beantworten: Welches Token tauchte in welchem Artikel und an welcher Stelle auf? Welchem Sprichwort entstammt es und welche Position nimmt es im Sprichwort ein?

  Dabei fand ein „Token-Matching“ statt, d. h. es wurde überprüft, wie viele Tokens rechts respektive links von dem sogenannten „Treffer“ eine Übereinstimmung mit den im jeweiligen Sprichwort vorkommenden Wörtern zeigen (vgl. Tabelle 4 auf der nächsten Seite).

  Bedauerlicherweise ist auch in dieser Phase der Untersuchung mit einem Informationsverlust zu rechnen: Werden bei den Einbettungen von Sprichwörtern in einen Kontext ihre Konstituenten mit anderen Flexionsmerkmalen versehen bzw. werden zwischen die Komponenten andere Wörter eingefügt, so sinkt die Wahrscheinlichkeit enorm, dass diese Sprichwörter vom Computer tatsächlich als solche erkannt werden.

← седяло → . . . които пише да би мирно седяло не би чудо видяло първият случай е през . . .
← седяло → да би мирно седяло не би чудо видяло

	| ← → |	
. . . които пише	да би мирно седяло не би чудо видяло	първият случай е през. . .
	да би мирно седяло не би чудо видяло	

Tabelle 4: Token-Matching

– Abschließend musste die anhand des Korpus vom Computer zusammengestellte Liste mit „Sprichwortkandidaten“ manuell nach solchen durchsucht werden.

5 Ergebnisse der Untersuchung

Wie bereits erwähnt, musste man bei der Anwendung der oben geschilderten Suchmethode mehrmals einen gewissen Informationsverlust in Kauf nehmen. Er war in erster Linie auf die Tatsache zurückzuführen, dass zum Zwecke der Untersuchung lediglich ein tokenisiertes und kein lemmatisiertes Korpus zur Verfügung stand. Dennoch konnten die ersten Ergebnisse zur Verwendung und Vorkommenshäufigkeit bulgarischer Sprichwörter in zeitgenössischen Zeitungstexten ermittelt werden, die sich wie folgt zusammenfassen lassen:

– Von den insgesamt 2301 Sprichwörtern, die auf ihre Verwendung und Frequenz geprüft wurden, kamen 225 im Korpus vor.

– Die manuelle Überprüfung der Trefferliste hat ergeben, dass sich außer den aufgelisteten Sprichwörtern (etwa 72,9% aller Treffer) im Korpus noch zahlreiche quantitative und qualitative Varianten (etwa 15,7%) und Modifikationen (Antisprichwörter) (etwa 11,4%) finden (vgl. Tabelle 5 auf der nächsten Seite).

– Die Top-10-Liste der bulgarischen Sprichwörter (ohne ihre Varianten und Modifikationen) gestaltete sich folgenderweise (vgl. Tabelle 6 auf der nächsten Seite).[9]

9 Sechs der zehn weiter oben aufgelisteten bulgarischen Sprichwörter haben totale bzw. beinahe totale Äquivalente im Deutschen: Нищо ново под слънцето *([Es gibt] nichts Neues unter der Sonne)*, По-добре късно, отколкото никога *(Besser spät als nie)*, Съединението прави силата *(Einigkeit macht stark)*, Глас народен, глас Божи *(des Volkes Stimme, die Stimme Gottes)*, Апетитът идва с яденето *(der Appetit kommt beim Essen)* und Кръвта вода не става *(Blut ist dicker als Wasser)*. Zwei Proverbien verfügen über partielle deutsche Äquivalente: Всяко зло за добро *(Auf Regen folgt Sonnenschein)* und Две дини под една

Sprichwort-Lemma	Exakte Treffer	Variante	Modifikation	Treffer insgesamt
[...]				
Апетитът идва с яденето.	24	1	2	27
Всеки дърпа чергата към себе си.	5	8	14	27
Кръвта вода не става.	23		3	26
Никой не е пророк в отечеството си.		25	1	26
След дъжд качулка.	25			25
Каквото било било.	16	8		24
Кучетата си лаят, керванът си върви.	6	13	5	24
Бедността не е порок.	6		15	21
Гайда къща не храни.		15	6	21
Рибата се вмирисва от главата.	5	8	6	19
[...]				

Tabelle 5: Zahl und Art der Treffer

Sprichwort-Lemma	Treffer
Нищо ново под слънцето.	1029
По-добре късно отколкото никога.	58
Всяко зло за добро.	48
Съединението прави силата.	45
Две дини под една мишница не се носят.	37
Глас народен глас божи.	36
Всяко чудо за три дни.	34
Апетитът идва с яденето.	27
Всеки дърпа чергата към себе си.	27
Кръвта вода не става.	26

Tabelle 6: Bulgarische Sprichwörter: Top-10-Liste

6 Ausblick

Mit dieser Pilotstudie ist die Untersuchung zur Ermittlung der Verwendung und Frequenz bulgarischer Sprichwörter in der geschriebenen Sprache keineswegs abgeschlossen. Um möglichst präzise und detaillierte Ergebnisse zu erzielen, müssen folgende wichtige Schritte eingeleitet werden:

мишница не се носят *(Man soll nicht zwei Hasen auf einmal jagen)*. Für die restlichen zwei Sprichwörter gibt es im Deutschen weder absolute noch partielle Äquivalente: Всяко чудо за три дни (Jedes Wunder dauert [höchstens] drei Tage lang) und Всеки дърпа чергата към себе си (Jeder zieht die Decke an sich/auf seine Seite).

– Erweiterung des Korpus um weitere Texte anderer Gattungen: Es steht bereits eine Sammlung von Texten der schönen Literatur bereit, geplant ist außerdem die Einbeziehung von wissenschaftlichen Texten, Internetforen, -blogs etc.

– Verfeinern des Suchalgorithmus: Die gezielte Suche würde sich noch effizienter gestalten, wenn auch Angaben über das Maß der Überstimmung der Sprichwörter- und Artikeltokens (in Prozent) vorliegen würden. Dann kämen auch die relativ hohen und nicht nur die hundertprozentigen Übereinstimmungen als Sprichwörterkonstituenten-Kandidaten in die engere Wahl. Ferner sollte der Suchalgorithmus imstande sein, jene Sprichwörter zu finden, in denen bestimmte Konstituenten substituiert oder neue Wörter eingefügt wurden.

Wie dem Gesagten zu entnehmen ist, stellt die hier präsentierte Pilotuntersuchung lediglich die Anfangsphase eines Projektes dar, das weiterer Schritte bedarf und im Endeffekt darauf abzielt, zuverlässige Informationen über den aktiven Gebrauch bulgarischer Proverbien zu liefern, die nicht nur in der Linguistik, Parömiologie und Parömiographie, sondern auch im Fremdsprachenunterricht ihre Anwendung finden können.

Literaturverzeichnis

Baur, R./Chlosta, Ch. (1996): „Sprichwörter: ein Problem für Fremdsprachenlehrer wie -lerner?", in: *Deutsch als Fremdsprache. Zeitschrift zur Theorie und Praxis des Deutschunterrichts für Ausländer, 33/2*; 91-102.

Burger, H./Häcki Buhofer, A. (Hrsg.) (2006): *Phraseology in Motion I*. Baltmannsweiler.

Burger, H./Häcki Buhofer, A./Gréciano, G. (Hrsg.) (2003): *Flut von Texten – Vielfalt der Kulturen.* Baltmannsweiler.

Carstensen, K. (Hrsg.) (2001): *Computerlinguistik und Sprachtechnologie: eine Einführung.* München, ²2004.

Čermák, F. (2003): „Paremiological Minimum of Czech: The Corpus Evidence". In: Burger et al. (Hrsg.) (2003); 15-31.

Čermák, F. (2006): „Statistical Methods for Searching Idioms in Text Corpora". In: Burger/Häcki Buhofer (Hrsg.) (2006); 33-42.

Ďurčo, P. (2001): „Bekanntheit, Häufigkeit und lexikographische Erfassung von Sprichwörtern". In: Häcki Buhofer et al. (Hrsg.) (2001); 99-106.

Ďurčo, P. (2006): „Methoden der Sprichwortanalysen oder auf dem Weg zum Sprichwörter-Optimum". In: Burger/Häcki Buhofer (Hrsg.) (2006); 3-20.

Grzybek, P. (1991): „Sinkendes Kulturgut? Eine empirische Pilotstudie zur Bekanntheit deutscher Sprichwörter", in: *Wirkendes Wort. Deutsche Sprache und Literatur in Forschung und Lehre, 41/2*; 239-264.

Häcki Buhofer, A./Burger, H./Gautier, L. (Hrsg.) (2001): *Phraseologiae Amor. Aspekte europäischer Phraseologie. Festschrift für Gertrud Gréciano zum 60. Geburtstag.* Baltmannsweiler.

Haß, U. (Hrsg.) (2005): *Grundfragen der elektronischen Lexikographie: elexiko - das Online-Informationssystem zum deutschen Wortschatz.* Berlin/New York.

Lemnitzer, L./Zinsmeister, H. (2006): *Korpuslinguistik: eine Einführung.* Tübingen.

Permjakov, G. (1985): *300 allgemeingebräuchliche russische Sprichwörter und sprichwörtliche Redensarten. Ein illustriertes Nachschlagewerk für Deutschsprechende.* Leipzig.
Permjakov, G. (1986) = Пермяков, Г. (1986): 300 общеупотребительных пословиц и поговорок (для говорящих на болгарском языке). Москва/София.
Permjakov, G. (1988) = Пермяков, Г. (1988): Основы структурной паремиологии. Москва.
Permjakov, G. (1997): „On the Question of a Russian Paremiological Minimum", in: *De Proverbio online. http://www.deproverbio.com/Dpjournal/DP,3,2,97/PERMINIMUM.htm*, gesehen am 06.02.2001.
Scherer, C. (2006): *Korpuslinguistik.* Heidelberg.
Standart = Стандарт. *http://paper.standartnews.com/bg/*, gesehen am 03.08.2008.
Stojkova, S. (2007) = Стойкова, С. (2007): Български пословици и поговорки. София.
Tóthné Litovkina, A. (1996): „Parömiológiai felmérés Magyarországon. (Milyen formában és változatban élnek a legismertebb közmondások, és mi határozza meg az ismeretüket?)", in: *Magyar nyelv, XCII/4*; 439-458.
Unicode Text Segmentation. http://unicode.org/reports/tr29/, gesehen am 30.05.2008.

Hrisztalina Hrisztova-Gotthardt
Foreign Language Centre
University of Pécs
Olga utca 1.X.32
7632 Pécs
Hungary
xpucu@freemail.hu

Body-Part Idioms across Languages: Lexical Analyses of VP Body-Part Idioms in English, German, Swedish, Russian and Finnish

Jussi Niemi/Juha Mulli/Marja Nenonen/Sinikka Niemi/Alexandre Nikolaev/Esa Penttilä

Der vorliegende Beitrag befasst sich mit dem Gebrauch der Somatismen, der mit Hilfe der Frequenzanalyse auf der Basis umfangreicher Korpora des Englischen, Deutschen, Schwedischen, Russischen und Finnischen untersucht wird. Die analysierten VP-Idiome, die aus einem Verb und mindestens einer Körperteilbezeichnung bestehen, zeigen auf, dass hier üblicherweise diejenigen Körperteilbezeichnungen vorkommen, die den frequentesten Körperteilbezeichnungen zugeordnet werden können. Substantive, die auf Mentales referieren, kommen kaum in Idiomen vor. Dies hat zur Folge, dass sowohl in idiomatischer als auch in nichtidiomatischer Sprache dieselben *kognitiven Domänen* als lexikalische Ressource fungieren. Unsere ausführliche Analyse zeigt zugleich, dass die Körperteilbezeichnungen in eng verwandten Sprachen nicht mit gleicher Häufigkeit vorkommen (beispielsweise Deutsch, Englisch, Schwedisch), obwohl man dies eigentlich erwarten könnte. Der gemeinsame historische Hintergrund impliziert nicht unbedingt Ähnlichkeiten im Hinblick auf die Frequenzen der Körperteilbezeichnungen als phraseologische Konstituenten.

1 Background

Across languages idioms typically use parts and processes of the human body to express various acts and mental or physical states (e. g., E. *haul ass, lose one's head*). However, it is surprising to observe that only a small number of comparative studies have been carried out in the sphere of body-part idioms, although they would be of interest, not merely to traditional linguists, but also to scholars in human cognition and cultural practices. In addition to the paucity of comparative studies, to the best of our knowledge, the existing cross-linguistic comparisons of lexical properties of phrasal idioms are methodologically unsatisfactory – to say the least – as they are, first of all, based on small corpora. For instance, the analyses of some languages in Akimoto (1994) are based on relatively narrow bilingual dictionaries. Moreover, Akimoto (1994) contains no quantitative data to support the frequency of use claims made by the author. In addition, since the meanings of idioms are generally hard to capture by using translation equivalents or longer paraphrases that

we typically encounter in dictionaries, we would like to claim that in order to arrive at comparative data in the languages under comparison, each language should be analyzed by a linguist with native or near-native mastery of the target language. Finally, since such a comparison *a fortiori* requires that a group of linguists work together, the members of the group should have a common lingua franca as their working language.

At this juncture, the reader should be warned about terminological shorthand that we use, very much so because of the established use of the term. Since we do not ascribe to Cartesian mind-body dualism, the term "body part" should be interpreted to cover both the physical body and our mental faculties with all their manifestations (e. g., memory, attention, affect). In consequence, in what follows, the term "body-part idiom" should be mentally transformed into "body/mind-part idiom".

2 Aims of the Comparative Study

The present study has three major aims. First, we aim to carry out an extensive cross-linguistic comparison of structurally similar (i.e., V + body-part N) body-part idioms that is amenable to solid quantitative and statistical analyses. In other words, we aim to go beyond the "gut feeling" approach that permeates the (quasi-)quantitative comparative idiom studies. A second, somewhat overlapping aim is to compare some lexical properties such as they appear in body-part idioms vis-à-vis in language in general. This second aim is motivated by the fact that we as humans are bipedal and manually dexterous animals who mostly rely on sense organs of distal perception (vision, audition). Thus, it is of some interest to cognitive linguists to test how these and other biological and psychological facts could be (semi)lexicalized into phrasal idioms. The third aim is to study the possible reflections of different cultural practices of the speaker communities in idiomatic language. Although English, German, Swedish, Russian and Finnish are basically SAE languages and although their major speaker communities culturally share the European Judeo-Christian tradition, it is expected that some differences in cultural practices may be detected in the data.

3 Data and Procedures

3.1 Structure of idioms

The verb plus complement noun structure appears to allow a high degree of productive variation in its verb and especially in its noun in all the five target languages, and presumably in all languages that carry this type of verb phrase. Thus, the common grammatical ground that was found the most suitable in the present context was that provided by the verb plus complement noun, in which the complement was to be a body-part noun (e. g., *kiss [some]one's ass*). Another motivation behind this decision was that most of the group members have previously analyzed body-part idioms in individual languages, and have

there extended their research interests beyond the present contrastive analysis (see, e. g., Nenonen 2002, 2007 for Finnish, Niemi 2004a,b,c, 2007 for Swedish and Mulli 2007 for German). Moreover, we have also compared cultural versus biological motivation of idioms across languges (English and Finnish) (Penttilä et al. 1998).

3.2 Data sources for idioms

As the type of idioms are typically semi-lexicalized items that often have a mono-lexical "counterpart" in the language, we used the following large dictionaries and/or specific idiom dictionaries to collect the idioms that were to be subsequently subject to our qualitative and quantitative analyses:

- English: Collins (1995), Fowler (1982), Makkai et al. (1984), Makkai et al. (1995), Seidl/McMordie (1992), and Wallace (1981)
- German: Duden (2002)
- Swedish: *Målande uttryck* (1990), Kari (1993b), Karlsson (1982–1987), and Romppanen et al. (1997)
- Russian: Евгеньева А.П. (1997)
- Finnish: *CD-perussanakirja* (1997); Kari (1993a), Kivimies (1964), Nurmi et al. (1991), and *Nykysuomen sanakirja I-VI* (1973)

The analyses of the dictionaries and the frequent joint meetings of the group eventually produced altogether about 5,300 idioms that fulfilled the analysis criteria (English: ca. 550, German: ca. 1,300, Swedish: ca. 1,800, Russian: ca. 800, and Finnish: ca. 830).

3.3 Data sources for frequency analyses

In each language, the idioms chosen from the dictionaries were subjected to frequency analyses using the following frequency data and/or full-text corpora:

- English: Leech et al. (2001): *Word frequencies in written and spoken English: Based on the British National Corpus* (ca. 100 million words)
- German: *IDS-Korpus* (*http://www.ids-mannheim.de/cosmas2/*, W - Archiv der geschriebenen Sprache: alle öffentlichen Korpora des Archivs W; at the time ca. 1 billion words)
- Swedish: Allén (1970): *Nusvensk frekvensordbok baserad på tidningstext, Vol. 2: Lemman* (ca. 1 million running words)
- Russian: *Russian National Corpus* (*http://www.ruscorpora.ru/en/corpora-intro.html*, ca. 150 million words)
- Finnish: *The Language Bank* (of Finland) (http://www.csc.fi, 131.4 million running words)

4 Results

4.1 Distribution of idioms per body-part noun

Our analysis covers the twenty most frequent body-part nouns occurring in the data in each language. Not unexpectedly, the body-part nouns are not evenly distributed within or across languages, as is vividly shown in the distribution of the number of idioms in function of the complement noun (table 1 on the facing page).[1] Although the majority of the top-twenty body parts were shared by most of the languages in our data (although only 11 by all), there are clear differences between languages, and consequently no two languages have exactly the same set of body-part nouns in their top 20 lists of frequency. Thus, due to these language-specific differences in the distribution of body-part nouns, our top 20 of each individual language eventually turned out to be the top 34 of the whole of our data.

4.2 Mental vs. non-mental nouns

Although concrete nouns are frequently used metaphorically to refer to mental states (e. g., E. *to catch someone's eye* 'to capture someone's attention'), it may be of interest to observe that – paradoxically enough – the target languages seldom use nouns with direct mental referents in their body-part idioms. Moreover, these languages use different lexical resources in idioms with mental nouns (table 2 on page 72).[2]

4.3 Nouns shared by all five languages in Top 20 idioms

It is counterintuitive that as few as 11 body-part nouns are shared by all five languages, when the 20 most frequent idioms per language are analyzed (table 3 on page 72). Moreover, our straightforward measure for overall body-part indexicality calculated in table 3 on page 72 shows that idiomatization also shows that humans are representatives of a species "with hands, heads/brains, and eyes". In other words, we conceptualize ourselves as manually manipulative and cognitive animals that use distal (visual) perception as their primary sensory channel for acquiring information about the external world.

4.4 Nouns in body-part idioms and in language in general

Not unexpectedly, there is a trend towards the following pattern: The more frequent the body-part noun is in an individual language, the more frequently it tends to be used in

---

1 In order for us to arrive at the quantitative data expressed in table 1 on the next page, several negotiations had to be carried out, since the human body tends to be segmented differently in different languages. Especially, the areas of the neck/throat and those of the upper extremities turned out be hard nuts to crack in this respect. Moreover, some animal-referent nouns, like the Finnish *nokka* 'beak', are used metaphorically for their human analogues (here: 'nose'). These instances are separately listed in the table.

2 Curiously enough, in the present data, German uses only one mental referent lexeme in these idioms, viz., 'soul' (cf. long-term influence of Christianity on Europeans).

	Body part	English	German	Swedish	Russian	Finnish
1	hand	19 (82)	13 (117)	16 (127)	17 (135)	9 (65)
2	head	10 (41)	12 (103)	8 (65)	10 (80)	12 (87)
3	eye	8 (34)	10 (92)	9 (76)	12 (91)	11 (77)
4	heart	13 (55)	6 (55)	7 (60)	5 (41)	6 (41)
5	foot/feet	4 (17)	6 (49)	6 (48)	7 (57)	7 (53)
6	ear	4 (17)	7 (63)	7 (53)	5 (39)	5 (36)
7	mouth	3 (14)	4 (39)	5 (37)	2 (19)	7 (50)
8	nose	3 (12)	5 (43)	6 (50)	4 (28)	4 (29)
9	blood	3 (12)	3 (26)	4 (32)	4 (34)	4 (27)
10	mind	7 (30)			5 (37)	6 (41)
11	tongue	2 (10)	2 (20)	3 (23)	4 (33)	4 (27)
12	finger	3 (11)	4 (38)	3 (24)	2 (16)	3 (23)
13	face	4 (19)	3 (28)	3 (21)	4 (28)	
14	leg	2 (9)	6 (51)	5 (41)		
15	soul		2 (15)		8 (60)	2 (16)
16	back		2 (18)	5 (42)	2 (12)	3 (18)
17	(front of) neck		5 (43)	2 (20)	2 (18)	
18	spirit				2 (14)	5 (33)
19	throat		2 (14)		2 (18)	2 (15)
20	tooth/teeth	3 (12)			3 (22)	
21	(back of) neck	3 (11)				2 (17)
22	arm	3 (11)		2 (19)		
23	hair	2 (9)		2 (17)	1 (5)	
24	ass		4 (32)			
25	nerve					3 (21)
26	heel	3 (11)				
27	body			2 (20)		
28	brain	2 (10)				
29	skin					2 (16)
30	beak					2 (16)
31	sense			2 (18)		
32	snout		2 (18)			
33	look			2 (16)		
34	behind		2 (16)			
	Total	(427)	(880)	(809)	(787)	(708)

Table 1: Proportional frequencies (%) and absolute frequencies of body-part nouns in VP idioms in the data (n=3611) (absolute frequencies in parentheses).

Mental noun	Language and percentage		
'brain'	English 2,3		
'sense'	Swedish 2,2		
'spirit'	Russian 1,8	Finnish 4,7	
'mind'	Russian 4,7	Finnish 5,8	English 7,0
'soul'	Russian 7,6	Finnish 2,3	German 1,7

Table 2: Frequency (%) of mental noun idioms in the five languages (derived from table 1 on the previous page).

Body part	English	German	Swedish	Russian	Finnish	Mean
'hand'	19,2	13,3	15,7	17,2	9,2	14,9
'head'	9,6	11,7	8,0	10,2	12,3	10,4
'eye'	8,0	10,5	9,4	11,6	10,9	10,0
'heart'	12,9	6,3	7,4	5,2	5,8	7,5
'foot'	4,0	5,6	5,9	7,2	7,5	6,0
'ear'	4,0	7,2	6,6	5,0	5,1	5,5
'mouth'	3,3	4,4	4,6	2,4	7,1	4,4
'nose'	2,8	4,9	6,2	3,6	4,1	4,3
'blood'	2,8	3,0	4,0	4,3	3,8	3,6
'tongue'	2,3	2,3	2,8	4,2	3,8	3,1
'finger'	2,6	4,3	3,0	2,0	3,2	3,0

Table 3: Body-part nouns shared by all five languages in their 20 most frequent body-part idioms (percentages, derived from table 1 on the preceding page).

Language	r	p-value	adjusted R^2
English	0,77	$<$ 0,000	0,57
German	0,89	$<$ 0,000	0,78
Swedish	0,86	$<$ 0,000	0,73
Russian	0,85	$<$ 0,000	0,70
Finnish	0,65	0,002	0,58

Table 4: Frequency correlation between body-part idioms and body-part lexemes (with R statistics for significance of correlation), without the Finnish 'mind'.

Language	En	Ge	Sw	Ru
Ge	0,79			
Sw	0,88	0,87		
Ru	0,83	0,83	0,91	
Fi	0,60	0,89	0,64	0,72

Table 5: Frequency correlation coefficients of use of body-part nouns across the five languages.

body-part idioms (correlations hover around 0.7–0.9; table 4 on the preceding page).[3] In other words, body-part idioms do not typically use infrequent archaic or Latinate, "learned" nouns, for instance. Instead they derive their nouns from the common lexical stock of the language.

Finally, although no two languages were found in the present sample that would have had exactly the same frequency of use order in their body-part idiom nouns (tables 1 on page 71 and 3 on the preceding page), there exists a high degree of correlation across all the five languages in this respect (table 5), since even the correlation coefficient ($r^2 = 0.60$) between the two languages with the lowest degree of correlation, viz., English and Finnish, is statistically significant ($p = 0.03$).

5 Conclusions

On the basis of our frequency analyses of large corpora of English, German, Swedish, Russian and Finnish VP idioms there is a strong general trend for these idioms to pick their body-part nouns from among the most frequent ones. Thus, we may claim that – across languages – *the same cognitive domains* tend to be used in the lexical resources for *idiomatic and in non-idiomatic language*. Yet it may seem surprising that nouns with actual mental referents are as rare as they are in body/mind-part idioms in these five languages. This, we propose, is explained by the fact that in idiomatic language in general, **metaphorical shifts** from corporeal (physical) domains to mental ones are often used for the cognitive faculties. For instance, the present languages show classic shift patterns like those in table 6 on the next page.

However, when the languages are analyzed in depth for the lexical frequencies of their VP idiom body-part nouns, we are in the position to claim that – in contrast to Akimoto (1994) – these extensive quantitative comparisons show that the nouns in body-part idioms do not have the same neat prevalence of occurrence in closely related languages (e. g., German, English, Swedish), as they appear to have when superficially scrutinized. Neither does a common cultural and cultural-linguistic background seem to necessarily imply a

3 The only exception here is the Finnish for 'mind' (*mieli*), which is frequently used as a formally semi-frozen adverb *mielellään*, meaning 'with pleasure'. If included, this lexeme would skew the present correlation for Finnish.

Mental domains	Physical nouns (exx.)
Intelligence	'head'
Attention	'eye'
Memory	'head', 'heart'
Emotion	'blood', 'heart'
Language (verbal output)	'tongue', 'mouth'
Language (auditory input)	'ear'

Table 6: Examples of metaphorical shifts in the present data from corporeal domains to mental ones.

strong similarity in the frequency of use of referents in body-part idioms (e. g., Swedish and Finnish; both official languages of Finland). These apparent paradoxes become understandable when we assume that many idioms are typically constructed very creatively. The prototypical idiom types here are those that are not based on fully specified lexemes but on constructional analogy. The English construction *to do a* [person's name] (and its Finnish counterpart *tehdä* [person's name]-plural)'to behave like person X', is a case in point (for the English construction, see Penttilä (2006), for the Finnish one, see Niemi et al., 1988). However, the present cross-linguistic comparison shows that creativity of idiom innovation is heavily constrained by the fact that our mind and body tend to be sources for our cognitive and linguistic extensions into the so-called external world.

Bibliography

Dictionaries

Allén, S. (1970): *Nusvensk frekvensordbok baserad på tidningstext.* Vol. 2: Lemman. Stockholm.
CD-perussanakirja (1997): *Kotimaisten kielten tutkimuskeskuksen julkaisuja 94.* Helsinki.
Collins Cobuild Dictionary of Idioms (1995). London.
Duden (2002): Wermke, M.; Kunkel-Razum, K.; Scholze-Stubenrecht, W. (eds.) (2002): *Redewendungen: Wörterbuch der deutschen Idiomatik.* Mannheim.
Fowler, W. S. (1982): *Dictionary of Idioms.* New edition. Walton-on-Thames.
Kari, E. (1993a): *Naulan kantaan – nykysuomen idiomisanakirja.* Helsinki.
Kari, E. (1993b): *Svenska här och nu.* Helsinki.
Karlsson, G. (ed.) (1982–1987): *Stora svensk-finska ordboken.* Helsinki.
Kivimies, Y. (1964): *Näinkin voi sanoa – suomen kielen fraseologiaa.* Helsinki.
Leech, G./Rayson, P./Wilson, A. (2001): *Word frequencies in written and spoken English: Based on the British National Corpus.* Harlow.
Makkai, A./Boatner, M.T./Gates, J.E. (1984): *Handbook of Commonly Used American Idioms.* Woodbury, NY.
Makkai, A./Boatner, M.T./Gates, J.E. (1995): *A Dictionary of American Idioms.* 3rd ed. Hauppauge, NY.
Målande uttryck – en liten bok med svenska idiom (1990). Språkdata, Göteborg.
Nurmi, T./Rekiaro, I./Rekiaro, P. (1991): *Suomen kielen sanakirja.* Helsinki.

Nykysuomen sanakirja I-VI (1973). Porvoo.
Romppanen, B./Cantell, I./Sundström, M.-P. (eds.) (1997): *Stora finsk-svenska ordboken.* Helsinki.
Seidl, J./McMordie, W. (eds.) (1992): *Oxford Pocket English Idioms.* Oxford.
Wallace, M.J. (1981): *Dictionary of English Idioms.* Glasgow and London.
Евгеньева, А.П. (ed.) (1997): Словарь русского языка. В 4-х т. РАН, Институт лингвистич. исследований; 4-е изд. Москва.

Theoretical literature

Akimoto, M. (1994): "A typological approach to idiomaticity", in: *The 20th LACUS forum, 1993.* Lake Bluff; 459–467.
Haukioja, T. (ed.) (1988): *Papers from the 16th Scandinavian Conference of Linguistics.* Turku.
Melander, B./ Melander Marttala, U./Nyström, C./Thelander, M./Östman, C. (eds.) (2004): *Svenskans beskrivning 26, Förhandlingar vid Tjugosjätte sammankomsten för svenskans beskrivning, Uppsala den 25—26 oktober 2002.* Uppsala.
Mulli, J. (2007): "Lexical observations on German idioms". In: Nenonen/Niemi (eds.) (2007); 193–199.
Nenonen, M. (2002): *Idiomit ja leksikko. Suomen kielen lausekeidiomien, syntaktisia, semanttisia ja morfologisia piirteitä* [Idioms and the lexicon. Syntactic, semantic, and morphological features of Finnish phrasal idioms]. Joensuu.
Nenonen, M (ed.) (2004): *Papers from the 30th Finnish Conference of Linguistics, Joensuu, May 15—16, 2003.* Joensuu.
Nenonen, M. (2007): "Prototypical idioms: Evidence from Finnish", in: *SKY Journal of Linguistics 20;* 309–330.
Nenonen, M./Niemi, S. (eds.) (2007): *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, May 19–20, 2006.* Joensuu.
Niemi, J./Nenonen, M./Penttilä, E. (1988): "Number as a marker of idiomaticity", in: Haukioja (ed.) (1988); 293–304.
Niemi, J./Odlin, T./Heikkinen, J. (eds.) (1998): *Language Contact, Variation and Change.* Joensuu.
Niemi, S. (2004a): "Kielitypologisia ja kognitiivisia havaintoja ruuminosaidiomeista", in: Nenonen (ed.) (2004); 144–150.
Niemi, S. (2004b): "Ruotsin ja suomen idiomatiikka yhteisen kulttuuritaustan indikaattorina", in: Nenonen (ed.) (2004); 151–165.
Niemi, S. (2004c): "Svenskans kroppsdelidiom ur ett språktypologiskt perspektiv", in: Melander et al. (eds.) (2004); 246–254.
Niemi, S. (2007): "Idiomatic and structurally related VPs in Swedish: a corpus analysis", in: Nenonen/Niemi (eds.) (2007); 235–242.
Penttilä, E. (2006): *It Takes an Age to Do a Chomsky: Idiomaticity and Verb Phrase Constructions in English.* Unpublished doctoral dissertation, University of Joensuu.
Penttilä, E./Nenonen, M./Niemi, J. (1998): "Cultural and biological bases of idioms: A crosslinguistic study", in: Niemi et al. (eds.) (1998); 234–245.

Jussi Niemi
University of Joensuu
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
jussi.niemi@joensuu.fi

Juha Mulli
University of Joensuu
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
juha.mulli@uef.fi

Marja Nenonen
University of Joensuu
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
marja.nenonen@uef.fi

Sinikka Niemi
University of Joensuu
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
sinikka.niemi@uef.fi

Alexandre Nikolaev
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
Alexandre.Nikolaev@joensuu.fi

Esa Penttilä
Department of Foreign languages and translation studies
Yliopistokatu 4
80101 Joensuu
Finland
esa.panttilae@uef.fi

Lexikalische Kollokationen und der Beitrag der Internet-Suchmaschine *Google* zu ihrer Erschließung und Beschreibung

Christine Konecny

This contribution is based on a narrow concept of collocations that situates them on a scale between idioms and free lexical combinations. Following Hausmann collocations are seen as hierarchically organised binary constructions consisting of a cognitively superordinate element, the *base*, and a cognitively subordinate element, the *collocator*. In contrast to research with an underlying broad concept of collocations, in which frequency has always played a considerable role, studies based on a more narrow definition have rarely taken this factor into account. Based on a semantic-conceptual approach and assuming a narrow conceptualisation of collocations, this contribution attempts to demonstrate if and to what extent frequency analyses with the internet search engine *Google* can contribute to the collection and improved description of lexical collocations. This will be illustrated with an analysis of selected Italian collocations, studying them from the perspective of the collocator and investigating which other bases this collocator can be connected with.

1 Theoretische Vorbemerkungen

1.1. Untersuchungsgegenstand des vorliegenden Beitrags sind die so genannten lexikalischen Kollokationen. Um diese sprachlichen Konstrukte einer näheren Betrachtung unterziehen zu können, stellt sich zunächst die berechtigte Frage, was unter einer (lexikalischen) Kollokation überhaupt zu verstehen ist. Für den Begriff *(lexikalische) Kollokation* wurden nämlich im Laufe der Kollokationsforschung bis dato zahlreiche unterschiedliche Definitionen vorgeschlagen; nicht umsonst ist daher bei Pöll (1996: 11) vom „schillernden Begriff ‚Kollokation'" die Rede. Dazu kommt, dass die einzelnen Definitionen, wie Scherfer (2001: 4) treffend feststellt, „gemeinhin entweder vage oder wenig ausgearbeitet" bleiben. Dies alles hat zur Folge, dass die einzelnen Forscher innerhalb der Kollokationsdiskussion vielfach aneinander vorbeireden, weil sie vermeintlich über denselben Gegenstandsbereich sprechen, ohne jedoch auf genau das gleiche zu referieren; wie Batteux (2000: 73) bemerkt, scheint aber „gerade die ungenaue Begriffsbestimmung der Grund für die Lebhaftigkeit der Diskussion um den Kollokationsbegriff" zu sein. Einig scheinen sich die einzelnen Vertreter der Kollokationsforschung lediglich darin zu sein, dass Kollokationen Lexemverbindungen

auf syntagmatischer Ebene und – sofern von einer weiten Phraseologismuskonzeption ausgegangen wird (cf. Konecny 2007: 3f.) – einen Teilbereich der linguistischen Subdisziplin *Phraseologie* darstellen.

1.2. Der Terminus *Kollokation* wurde in den 50er Jahren des 20. Jahrhunderts im Rahmen des Britischen Kontextualismus von John Rupert Firth geprägt. Das Phänomen als solches war allerdings schon früher wahrgenommen, jedoch nicht mit dem Etikett *Kollokation* benannt worden. In diesem Zusammenhang ist etwa Charles Bally zu nennen, der im Jahre 1909 eine bereits sehr ausgefeilte Klassifikation von Wortverbindungen entwickelte, welche u.a. die mit den Kollokationen vergleichbaren *groupements usuels* enthält. Des Weiteren sei auf Walter Porzig mit seinem 1934 publizierten Artikel „Wesenhafte Bedeutungsbeziehungen" sowie auf Eugenio Coseriu mit seinem Konzept der lexikalischen Solidaritäten (1967) hingewiesen, welche sich in den entsprechenden Beiträgen eingehend mit den zwischen einzelnen Lexikoneinheiten bestehenden syntagmatischen Relationen beschäftigen und daher ebenfalls als Vorläufer des Kollokationsbegriffes gelten können. Was die verschiedenen, seit der eigentlichen Einführung des Terminus entwickelten Kollokationskonzepte betrifft, so reichen diese von sehr weiten wie jenem der Vertreter des Britischen Kontextualismus (cf. Firth 1957; Halliday 1966; Sinclair 1966) sowie der Computer- und Korpuslinguisten, gemäß welchem unter einer Kollokation jegliches Miteinandervorkommen von lexikalischen Einheiten in einem Korpus zu verstehen ist, bis hin zu sehr engen wie jenem von Hausmann (1979; 1984; 1985; 2004), gemäß welchem Kollokationen Lexemverbindungen auf dem Kontinuum zwischen Idiomen auf der einen und freien Wortverbindungen auf der anderen Seite darstellen. Für einen Überblick über verschiedene im Laufe der Kollokationsforschung propagierte Kollokationsauffassungen sei auf die Ausführungen in Konecny (2007: 11-72) sowie des Weiteren auf Bahns (1997: 9-60), Bischof (2007: 25-60), Gitsaki (1999: 5-26), Gładysz (2003: 13-39), Lehr (1998: 257-261) und Nesselhauf (2005: 11-24) verwiesen.

1.3. Gegenständlichem Beitrag liegt ein *enges* Kollokationsverständnis zu Grunde, welches sich zu einem Großteil an jenem von Hausmann orientiert. Hausmanns Ansatz folgend (1979: 191) wird angenommen, dass eine lexikalische Kollokation eine hierarchisch organisierte binäre Konstruktion darstellt, welche sich aus einem übergeordneten Element, der Basis, und einem untergeordneten Element, dem Kollokator, zusammensetzt: In it. *il latte caglia* (‚die Milch gerinnt') ist *latte* die Basis und *caglia* der Kollokator, in *abbracciare una professione* (‚einen Beruf ergreifen') ist *professione* die Basis und *abbracciare* der Kollokator, in *prezzi salati* (‚gesalzene Preise') ist *prezzi* die Basis und *salati* der Kollokator. Die Grundlage für die Unterscheidung zwischen Basis und Kollokator bildet kein syntaktisch-morphologisches Kriterium[1], sondern ein kognitives und primär in fremdsprachendidaktischer und lernerlexikographischer Hinsicht relevantes. Die Berücksichtigung dieses Kriteriums wird laut Hausmann (1984: 400ff.; 1985: 121; 2004: 312ff.) vor allem dann nahegelegt, wenn man sich die kognitiven Prozesse vor Augen führt, welche bei einem Sprecher/Schreiber ablaufen, der als L2-Lerner einen Text in der Fremdsprache produzieren möchte. Hier könne man beobachten, dass die Enkodierung einer Äußerung

1 Ausgehend von der Valenztheorie z. B. müsste in einer Substantiv-Verb-Kollokation das Verb (nicht das Substantiv) als das die Konstruktion bestimmende Element angesehen werden.

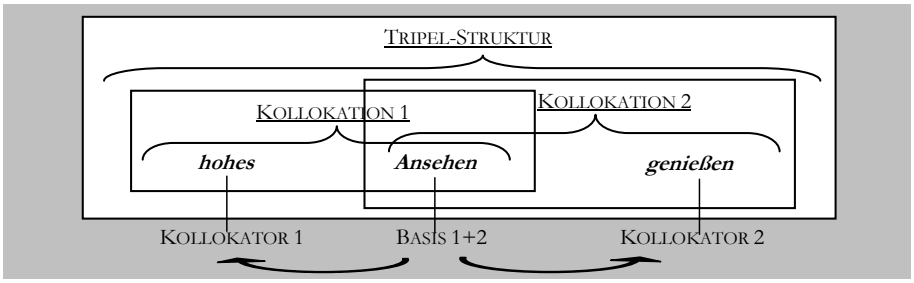

Abbildung 1: Die Tripel-Struktur *hohes Ansehen genießen*

über die Basis zum Kollokator erfolge. Wenn jemand z. B. über einen Junggesellen schreibt oder spricht, setze er das passende Adjektiv *eingefleischt* hinzu. Es sei auch möglich, dass er nach dem Adjektiv sucht, weil es ihm gerade nicht einfällt oder er es nicht weiß. Es sei hingegen nicht denkbar, dass der Textproduzent bereits *a priori* den Kollokator *eingefleischt* im Kopf hat und nach der Basis *Junggeselle* sucht. Immer Hausmann folgend, bestehe ein weiterer Unterschied zwischen Basis und Kollokator darin, dass die Basis meist in ihrer wörtlichen Bedeutung verwendet wird, während die Bedeutung des Kollokators innerhalb der Kollokation von seiner Ausgangsbedeutung auch abweichen kann (cf. Konecny 2007: 49; im Druck b; c). In der Regel sei diese aber aus dem Zusammenhang mit der Basis erschließbar, d. h. die Gesamtbedeutung der Kollokation ergäbe sich in diesem Fall immer noch aus der Summe der Bedeutungen von Basis und Kollokator.

1.4. Geht man von der Hausmann'schen Definition aus, stellen Kollokationen im Regelfall zwar binäre Entitäten dar, bisweilen können zwei Kollokationen aber auch zu einer Tripel-Struktur im Sinne von Hausmann (2004: 316) verschmelzen, in der eine Kollokation in eine andere, übergeordnete kollokative Struktur eingebettet wird, wie im Falle der dt. Kollokationen *massive/scharfe Kritik üben*, *ein umfassendes Geständnis ablegen*, *reißenden Absatz finden* und *hohes Ansehen genießen* (cf. Abbildung 1).

1.5. Nach der engen Konzeption handelt es sich bei Kollokationen – anders als bei Idiomen – um Wortverbindungen, deren Kenntnis unerlässlich ist, um sich in der jeweiligen Sprache der Norm (im Sinne von Coseriu 1970: 40) entsprechend ausdrücken zu können: Lernt man Italienisch, sollte man z. B. wissen, dass ein Nagel im Italienischen „eingepflanzt" wird (*piantare un chiodo*), ein wackelnder Stuhl „hinkt" (*la sedia zoppica*), ein wackelnder Zahn „tanzt" (*il dente balla*) oder man einen verpassten Zug „verliert" (*perdere il treno*). Für Fremdsprachenlerner sind Kollokationen daher besonders tückisch: Auch wenn sie die entsprechende Fremdsprache bereits gut beherrschen, sind sie gerade im Bereich der Kollokationen immer wieder der Gefahr von Interferenzfehlern ausgesetzt, weil sie normalerweise unbewusst dazu tendieren, die Wortverbindungen wörtlich von der Mutter- in die Fremdsprache zu übertragen und damit gegen die lexikalischen Kollokationsrestriktionen der Zielsprache zu verstoßen. Die sich daraus ergebende, zweifellos große

Bedeutung der Kollokationen im Spracherwerb, in der Fachdidaktik, der Translatologie und der (Lerner-)Lexikographie wird in diesem Beitrag jedoch ausgeklammert. Statt dessen sollen in erster Linie die semantisch-begrifflichen Aspekte der Kollokationen und damit das Wesen der Kollokationen an sich sowie die Gründe für die starke Kohäsion zwischen den Kollokationsbestandteilen im Mittelpunkt der Betrachtungen stehen; in diesem Zusammenhang wird speziell der Frage nachgegangen, ob bzw. inwieweit Frequenzanalysen mittels der Internet-Suchmaschine *Google* einen Beitrag zur Erschließung und besseren Beschreibung von lexikalischen Kollokationen leisten können.

2 Das Kriterium der Frequenz im Rahmen der engen Kollokationskonzeption

2.1. Geht man von einem engen Kollokationsbegriff aus, ist das Kriterium der Frequenz zur Identifikation von Kollokationen im Allgemeinen ungeeignet, was u.a. Hausmann (1985: 124ff.) zu Recht betont (cf. Konecny 2007: 110ff.; im Druck b). Einen Beweis dafür stellt z. B. die Tatsache dar, dass bei der Suche mit *Google* am 27.03.2009 für *il rubinetto perde* (‚der Wasserhahn tropft') nur ca. 650, für *stipulare un contratto* (‚einen Vertrag schließen') hingegen ca. 56.700 *tokens* ermittelt werden konnten, dass aber dennoch beide Verbindungen als Kollokationen zu werten sind. Da die Frequenz einer Wortgruppe stets auch mit der Häufigkeit des bezeichneten Sachverhaltes zusammenhängt (cf. Irsula Peña 1994: 35ff.; Klotz 2000: 91f.; Wotjak 1994: 653), kann es zudem vorkommen, dass konventionell nicht festgefügte Wortverbindungen kontextbedingt häufiger auftreten als die konventionell festgefügten Kollokationen, d. h. usuell konsolidierte und daher rekurrent in der entsprechenden Kombination auftretende Verbindungen. So konnten z. B. am 27.03.2009 für die freie Wortverbindung *mangiare un panino* (‚ein Brötchen essen') ca. 44.100 und damit bedeutend mehr *tokens*, als für die Kollokation *il rubinetto perde* ermittelt werden. Für die Klassifikation einer Wortverbindung als Kollokation entscheidend ist deshalb laut Hausmann (1985: 124) nur ihre Disponibilität, d. h., dass sie als Verbindung im mentalen Lexikon der Sprecher verfügbar und jederzeit abrufbar sein muss, unabhängig davon, wie oft sie tatsächlich abgerufen wird.

2.2. Im Gegensatz zur weiten Kollokationskonzeption, innerhalb welcher dem Kriterium der Frequenz seit jeher größte Aufmerksamkeit zuteil wurde, hat dieses innerhalb des engen Verständnisses aus den genannten Gründen kaum Beachtung gefunden. Meines Wissens wurde im Rahmen der engen Auffassung bisher nur unter zwei Gesichtspunkten auf die Relevanz der Frequenz aufmerksam gemacht (cf. Konecny im Druck b):

2.2.1. Der erste dieser Gesichtspunkte betrifft den diachronen Aspekt der Kollokationen (cf. Konecny 2007: 113): Es ist anzunehmen, dass sich einige Kollokationen diachron durch die Konventionalisierung einer häufig vorkommenden freien Wortverbindungen entwickelt haben, worauf u.a. Marello (2000: 205) und Muljačić (1991: 185) hinweisen. Unter diachronem Vorzeichen wären daher Frequenzanalysen von historischen Textkorpora von großem Interesse, weil so nachgewiesen werden könnte, in welchen Texten und ab wann

Teilausschnitt aus der Makrostruktur des Substantivs *Angst*

bekommen	*haben*	*machen*	*äußern*	*überwinden*	...
kriegen	*empfinden*	*einflößen*	*spiegeln*		
	fühlen	*einjagen*	*verraten*		
	merken	*hervorrufen*			
	spüren	*auslösen*			
	verspüren	*erwecken*			
		säen			

VERTIKALE EBENE

HORIZONTALE EBENE

Abbildung 2: Teilausschnitt aus der Makrostruktur des Substantivs *Angst* (nach Irsula Peña 1994: 34)

genau eine bestimmte, heute kollokative Verbindung vermehrt Verbreitung gefunden hat bzw. ab wann früher zulässige Verbindungen „verdrängt" wurden.

2.2.2. Bezüglich des synchronen Aspektes weist Irsula Peña (1994: 33ff.) darauf hin, dass eine Kollokation nicht nur über eine so genannte „Mikrostruktur" verfügt, die aus der Basis und dem jeweiligen Kollokator besteht, sondern auch in eine größere „Makrostruktur" eingebettet ist, in deren Zentrum dieselbe Basis steht und welche auch alle anderen potentiellen Kollokatoren dieser Basis umfasst (cf. Abbildung 2). Auf der horizontalen Ebene der Makrostruktur stünden sich dabei verschiedene Kollokatoren gegenüber, die gemeinsam mit der Basis jeweils unterschiedliche Szenen versprachlichen, wie im Falle der Basis *Angst* die Kollokatoren *bekommen*, *haben* und *machen*. Die vertikale Ebene umfasse dagegen die verschiedenen Möglichkeiten der Versprachlichung *innerhalb* einer Szene. Auf horizontaler Ebene würden Frequenzuntersuchungen nur wenig Sinn machen, weil die zur Verfügung stehenden Verbindungen in der Regel so häufig vorkommen wie der bezeichnete Sachverhalt wahrscheinlich ist. Auf vertikaler Ebene hingegen könnten sie u.U. aussagekräftige Ergebnisse liefern, und zwar dann, wenn zur Versprachlichung eines Sachverhalts mehrere denotativ synonyme Verbindungen existieren, denn hier impliziere hohe Frequenz tendenziell zugleich auch „hohe Typikalität". Laut Irsula Peña (1994: 36f.) ist eine Frequenzzählung auf vertikaler Ebene allerdings nur dann sinnvoll, „wenn sie außer dem jeweiligen Sachverhalt auch die betreffende Textsorte bzw. Kommunikationssituation berücksichtigt."

2.3. In diesem Beitrag soll anhand einer Analyse fünf ausgewählter italienischer Kollokationen untersucht werden, ob Frequenzanalysen im Rahmen eines engen Kollokationsbegriffes nicht auch noch in einer *dritten* Hinsicht interessante Ergebnisse liefern könnten, und zwar dann, wenn eine Kollokation ausgehend vom Kollokator (im Hausmann'schen Sinne) dahingehend untersucht wird, mit welchen anderen Basen dieser in der jeweils aktualisierten Bedeutung noch verbunden werden kann.

3 Zur methodischen Vorgehensweise bei der Beispielanalyse

Die von mir angewandte Methodik bei der empirischen Untersuchung in 4. kann insgesamt als eklektisch bezeichnet werden, weil sie sich verschiedener, vor allem semantisch und kognitiv ausgerichteter Ansätze bedient. Als Korpusgrundlage dient mir das *World Wide Web*, das ich über die italienische Version der Suchmaschine *Google* auf Kontexte für die gewählten Kollokationen hin durchsucht habe.[2] Dabei wurde jeweils von einem bestimmten Kollokatorlexem ausgegangen, das – in den meisten Fällen – mit unterschiedlichen Basen verbunden werden kann. Die *Google*-Suche erfolgte dabei stets in zwei Etappen: (1) Um überhaupt zu einer Liste verschiedener möglicher Basen des jeweiligen Kollokators zu gelangen, wurde zuerst mit der „normalen" einfachen Suche gearbeitet. Bei dieser wurde der Kollokator in unterschiedlichen morpho-syntaktischen Formen (z. B. *digrignare, digrignato* usw.) in *Google* eingegeben; anschließend wurden verschiedene Basis-Substantive durch manuelle Durchsicht aller *Google*-Treffer (bzw. der Vorschau für die einzelnen Treffer) ermittelt, wobei mit den unter 4.2.1-4 jeweils tabellarisch aufgelisteten Basen *nicht* der Anspruch auf Vollständigkeit erhoben werden soll. (2) In einem zweiten Schritt folgte sodann die eigentliche Frequenzanalyse: Bei dieser wurde nach verschiedenen morpho-syntaktischen Formen des Kollokators in Kombination mit verschiedenen möglichen Basen gesucht, wobei die Ergebnisse in Frequenztabellen (cf. 4.2.1-5) aufgelistet werden. In diesen wird jeweils jene Zahl angegeben, die bei erfolgter Suche (in der Regel) mit vorausgehendem „circa" erscheint (cf. Abbildung 3). Bei dieser zweiten Analyse wird allerdings *nicht* mehr mit der einfachen Suche, sondern der Phrasensuche in Anführungszeichen gearbeitet. Bei der it. Kollokation *abbracciare una professione* könnte die einfache Suche etwa so aussehen, dass man *abbracciare* und *professione* einfach hintereinander in das Suchfeld eingibt. In diesem Fall fügt *Google* automatisch den Bool'schen Operator UND zwischen den Suchbegriffen ein. Eine solche Suche hat aber den Nachteil, dass hier auch Ergebnisse

2 Die Gründe für den Gebrauch von *Google* sind vielfältig. Erstens erfolgt die *Google*-Suche extrem schnell und ist äußerst einfach. Als hilfreich erweist sich die Tatsache, dass für jedes Suchergebnis eine Vorschau geboten, d. h. ein kurzer Auszug aus dem betreffenden Text angezeigt wird, wobei die Suchbegriffe optisch durch Fettdruck hervorgehoben sind. Die *Google*-Suche hat weiter den Vorteil, dass die Resultate in der Regel weder in diaphasischer noch in diastratischer oder diatopischer Hinsicht Beschränkungen aufweisen. Dies trifft auch auf die diamesische Ebene zu: Dadurch, dass in die Ergebnislisten z. B. auch Chat-Foren miteinbezogen werden und persönliche Statements im Internet in erster Linie die gesprochene Sprache reflektieren, wird bei den untersuchten Kollokationen automatisch sowohl die geschriebene als auch die gesprochene Dimension berücksichtigt. Ob eine bestimmte Wortverbindung im Italienischen möglich ist, kann man so am schnellsten und zuverlässigsten mit *Google* herausfinden. Während die betreffende Kombination z. B. in der LIZ (*Letteratura italiana Zanichelli*; 2001) überhaupt nicht vorkommen muss, etwa weil es sich um eine erst seit kurzem übliche Verbindung handelt (die in der LIZ erfassten Texte reichen nur bis Gabriele D'Annunzio) oder weil der damit bezeichnete Sachverhalt in den verzeichneten Texten nicht vorkommt, ist bei *Google* die Wahrscheinlichkeit äußerst gering, dass kein Treffer erzielt wird, wenn die Verbindung tatsächlich möglich ist. Ein weiterer Vorzug von *Google* besteht darin, dass auch dem Kriterium der Aktualität Rechnung getragen wird, weil in die Ergebnislisten automatisch stets auch die neuesten Websites miteinbezogen werden.

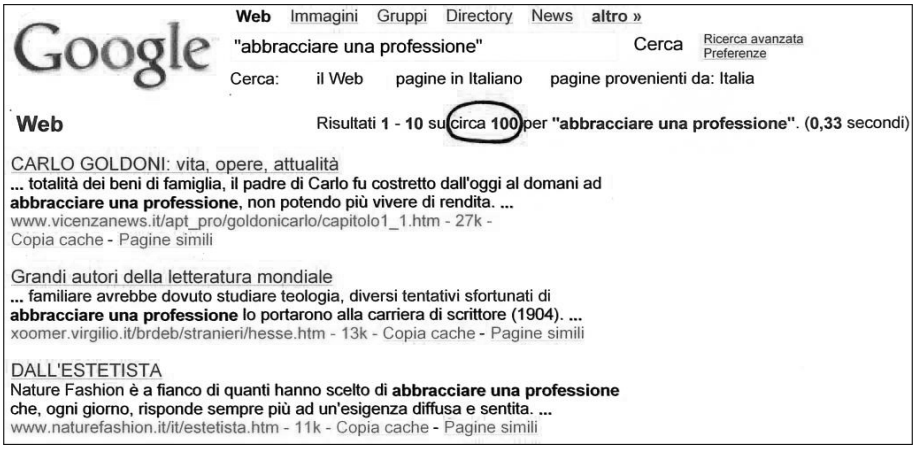

Abbildung 3: Frequenzzählungen mit Phrasensuche in Anführungszeichen mittels *Google*

miteinbezogen werden, in denen zwischen den Suchbegriffen gar keine Verbindung in Form einer Kollokation besteht, weil sie weder syntaktisch noch inhaltlich aufeinander bezogen sind (cf. Konecny 2007: 267). Die Phrasensuche in Anführungszeichen mit mehreren Suchbegriffen hintereinander hat demgegenüber den Vorteil, dass die Suchergebnisse exakt dem eingegebenen Wortlaut in genau der angegebenen Reihenfolge entsprechen. Es ist daher – wenn auch nicht stets, so doch im Normalfall – garantiert, dass die Suchbegriffe tatsächlich eine zusammengehörige Wortgruppe bilden. Allerdings sind die Ergebnisse hier immer genau auf die jeweils eingegebene Form beschränkt. Bei der Eingabe von „abbracciare una professione" wird z. B. nur nach Texten gesucht, in denen vor *professione* der unbestimmte Artikel *una* steht. Bei den Beispielanalysen in 4. wird daher bei Verb-Substantiv-Kollokationen meist sowohl eine Suche mit dem unbestimmten als auch mit dem bestimmten Artikel durchgeführt. Bei den Verbformen werden neben dem Infinitiv auch andere Formen (z. B. das Partizip Perfekt) berücksichtigt. Bei Substantiv-Adjektiv-Kollokationen wird sowohl nach der singularischen als auch der pluralischen Form gesucht.

Die bei der *Google*-Suche ermittelten Frequenzen können aber schon unmittelbar darauf andere sein, weil es sich beim Internet bekanntlich um ein sich ständig im Wandel befindliches Medium handelt, bei dem immer wieder neue Seiten hinzugefügt und andere entfernt werden (cf. Konecny 2007: 270). Deshalb ist in den Frequenztabellen immer auch das Abrufdatum angegeben. Darüber hinaus werden bei *Google* einige Texte doppelt oder sogar mehrfach gezählt, etwa dann, wenn es den betreffenden Text zugleich in verschiedenen Dokumenttypen gibt, z. B. sowohl im PDF-Format als auch als HTML-Website. Die angegebenen Frequenzen sind aus diesen Gründen nicht als absolut, sondern als relativ anzusehen, stellen also eine Art Richtwert dar: Der Wert kann sich zwar laufend ändern, es

ist an ihm aber dennoch eine gewisse Tendenz erkennbar, d. h. man sieht, ob die Verbindung insgesamt nur sehr selten (z. B. 10 Mal) vorkommt, oder aber ein paar 100 oder mehr als 10.000 Mal.

4 Analyse ausgewählter lexikalischer Kollokationen des Italienischen

4.1. Der folgenden Auswahl fünf italienischer Kollokationen liegt ein morpho-syntaktisches bzw. strukturelles Kriterium zu Grunde, insofern ausschließlich solche Kollokationen untersucht werden, deren Basen Substantive darstellen, was prinzipiell auf folgende drei Strukturmuster zutrifft: (a) „Verb plus Substantiv (direktes Objekt)", (b) „Substantiv (Subjekt) plus Verb" und (c) „Substantiv plus Adjektiv (Attribut)". Die Kollokationen *digrignare i denti* (4.2.1) und *marinare la scuola* (4.2.4) gehören dem Typ (a) an, wobei im Falle von *digrignare* grundsätzlich von einem pluralischen Objekt (*i denti*) als prototypischer Basis auszugehen ist, im Falle von *marinare* hingegen von einem singularischen (*la scuola*). *Il sole tramonta* (4.2.2) ist dem o.a. Typ (b) zuzuordnen. *Odio profondo* (4.2.3) und *notte bianca* (4.2.5) stellen Kollokationen Typs (c) dar, wobei bei ersterem sowohl eine post- als auch eine pränominale, bei letzterem hingegen ausschließlich eine postnominale Position des Adjektivs möglich ist.[3]

4.2. Die von Hausmann propagierte Unterscheidung zwischen „Basis" und „Kollokator" (s. o.) wird für die folgende Analyse übernommen, weil mir diese nicht nur in sprachdidaktischer und lexikographischer, sondern vor allem auch in linguistisch-theoretischer Hinsicht sinnvoll zu sein scheint. Wie nämlich u.a. Hausmann (1979: 191) selbst feststellt, liegt bei einer lexikalischen Kollokation eine „orientierte" Kombination vor[4], deren Bestandteile nicht gleichwertig sind. Die Basis als das hierarchisch übergeordnete Element sei dabei als semantisch autonom zu betrachten, der Kollokator als das hierarchisch untergeordnete Element hingegen als semantisch relational (cf. auch Croft 2003: 90; Ramat 2005: 98). Siller-Runggaldier (im Druck) zufolge wird dabei „die Wahl des relationalen Ausdrucks [...] durch die autonome Komponente bestimmt und nicht umgekehrt, und die Verbindung selbst ist nur aufgrund der Relationalität des relationalen Gliedes möglich." Bei den zu untersuchenden fünf Kollokationen ist die Basis stets durch die Kategorie *Substantiv* repräsentiert, deren autonomer Charakter laut Siller-Runggaldier (ibid.) mit der „wortartspezifischen Eigenschaft" zusammenhängt, „dass Substantive grundsätzlich referieren, Adjektive und Verben hingegen nicht", bzw. – mit den Worten von Hausmann (1985: 119) ausgedrückt – , dass es „die Substantive sind, welche die Dinge und Phänomene dieser Welt

3 Zwei weitere syntaktisch-morphologische Kollokationstypen, die in der vorliegenden Analyse ausgeklammert werden, sind z. B. „Verb (= Basis) plus Adverb oder Präpositionalsyntagma (Modaladverbiale)" (z. B. *pentirsi amaramente (di qcs.)* ‚(etw.) bitter bereuen', *bastonare / picchiare (qcn.) di santa ragione* ‚(jmdn.) ordentlich verprügeln') und „Adjektiv / Partizip (= Basis) plus Adverb oder Präpositionalsyntagma (Attribut)" (z. B. *ricco sfondato* ‚steinreich', *nuovo di zecca* ‚funkelnagelneu') (cf. Konecny 2007: 195-207).

4 Mann (1999: 6) spricht in diesem Zusammenhang auch von „polaren Wortverbindungen", Scherfer (2002: 233) hingegen von der „Direktionalität kollokationeller Relationen".

	„digrignare…“	*„digrignato…“*	*„digrigno/-ò…“*[a]	*„digrigna…“*
i denti	8730	622	3200	6030
la bocca	9	4	634	485
il muso	1	0	0	2
il ceffo	1	0	0	0
la mascella	115	1	4	5
le gengive	8	0	4	164
il volto	2	3	1	4
il viso	3	4	4	3
la faccia	5	0	0	6

a Zwischen den Formen *digrigno* (1. Person Singular des Präsens – ‚ich fletsche…‘) und *digrignò* (mit Akzent; 3. Person Singular des *Passato remoto* – ‚er / sie fletschte…‘) wird bei der *Google*-Suche kein Unterschied gemacht, weswegen die betreffenden Frequenzzahlen die *tokens* beider Formen umfassen.

Tabelle 1: Frequenzanalyse möglicher Basen zu *digrignare* mittels *Google* [24.06.2009]

ausdrücken, über die es etwas zu sagen gibt.“[5] Bei den untersuchten Kollokationen stellt die Basis daher immer eine referentielle Größe dar, welche dazu dient, etwas zu benennen bzw. sprachlich zum Gegenstand/Argument zu machen (cf. z. B. auch Detges 1991: 254f.), weshalb mir der Gebrauch des Hausmann'schen Terminus „Basis“ für den substantivischen Bestandteil der zu untersuchenden Kollokationen durchaus gerechtfertigt erscheint.

4.2.1. Das erste Beispiel ist *digrignare i denti* (‚die Zähne fletschen‘; Tabelle 1; ausführlich analysiert in Konecny 2007: 271-276). Die in Battaglia/Barberi Squarotti (1961-2002: s.v.) zu findende Definition für *digrignare* lautet „mostrare i denti arrotandoli rabbiosamente in atto di mordere“. *Digrignare* ist also ein Lexem mit einer sehr spezifischen Bedeutung, das typischerweise mit *denti* kollokiert, weswegen von einer Relation der Implikation im Sinne Coserius (1967: 299) gesprochen werden kann, d. h., dass das Substantiv *denti* in der Bedeutung von *digrignare* bereits implizit mitgesetzt ist. Wie die Analyse mit *Google* zeigt, handelt es sich bei *digrignare* aber um keine unikale Komponente, denn neben *denti* können bisweilen auch andere Substantive als Basen auftreten, wie z. B. *bocca, muso, mascella, gengive, volto* usw., deren Gebrauch auf eine Relation der Kontiguität mit *denti* zurückzuführen ist, insofern ihre Referenten dem selben außersprachlichen Bereich, und

5 Aus diesem Grund stellt das Substantiv laut Hausmann (1985: 119) auch die wichtigste Basiswortart dar. Ein Adjektiv oder Verb komme hingegen nur dann als Basis einer Kollokation in Frage, wenn es durch ein Adverb näher bestimmt ist. Im Gegensatz zu den Substantiven besteht die primäre Funktion von Adjektiven darin, einem Referenten bestimmte Eigenschaften zuzuweisen (sie können damit aber auch den Referenzbereich des Bezugssubstantivs einschränken, wie es bei Substantiv-Adjektiv-Komposita der Fall ist), wobei Adjektive jedoch nichts zum Gegenstand machen, sondern typischerweise in Abhängigkeit von Substantiven fungieren. Verben dienen hingegen in erster Linie der Sachverhaltsdarstellung, d. h. durch sie werden Entitäten explizit Eigenschaften oder Relationen zugeschrieben. (cf. Detges 1991: 254f.; Konecny 2007: 524)

	„. . . tramonta“	*„. . . è tramontato/a“*	*„tramontato/a. . . “*	*„. . . che tramonta“*
il sole	97600	4470	2640	7660
la luna	1210	3820	565	208
la/una stella	9 / 94	3 / 4	235 / 10	1 / 7
la stella polare	1	0	1	0
l’/un astro	41 / 24	4 / 1	125 / 0	588 / 86
il/un giorno	(306 / 3)[a]	(9 / 1)	(319 / 6)	(356 / 6)
la/una sera	(488 / 0)	(0 / 0)	(0 / 0)	(1 / 1)
la/una luce	54 / 1	1 / 0	267 / 0	5 / 4
la luce del sole	7	0	1	333
il/un pianeta	57 / 1	2 / 0	0 / 0	1 / 1
Nettuno	499	0	0	1
Venere	218	0	4	116
Giove	235	7	0	7
Saturno	35	2	0	5
Marte	52	0	1	5
Mercurio	91	4	0	2

a Diese Zahlen sind eingeklammert, weil die Nominalsyntagmen *il/un giorno* und *la/una sera* in einigen der mit Hilfe von *Google* ermittelten Belege nicht als Subjekt, sondern als Temporaladverbiale zu interpretieren sind und daher nicht die Basis zu *tramontare* darstellen (dies beispielsweise in einem Satz wie *La mattina sorge il sole e la sera tramonta.*), wodurch sich die Frequenzzahlen entsprechend erhöhen. Es liegt hier also einer der wenigen Ausnahmefälle vor, in denen die eingegebenen Suchbegriffe nicht zwingendermaßen ein zusammengehöriges Syntagma bilden.

Tabelle 2: Frequenzanalyse möglicher Basen zu *tramontare* mittels *Google* [27.03.2009]

zwar dem Körperteil ‚menschlicher Kopf‘, angehören. Wie die Tabelle 1 zeigt, stellt *denti* aber die weitaus häufigste und somit prototypische Basis dar. Zwar gehen auch die anderen Basen kollokationelle Verbindungen mit *digrignare* ein, diese sind jedoch als von der prototypischen Kollokation abgeleitete, periphere Kollokationen anzusehen, die zwar theoretisch möglich, in der Praxis jedoch viel weniger frequent sind.

4.2.2. Die zweite untersuchte Kollokation ist *il sole tramonta* (‚die Sonne geht unter‘; Tabelle 2; ausführlich analysiert in Konecny 2007: 346-360). Der Kollokator *tramontare* stellt ein Verb mit sehr spezifischer Bedeutung (cf. Battaglia/Barberi Squarotti 1961-2002: s.v.) und dementsprechend eingeschränktem Kombinationsradius[6] dar. Das Substantiv *sole* weist innerhalb der verschiedenen Basen zwar die höchste Frequenz auf, es gibt aber mehrere andere Basen, die ebenfalls relativ häufig vorkommen, weswegen bei *sole* wohl eher nicht von einer prototypischen Basis zu sprechen ist. Was den Zusammenhang der mit

6 Der „Kombinationsradius“ (auch: „Kombinationsbereich“) eines bestimmten Kollokator- (oder sonstigen) Lexems umfasst all jene Lexeme bzw. Klassen von Lexemen, die mit diesem eine syntagmatische Verbindung eingehen können (cf. Konecny 2007: 244).

	postnominaler Kollokator („*. . . profondo/-a*“)	pränominaler Kollokator („*profondo/-a. . .* “)
odio	20 700	5330
amore	89 600	28 400
dolore	12 200	31 900
sentimento	16 500	16 000
desiderio	10 100	15 300
tristezza	4720	24 900
orrore	6310	1070
umiltà	724	6630
pietà	1890	4000
pena	1050	1640
disgusto	504	4470
sgomento	130	1430
affezione	253	346
presentimento	143	31
melancolia	9	30

Tabelle 3: Frequenzanalyse möglicher Basen zu *profondo* mittels *Google* [27.03.2009]

tramontare möglichen Basen betrifft, so gehören diese zum Großteil einem gemeinsamen semantischen Feld an, als dessen Archilexem das Kompositum *corpo celeste* oder das Substantiv *astro* fungiert. Es kann daher von einer Selektion im Sinne Coserius (1967: 299) oder zumindest von einer selektionsähnlichen Relation gesprochen werden. Eine Ausnahme bilden lediglich die Basen *giorno, sera* und *luce*, bei denen sich die Möglichkeit der Verbindung mit *tramontare* aber auf Grund einer auf außersprachlichen Faktoren basierenden Verwandtschaft mit *sole* ergibt: Mit dem Sonnenuntergang, welcher meist im Laufe des Abends stattfindet, geht in der Regel der Tag zu Ende; damit verschwindet auch das Sonnenlicht und es wird dunkel.[7]

4.2.3. Das dritte Beispiel ist *odio profondo* (‚[abgrund]tiefer Hass‘; Tabelle 3; cf. Konecny 2007: 416-432; im Druck a). Beim Kollokator *profondo* handelt es sich um ein stark polysemes Adjektiv mit der Ausgangsbedeutung „che presenta una notevole distanza fra il pelo dell‘acqua e il fondo“ (Battaglia/Barberi Squarotti 1961-2002: s.v.), in welcher es sich auf die Distanz von der Wasseroberfläche bis zum Grund des Wassers bezieht. Innerhalb der Kollokation *odio profondo* weist es die metaphorische Bedeutung „sentito vivamente e appassionatamente, provato con grande intensità“ (ibid.) auf, bezieht sich also auf die Intensität eines Gefühls. *Odio profondo* ist somit eine Kollokation, deren Kollokator im Vergleich zu seiner Ausgangsbedeutung eine semantische Weiterentwicklung erfahren hat. Diese besteht darin, dass es ein bestimmtes Sem der Ausgangsbedeutung (jenes der konkreten lokalen Dimension) verliert, während ein anderes, in der Ausgangsbedeutung ebenfalls

7 Zur Möglichkeit der Verwendung von *tramontare* in der metaphorischen Bedeutung ‚perdere vigore, autorità, andare in declino‘ und der in diesem Fall mit dem Verb kollokierenden Substantive (z. B. *un’idea/la vita/la guerra tramonta*) cf. Konecny (2007: 348f.).

	„marinare…“	*„marinato…“*	*„marinavo…“*	*„marinava…“*
la scuola	13700	5190	553	808
il liceo	7	8	4	3
l'università	128	49	3	2
la lezione/le lezioni	176 / 659	464 / 90	2 / 91	4 / 91
il corso/i corsi	40 / 42	281 / 3	1 / 3	1 / 2
il seminario	1	3	0	0
l'esame/gli esami	4 / 1	1 / 0	2 / 0	0 / 0
il lavoro	325	794	4	6
l'ufficio	92	331	1	3
la fabbrica	2	0	0	1
la riunione	3	7	0	0
la palestra	10	66	1	0
la messa	145	9	4	4
il concerto	3	4	0	0
il cinema	3	2	0	1

Tabelle 4: Frequenzanalyse möglicher Basen zu *marinare* mittels *Google* [27.03.2009]

vorhandenes Sem (jenes des hohen Grades bzw. der besonderen Dimension) metaphorisch auf den abstrakten Frame ‚Gefühle' übertragen wird. Diese Metapher ist heute aber bereits vollkommen lexikalisiert, weswegen die betreffenden Kollokationen bereits zu den freien Wortverbindungen hin tendieren. Die verschiedenen möglichen Basen gehören dem gemeinsamen semantischen Feld ‚Gefühle' an. Innerhalb dieser Basen zeichnet sich allerdings nicht nur *odio* durch eine hohe Frequenz aus, sondern es kommen auch zahlreiche andere Basen häufig vor, was bedeutet, dass es wiederum keine prototypische Basis gibt.

4.2.4. Die vierte analysierte Kollokation ist *marinare la scuola* (‚die Schule schwänzen'; Tabelle 4; cf. Konecny 2007: 396-415; im Druck a). Die Ausgangsbedeutung von *marinare* entspricht jener des dt. Verbs *marinieren*, in welcher es mit Substantiven mit dem Sem [+essbar] verbindbar ist, und zwar vorwiegend mit solchen, die einen Fisch oder ein Stück Fleisch bezeichnen (die betreffenden Verbindungen sind aber nicht als Kollokationen, sondern als freie Wortverbindungen aufzufassen). In der Kollokation *marinare la scuola* ist das Verb in einer sekundären, auf einer Metapher basierenden Bedeutung verwendet: Die Similarität zwischen Bildspender- und -empfängerbereich besteht darin, dass die Schule bzw. die jeweilige Verpflichtung als etwas angesehen wird, das beim Schwänzen für einige Zeit beiseite gelegt wird, ähnlich wie ein Fisch, der in eine Marinade eingelegt und in dieser über einen gewissen Zeitraum liegen gelassen wird. Diese Metapher ist synchron aber bereits verblasst und der Zusammenhang zwischen den beiden Bedeutungen somit nicht mehr transparent. Von den möglichen Basen zu *marinare* weist *scuola* die höchste Frequenz auf, so dass von ihr als der prototypischen Basis ausgegangen werden kann. *Liceo* (‚Gymnasium') ist mit *scuola* über eine Hyponymie-Relation verbunden. Die meisten anderen Basen sind mit *scuola* im engeren oder weiteren Sinn begrifflich verwandt, insofern sie entweder dem Frame ‚Schule' bzw. den benachbarten Frames ‚Universität' und ‚Arbeit' angehören oder auf

	Belege
„notte bianca“	625 000
„notti bianche“	139 000

Tabelle 5: Frequenzanalyse der Kollokation *notte bianca* mittels *Google* [27.03.2009]

nicht mehr institutionell festgelegte, sondern ganz allgemein von außen aufgedrängte oder freiwillig eingegangene Verpflichtungen (*riunione, palestra, messa, concerto*) bezogen sind. Eine Ausnahme stellt *cinema* dar, denn hier kann wohl nicht mehr von einer eingegangenen Verpflichtung die Rede sein: Es kommt zu einer Aufweichung des Merkmals ‚verpflichtend‘ und zugleich zu einer Entwicklung hin zur Bedeutung ‚absichtliches Fernbleiben von jedweder Tätigkeit in einem öffentlichen Bereich‘. D. h. also, dass durch einen Verlust bestimmter Seme die Basisklassen sich verändern können und durch Reanalyse auch das Verb in seiner Semstruktur modifiziert wird (cf. Siller-Runggaldier 2008: 592; 596). Es wäre daher denkbar, dass *marinare* in Zukunft mit weiteren Basen verbunden werden kann, die keine Verpflichtung bezeichnen, und dass es so eine Bedeutungserweiterung erfährt, auf Grund welcher es schließlich die allgemeinere Bedeutung ‚etw. vermeiden/auslassen/nicht besuchen‘ zum Ausdruck bringt.

4.2.5. Das fünfte Beispiel ist *notte bianca* (‚eine schlaflose/durchwachte Nacht‘; auch: *notte in bianco*; Tabelle 5; cf. Konecny 2007: 510). Das Adjektiv *bianco*, welches die Ausgangsbedeutung ‚weiß‘ hat, ist innerhalb dieser Kollokation in der Bedeutung „trascorsa vigilando“ (Battaglia/Barberi Squarotti 1961-2002: s.v.) verwendet. Die ursprüngliche Motivation der Kollokation ist heute bereits verblasst und die Bedeutung des Kollokators *bianco* somit nicht mehr transparent, weswegen in diesem Fall von einer teilidiomatischen Kollokation gesprochen werden kann. Mit dieser Tatsache hängt auch zusammen, dass *bianco* in der genannten Bedeutung einen unikalen Kollokationsradius aufweist, d. h. einzig und allein mit *notte* verbunden werden kann und keine Ausdehnung auf weitere Basen erfahren hat (dies ist u.a. daran erkennbar, dass die Verbindung im Wörterbuch von Battaglia/Barberi Squarotti [1961-2002] s.v. *bianco* unter dem Punkt „Locuzioni“ als feststehender Ausdruck angeführt ist). Aus der Tatsache, dass es keine weiteren Basen gibt, folgt, dass eine Frequenzanalyse wie in Tabelle 5 in diesem Fall nicht zielführend ist.[8]

5 Schlussfolgerungen

5.1. Wie die Analysen bestätigen, kann sich die Untersuchung einer Kollokation ausgehend vom Kollokator sowohl in sprachdidaktischer als auch in linguistisch-theoretischer Hinsicht als sehr „fruchtbar“ erweisen. In diesem Fall sollte die Kollokation aber nicht isoliert betrachtet werden, sondern stets unter Berücksichtigung des gesamten Basisfeldes, d. h.

8 Ähnliches wie für *notte bianca* gilt im Deutschen übrigens für die Kollokation *blinder Passagier* (cf. hierzu die Ausführungen in Konecny im Druck b).

Teilausschnitt aus der Makrostruktur des Verbs *piantare*

VERTIKALE EBENE

una pianta	*una vigna*	*un chiodo*	*una siringa*	*un'arma*	...
un fiore	*un vigneto*	*un palo*	*un ago*	*un coltello*	
le rose	*una possessione*	*una croce*		*una spada*	
i garofani	*una pianura*	*un'asta*		*un pugnale*	
un albero	*un giardino*	*uno scudo*		*una freccia*	
un abete	*un bosco*	*una bandiera*		*una lancia*	
una palma	*una foresta*	*un'insegna*		*una proiettile*	
il seme	*un orto*	*un vessillo*		*una pallottola*	
le verdure	*un prato*			*i denti*	
i pomodori				*le unghie*	

HORIZONTALE EBENE

Abbildung 4: Teilausschnitt aus der Makrostruktur des Verbs *piantare*

aller möglichen Basen des Kollokators (cf. Konecny 2007: 501; im Druck b). Nur auf diese Weise kann nämlich festgestellt werden, welchen Kombinationsradius letzterer aufweist, d. h. mit welchen und wie vielen verschiedenen Basen er verbunden werden kann und wie diese in semantischer und/oder begrifflicher Hinsicht untereinander zusammenhängen. Das Basisfeld eines Kollokators kann mit den Worten von Irsula Peña (1994: 33ff.) auch als dessen *Makrostruktur* bezeichnet werden. Irsula Peña verwendet diesen Terminus zwar offenbar nur im Zusammenhang mit der Basis, geht eine Kollokationsanalyse jedoch vom Kollokator aus, so ist vorrangig *dessen* Makrostruktur und nur in zweiter Linie jene der Basis von Interesse. Ein Teilausschnitt aus der Makrostruktur des italienischen Kollokatorlexems *piantare* (untersucht in Konecny 2007: 501) kann schematisch wie in Abbildung 4 dargestellt werden.

5.2. Kollokatoren können – wie bereits Siller-Runggaldier (2008) gezeigt hat – ein ganzes Feld von Basen um sich herum scharen, allerdings nicht vollkommen idiosynkratisch, denn die Basen müssen untereinander semantisch oder begrifflich irgendwie verwandt sein. Typische semantische Relationen zwischen den Basen sind Hypero-/Hyponymie (*il sole/l'astro tramonta*; *marinare la scuola/il liceo*), Ko-Hypoymie (cf. *commettere un omicidio/un furto/una rapina* in Konecny im Druck a; b) oder Quasi-Synonymie (cf. *digrignare il volto/il viso/la faccia* sowie die Beispiele *ammazzare la noia/il tedio; commettere un errore/uno sbaglio/una mancanza* in Konecny im Druck a; b). Ferner können die Basen einem gemeinsamen semantischen Feld angehören und durch dasselbe Archilexem/-semem zusammengehalten werden (*il sole/la luna/il pianeta tramonta* – Archilexem: *corpo celeste/astro*; *odio/amore/dolore profondo* – Archilexem: *sentimento*). Eine begriffliche Verwandtschaft ist gegeben, wenn die Basis eine metaphorische (*acqua profonda – odio profondo; marinare un pesce – marinare la scuola*) oder metonymische (*digrignare i denti – digrignare la bocca/la mascella/le gengive* etc.) Weiterentwicklung gegenüber ihrer Ausgangsbedeutung erfahren

hat oder wenn eine außersprachlich bedingte Affinität zwischen den Basen gegeben ist, die durch einen gemeinsamen *Frame* vorgegeben ist (*marinare la scuola/una lezione/un corso/un esame*). Die Prozesse, welche zu bestimmten semantischen Weiterentwicklungen führen, folgen also stets bestimmten Regularitäten und sind daher nicht völlig beliebig. Wenn diesbezüglich überhaupt von Idiosynkrasie gesprochen werden kann, dann nur insofern, als die betreffende Einzelsprache frei zwischen den verschiedenen Prozessen der Versprachlichung und der Basisfelderweiterung „wählen" kann (cf. Konecny 2007: 520).

5.3. Es konnte nachgewiesen werden, dass Kollokationen insgesamt keine starren Konstrukte, sondern äußerst dynamische Lexemverbindungen darstellen, die offenbar einem fortwährenden Modifikationsprozess unterliegen. Dieser Prozess spielt sich vor allem im Bereich der Basen ab, wirkt sich aber auch auf den Kollokator aus, wenn dieser auf Grund veränderter Basen semantisch reanalysiert werden muss und damit seinerseits für neue Basisfelder offen wird (cf. Siller-Runggaldier 2008: 592; 596). Kollokationen sind daher als dynamisches, multifaktorielles Phänomen aufzufassen, das nach unterschiedlichen Kriterien und ausgehend von unterschiedlichen Theorien untersucht werden muss, damit es als solches überhaupt fassbar wird; der von mir gewählte eklektische Beschreibungsansatz stellt daher m.E. eine geeignete Basis dafür dar. Eine klare Abgrenzung der Kategorie *Kollokation* erweist sich vor diesem Hintergrund als nicht mehr möglich. Am besten nähert man sich den Kollokationen wohl durch die Beschreibung ihrer Charakteristika, die aber nicht als notwendige und hinreichende Merkmale anzusehen sind, sondern als solche einer nicht diskreten syntagmatischen Kategorie, die in vielen Fällen eine prototypische Organisation aufweist, d. h. aus zentraleren und periphereren Vertretern besteht.

5.4. Bei der Analyse des Basisfeldes eines Kollokators kommt dem Kriterium der *token*-Frequenz in den meisten Fällen eine entscheidende Rolle zu.[9] Diese erlaubt nämlich wichtige Rückschlüsse auf die interne Struktur des Basisfeldes. Gibt es zu einem Kollokator eine prototypische Basis, so ist diese durch die höchste *token*-Frequenz erschließbar und zeichnet sich durch eine besondere kognitive Salienz aus. Zum Zweck der Unterscheidung von anderen Arten syntagmatischer Wortverbindungen stellt die Frequenz allerdings kein brauchbares Kriterium dar, weil beispielsweise auch rekurrente freie Wortverbindungen häufig auftreten können, sodass Frequenz in diesem Fall kategorial nicht distinktiv ist. Daraus folgt, dass die Frequenz lediglich zur Ausfindigmachung zentraler und peripherer Basen für einen bestimmten Kollokator und somit nur *innerhalb* gegebener kollokationeller Strukturen von Relevanz ist, sich aber nicht zur Identifizierung bzw. allgemeinen Überprüfung des Kollokationsstatus bestimmter Verbindungen eignet. Eine Frequenzanalyse mittels der Internet-Suchmaschine *Google* ist zur Erschließung von Kollokationen selbst also nicht geeignet, kann aber sehr wohl einen bedeutenden Beitrag zum Aufzeigen der Charakteristika von Kollokationen und somit zu deren Beschreibung leisten.

---

9 Die einzigen Ausnahmefälle, in welchen Frequenzanalysen keinen Sinn machen, sind teilidiomatische Kollokationen des Typs *notte bianca*, weil hier, wie in 4.2.5. festgestellt, für den Kollokator (*bianco*) im Rahmen der in der Kollokation aktivierten, demotivierten Bedeutung insgesamt nur ein einziges Kombinationselement (*notte*) in Frage kommt.

Literaturverzeichnis

Bahns, J. (1997): *Kollokationen und Wortschatzarbeit im Englischunterricht.* Tübingen.

Bally, Ch. (1909): *Traitè de stylistique française 1*. Heidelberg.

Battaglia, S./Barberi Squarotti, G. (1961-2002): *Grande dizionario della lingua italiana.* Torino.

Batteux, M. (2000): *Die französische Synonymie im Spannungsfeld zwischen Paradigmatik und Syntagmatik*. Wien.

Bazell, C. E./Catford, J. C./Halliday, M. A. K./Robins, R. H. (Hrsg.) (1966): *In Memory of J. R. Firth.* London.

Bergenholtz, H./Mugdan, J. (Hrsg.) (1985): *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984.* Tübingen.

Bischof, B. (2007): *Französische Kollokationen diachron. Eine korpusbasierte Analyse.* Stuttgart (Diss.). http://elib.uni-stuttgart.de/opus/volltexte/2008/3418/ [19.03.2009].

Coseriu, E. (1967): „Lexikalische Solidaritäten", in: *Poetica, 1/3*; 293-303.

Coseriu, E. (1970): *Einführung in die strukturelle Betrachtung des Wortschatzes.* Tübingen.

Cresti, E. (Hrsg.) (2008): *Prospettive nello studio del lessico italiano. Atti del IX Congresso della Società Internazionale di Linguistica e Filologia Italiana (SILFI), Firenze, 14-17 giugno 2006.* Firenze.

Croft, W. (2003): „Il ruolo dei domini semantici nell'interpretazione di metafore e metonimie". In: Gaeta/Nuraghi (Hrsg.) (2003); 77-100

Danler, P./Iliescu, M./Siller-Runggaldier, H. (Hrsg.) (im Druck): *Kongressakten des XXV Congrès International de Linguistique et Philologie Romanes (CILPR 2007), Innsbruck, 02.09.2007-08.09.2007.* Tübingen.

Detges, U. (1991): „Französische Funktionsverbfügungen vom Typ *être Präp. N*. Zum Verhältnis von lexikalischer Kategorie und propositionaler Funktion". In: Koch/Krefeld (Hrsg.) (1991); 253-277.

Firth, J. R. (1957): „Modes of Meaning. Essays and studies (The English Association)". In: Firth (Hrsg.) (1957); 190-215.

Firth, J. R. (Hrsg.) (1957): *Papers in Linguistics 1934-1951.* London.

Gaeta, L./Nuraghi, S. (2003): *Introduzione alla linguistica cognitiva.* Roma.

Gautier, L./Mejri, S. (Hrsg.) (im Druck): *Les collocations dans les discours spécialisés. Workshop im Rahmen der Internationalen Konferenz Europhras 2008 zum Thema „Phraseologie global – areal – regional", Helsinki, 13.08.2008-16.08.2008.* Dijon.

Gitsaki, C. (1999): *Second language lexical acquisition. A study of the development of collocational knowledge.* San Francisco/Calif [u.a.].

Gładysz, M. (2003): *Lexikalische Kollokationen in deutsch-polnischer Konfrontation.* Frankfurt a.M. [u.a.].

Halliday, M. A. K. (1966): „Lexis as a linguistic level". In: Bazell et al. (Hrsg.) (1966); 148-162.

Hausmann, F. J. (1979): „Un dictionnaire des collocations est-il possible?", in: *TraLiLi, 17/1*; 187-195.

Hausmann, F. J. (1984): „Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", in: *Praxis des neusprachlichen Unterrichts, 31*; 395-406.

Hausmann, F. J. (1985): „Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels". In: Bergenholtz/Mugdan (Hrsg.) (1985); 118-129.

Hausmann, F. J. (2004): „Was sind eigentlich Kollokationen?". In: Steyer (Hrsg.) (2004); 309-334.

Irsula Peña, J. I. (1994): *Substantiv-Verb-Kollokationen. Kontrastive Untersuchungen Deutsch-Spanisch.* Frankfurt a.M. [u.a.].

Klotz, M. (2000): *Grammatik und Lexik. Studien zur Syntagmatik englischer Verben.* Tübingen.

Koch, P./Krefeld, T. (Hrsg.): *Connexiones Romanicae. Dependenz und Valenz in romanischen Sprachen.* Tübingen.
Kolboom, I./Kotschi, T./Reichel, E. (Hrsg.): *Handbuch Französisch. Sprache – Literatur – Kultur – Gesellschaft. Für Studium, Lehre, Praxis.* Berlin.
Konecny, C. (2007): *Kollokationen. Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand des Italienischen.* Innsbruck (Diss.).
Konecny, C. (im Druck a): „Le collocazioni lessicali – proposta per una classificazione semantica". In: Danler et al. (Hrsg.) (im Druck).
Konecny, C. (im Druck b): „Le critère de fréquence est-il fiable pour la mise en évidence de collocations spécialisées?". In: Gautier/Mejri (Hrsg.) (im Druck).
Konecny, C. (im Druck c): „Divergenze e convergenze in collocazioni lessicali italiane e tedesche". In: Lavric/Pöckl (Hrsg.) (im Druck).
Lavric, E./Pöckl, W. (Hrsg.) (im Druck): *Akten der VI. Internationalen Arbeitstagung „Romanisch-deutscher und innerromanischer Sprachvergleich", Innsbruck, 03.09.2008-05.09.2008.* Frankfurt a.M. [u.a.].
Lehr, A. (1998): „Kollokationen in Langenscheidts Großwörterbuch Deutsch als Fremdsprache". In: Wiegand (Hrsg.) (1998); 256-279.
LIZ (2001) = Stoppelli, P./Picchi, E. (2001): *Letteratura italiana Zanichelli – LIZ 4.0. CD-Rom dei testi della letteratura italiana.* Bologna.
Lorenz-Bourjot, M./Lüger, H.-H. (Hrsg.) (2001): *Phraseologie und Phraseodidaktik.* Wien.
Mann, D. (1999): *Verb-Nomen-Kollokationen: Französische Sprechakte mit einem kontrastiven Ausblick auf das Deutsche.* Stuttgart.
Marello, C. (2000): *Le parole dell'italiano. Lessico e dizionari.* Bologna.
Muljačić, Ž. (1991): *Scaffale italiano. Avviamento bibliografico allo studio della lingua italiana.* Firenze.
Nesselhauf, N. (2005): *Collocations in a Learner Corpus.* Amsterdam/Philadelphia.
Pöll, B. (1996): *Portugiesische Kollokationen im Wörterbuch. Ein Beitrag zur Lexikographie und Metalexikographie.* Bonn.
Porzig, W. (1934): „Wesenhafte Bedeutungsbeziehungen", in: *Beiträge zur Geschichte der deutschen Sprache und Literatur, 58*; 70-97.
Ramat, P. (2005): *Pagine linguistiche. Scritti di linguistica storica e tipologica.* Roma/Bari.
Scherfer, P. (2001): „Zu einigen wesentlichen Merkmalen lexikalischer Kollokationen". In: Lorenz-Bourjot/Lüger (Hrsg.) (2001); 3-19.
Scherfer, P. (2002): „Lexikalische Kollokationen". In: Kolboom/Kotschi/Reichel (Hrsg.) (2002); 230-237.
Sandig, B. (Hrsg.): *Europhras 1992, Saarbrücken. Tendenzen der Phraseologieforschung*. Bochum.
Siller-Runggaldier, H. (im Druck): „Syntagmatik und Ontologie: Zweigliedrige Lexemverbindungen im interlingualen Vergleich (Deutsch, Italienisch, Französisch, Ladinisch)". In: Lavric/Pöckl (Hrsg.) (im Druck).
Siller-Runggaldier, H. (2008): „Le collocazioni lessicali: strutture sintagmatiche idiosincratiche?". In: Cresti (Hrsg.) (2008); 591-598.
Sinclair, J. (1966): „Beginning the study of Lexis". In: Bazell et al. (Hrsg.) (1966); 410-430.
Steyer, K. (Hrsg.) (2004): *Wortverbindungen – mehr oder weniger fest.* Berlin/New York.
Wiegand, H. E. (Hrsg.) (1998): *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von ‚Langenscheidts Großwörterbuch Deutsch als Fremdsprache'.* Tübingen.
Wotjak, G. (1994): „Nichtidiomatische Phraseologismen. Substantiv-Verb-Kollokationen – ein Fallbeispiel". In: Sandig (Hrsg.) (1994); 651-677.

Christine Konecny
Institut für Romanistik
Universität Innsbruck
Innrain 52d
6020 Innsbruck
Österreich
Christine.Konecny@uibk.ac.at

The Canonical Form in Murky Waters: Idiom Variation and the Croatian National Corpus

Jelena Parizoska

Die formale Stabilität gilt in der Phraseologie als eine der konstitutiven Eigenschaften von Idiomen. Textlinguistische und korpusbasierte Studien haben diese Ansicht relativiert, sodass man in der Phraseologieforschung nur von der „relativen" Stabilität sprechen kann. Viele Idiome weisen Variationen in ihrer Form auf, wie etwa das kroatische Idiom *loviti u mutnom* (wörtl. im Trüben fangen; ‚etwas auf unehrliche Weise bekommen'), das häufig als *lov/loviti/lovac u mutnom* vorkommt.

Das Ziel dieser Arbeit ist anhand von Belegen aus dem kroatischen Nationalkorpus zu zeigen, dass *loviti u mutnom* ein „schematisches Idiom" darstellt, dessen Variationen seinen konzeptuellen Kern *mutna voda* (‚trübes Wasser') systematisch reflektieren. Die Ergebnisse zeigen, dass verschiedene lexikalische und syntaktische Realisationen des Idioms durch Extensionen und Elaborationen des Modells beschränkt sind. Die Variabilität des betreffenden Idioms wirft die Frage nach den anwendbaren Kriterien für die Formulierung der „kanonischen" Form bzw. der Wörterbuch-Nennform der Phraseologismen unter Berücksichtigung der Frequenz und der Prototypizität.

1 Introduction

Idioms are traditionally defined as units which have a fixed form. This notion has been relativized, especially more recently in light of strong empirical evidence from corpus-based research, which has shown that many idioms occur in one or more lexical and/or syntactic forms (Moon 1998; Cignoni et al. 2002; Omazić 2003; Langlotz 2006). It has also been shown that a number of idioms are schematic in nature. A schematic idiom represents a syntactic construction containing a minimum number of lexical items which activates the meaning of the expression as a whole and serves as the basis for various realizations. The type of realizations which occur within such a unit are tied to and constrained by the underlying cognitive mechanism or mechanisms motivating the idiom. Research done in cognitive linguistics (Lakoff 1987; Gibbs 1990, 1994; Gibbs/O'Brien 1990; Kövecses/Szabó

1996) has shown that the cognitive mechanisms that motivate most idioms are metaphor, metonymy and *idealized cognitive models* (ICMs; cf. Lakoff 1987).[1]

Examples of idioms which have several different forms can be found in many languages, including Croatian. A case in point is the expression *loviti u mutnom* (lit. to hunt in the murk-LOC; 'to gain an unfair advantage from a difficult situation'), which appears in the Croatian National Corpus as *lov/loviti/lovac u mutnom* (lit. a hunt/to hunt/a hunter in the murk). Such relatively variable idioms raise the issue of what has been traditionally known as the canonical form – a kernel construction which serves as the basis for variant lexical and syntactic realizations that are derived from it.

This paper addresses the issue of the criteria used in establishing the canonical form and the dictionary citation-form which reflects the standard usage of an idiom. We argue that three factors are crucial: corpus evidence, the grammatical structure of an idiom and its conceptual motivation. We will try to prove this on the example of the Croatian expression *loviti u mutnom*. The aim is to show that *loviti u mutnom* is a schematic idiom which varies according to the extensions and elaborations of the underlying cognitive mechanism and/or combinations with other mechanisms. More specifically, on the basis of the data from the Croatian National Corpus we will prove that the ICM MUTNA VODA ('murky water') constitutes the conceptual core of the idiom (Langlotz 2006) which acts as the focus for various lexical and syntactic realizations. We will also show that variations are not unpredictable and that all the changes in the lexical make-up and syntactic structure of the given idiom are compatible with and constrained by its conceptual core.

The paper is organized as follows: The first section gives an overview of the issues involved in the variability of idioms and its link to conceptual motivation. The second section discusses the use of computer corpora in Croatian phraseology and the criteria used in determining the citation-form. The third section presents the results of the study of the idiom *loviti u mutnom* in the Croatian National Corpus. The fourth section discusses the results of the study. The final section is the conclusion.

2 Idiom Variability and Conceptual Motivation

Idioms exhibit various degrees of lexico-syntactic variability. It has therefore been suggested that they should be viewed on a scale (Cowie et al. 1985: xii): from items that are completely frozen (e. g. *spick and span*, which cannot be used in any other form) to items which are lexically and syntactically flexible, such as open collocations. Idioms placed closer to the centre exhibit restricted variance, they allow variation of some slots, but resist change to others. Such constructions have been variously termed *formal idioms* (Fillmore et al. 1988), *idiom schemas* (Moon 1998) and *schematic idioms* (Croft/Cruse 2004). They are lexically and syntactically flexible, with open slots that can be filled by a range of semantically related items. Here is an example (Moon 1998: 161):

1 Lakoff defines idealized cognitive models (ICMs) as a way in which we organize knowledge, and enumerates five basic types: propositional, image-schematic, metaphorical, metonymic, and symbolic (1987: 284).

(1) fan the fire of SOMETHING
fan the fires of SOMETHING
fan the flames (of SOMETHING)
add fuel to the fire
add fuel to the flame
add fuel to the flames
fuel the fire
fuel the fires
fuel the flame
fuel the flames (of SOMETHING)

The expressions illustrated in (1) show that some parts of a schematic idiom are lexically open and they may be filled by a range of words which belong to a restricted set. Studies of idiom variation (Gibbs/Nayak 1989; Gibbs/Nayak/Cutting 1989; Gibbs et al. 1989; Cacciari/Glucksberg 1991; Glucksberg 1993; Nunberg et al. 1994) found a close connection between the conceptual motivation of idioms and their lexical and syntactic variability. The meaning of the schematic idiom illustrated in (1) ('to make an argument or bad situation worse') is motivated by the conceptual metaphor ANGER IS FIRE. The variant lexical realizations are directly associated with the construction and must be semantically and syntactically appropriate for the slot (adding fuel, fuelling the fire and fanning the fire all highlight the increase in the intensity of the anger).

Moon (1998: 165) goes so far as to say that all conceptually motivated idioms represent schemas: they are motivated by the same underlying cognitive mechanism and are associated with particular lexical realizations, but do not have fixed lexis or fixed syntactic structure. This can be tied in with Langacker's definition of an idiom which "may be recognized as a unit that is to some degree independent of a specific overt morphemic arrangement, even if one such arrangement is far more similar and hence more 'usual' than the others" (Langacker 1987: 25). Thus, if an idiom possesses several lexical and/or syntactic variations, one of these forms is conventionalized and registered in idiom dictionaries as the canonical (or standard) form associated with an idiomatic meaning.

Research-based evidence shows that many idioms exhibit a considerable potential for variation. More importantly, a number of idioms has been shown to be linked together by a common cognitive mechanism, without having a fixed lexical or syntactic structure. This raises the issue of the criteria used in choosing a construction that is listed in idiom dictionaries as the canonical form.

3 Corpus for Croatian Phraseology

The development of corpus linguistics has led to an increased use of electronic corpora in the study of idioms and lexicography. This has been a long-established practice in English lexicography. For instance, *Collins Cobuild Dictionary of Idioms* (CCDI) is based on the *Bank of English*, now consisting of 500 million words. A major advantage of such large

corpora compared to hand-collected data is that they provide accurate information on the frequencies and distribution of idioms as well as their collocational potential and the forms they occur in. Electronic corpora have proved to be particularly useful in distinguishing the common and current use of idioms from marginal or obsolete ones. The authors of CCDI point out that "idioms are comparatively infrequent, and it is only by having a very large corpus that we have sufficient evidence to describe idioms accurately and with confidence" (2002: vi).

There is a large corpus available for Croatian – the Croatian National Corpus with some 101 million tokens (*http://www.hnk.ffzg.hr*). The Croatian National Corpus is still under construction, and although it is not yet balanced (70% of the texts are news paper texts), it nevertheless provides a very useful tool for language analysis. Despite this, it has not yet been systematically used in any of the recent monolingual general dictionaries or dictionaries of idioms. In Croatia there are two monolingual dictionaries of idioms: *Frazeološki rječnik hrvatskoga ili srpskog jezika* (Matešić 1982) and *Hrvatski frazeološki rječnik* (Menac et al. 2003). Both of them are based on data that have been manually collected from a range of sources, primarily contemporary prose and journalism. Studies of idiom variation in Croatian (e. g. Fink 1997; Kovačević/Mihaljević 2004) have relied on the dictionary citation-form as the canonical form, from which variations are derived and to which they can be related. However, corpus-based research on idiom variation has shown that there may be a certain discrepancy between the forms listed in the dictionaries and the data found in the Croatian National Corpus (Stanojević et al. 2009). Given that the corpus provides more comprehensive data than can ever be hoped to be collected manually, it would seem reasonable that the corpus has to be included as one of the primary sources.

A case in point is the idiom *loviti u mutnom*, which we will turn to in the remainder of the paper. In Menac et al. (2003: 181) this idiom is listed under MUTAN ('murky', adj.) as an expression with the structure VP + PP (*loviti u mutnom*, lit. to hunt in the murk-LOC) with the meaning explained as 'using suspicious circumstances to one's own advantage, getting rich (achieving success or other goals) in an illegal manner'. In Matešić (1982) idioms with very similar meaning, lexical make-up and structure (VP + PP and NP + PP respectively) are listed under three separate entries, i.e. as three individual expressions:

(2) a. MUTAN ('murky', adj.)
loviti u mutnom (lit. to hunt in the murk-LOC)
'to gain an unfair advantage from a difficult situation'
b. LOV ('hunt', n.)
lov u mutnom (lit. a hunt in the murk-LOC)
'an unfair advantage gained from a difficult situation'
c. VODA ('water', n.)
loviti (pecati) u mutnoj vodi (lit. to hunt (fish) in murky water-LOC)
'to try to gain an advantage from a difficult situation'

Examples in (2) illustrate several significant issues. Firstly, given that their meanings and lexical make-up are similar, it would be reasonable to assume that they are not separate expressions but specific realizations of a single schematic idiom. If this is the case, then the

following questions arise: What is the common motivating element and how is it related to the variants? Finally, this raises the issue of the criteria used in establishing the canonical form and the dictionary citation-form in terms of frequency and prototypicality.

4 Corpus Research and Results

In order to show that the expression *loviti u mutnom* is a schematic idiom with a relatively flexible structure and that variations reflect the conceptual motivation systematically, we performed a corpus study. First, we obtained a sample of 791 tokens of the adjective *mutan* ('murky') from the Croatian National Corpus. Since Croatian is a flective language, queries were built in such a way to identify all forms of the lexical items making up the idiom. The positive filter was used to construct complex queries. We looked for occurrences of two patterns (which are common in all the dictionary entries listed above): *u* ('in') and *mutan* ('murky') and *mutan* and *voda* ('water') within the span of 5 words with a two-way link in-between. We manually sorted the results, eliminating instances of *mutan* used in the literal sense. In this way we obtained a sample of 143 examples from the Croatian National Corpus.

Our results show that all the expressions share a common element which reflects the CONTAINER image schema.[2] In the vast majority of the examples (96%), this element is expressed by a locative prepositional phrase, an accusative prepositional phrase or a noun phrase in the instrumental case:

(3)

u mutnom	
in murky-LOC sg. n.	
u mutnoj	*vodi*
in murky-LOC sg. f.	water-LOC sg. f.
u mutnim	*vodama*
in murky-LOC pl. f.	water-LOC pl. f.
po mutnim	*vodama*
on murky-LOC pl. f.	water-LOC pl. f.
u mutne	*vode*
in murky-ACC pl. f.	water-ACC pl. f.
mutnim	*vodama*
murky-INSTR pl. f.	water-INSTR pl. f.

In 86% of the cases the common element is expressed by a phrase containing the preposition *u* ('in') followed by a nominal element in the locative case (*u mutnom* (81%), *u mutnoj vodi* (3%) and *u mutnim vodama* (2%)). The occurrence of the preposition *u* suggests that the CONTAINER schema plays the crucial role in the configuration (Rudzka-Ostyn 2003) – the preposition *u* is prototypically used with objects which are conceptualized as containers.

2 A general description of image schemas is given in Johnson (1987).

In Croatian the construction *u* + locative is used to refer to a situation in which one entity, the trajector, is contained within another, the landmark (Šarić 2008: 81).

The expressions containing the construction *po* + locative (*po mutnim vodama*) and the construction *u* + accusative (*u mutne vode*) specify the direction of a moving trajector relative to a landmark. More specifically, *po* + locative signals the movement of a trajector on the surface of a landmark, while the construction *u* + accusative expresses the movement of a trajector towards the interior of a landmark.

In Croatian and other Slavic languages (for Croatian see Silić/Pranjković 2005; for Russian see Janda 1993), spatial relations can be expressed by nouns in oblique cases without the use of prepositions, e. g. the instrumental case (*mutnim vodama*). The instrumental of space prototypically marks a location, i.e. the physical setting for an action described by a verb of motion (Janda 1993: 166).

As illustrated by the examples in (3), the element that the researched expressions have in common is a location which functions as a landmark. In the majority of cases (86%) it is expressed by a locative prepositional phrase, which specifies the position of the trajector relative to a specific location. The nominal part of the prepositional phrase (*mutno*) points to the interaction of the CONTAINER schema with the metaphor KNOWING IS SEEING.

Some (albeit few) corpus examples show that in fact the element *mutno* refers to an expanse of water. This is further supported by cross linguistic evidence from Russian and Polish, where the canonical forms of similar idioms explicitly mention water (Russian ловить рыбу в мутной воде (lit. to catch fish in murky water) (Birikh et al. 2005); Polish łowić ryby w mętnej wodzie (lit. to catch fish in murky water) (Bąba/Liberek 2002)). Moreover, the origin of the Russian expression can be traced back to the act of deliberately disturbing the water while casting nets because it clouds the vision of the fishes and makes them easy prey (Birikh et al. 2005: 617). The liquid – water – is conventionally metonymically construed as a container (CONTAINED FOR CONTAINER). Finally, Croatian further schematizes the expression by frequently leaving out the component *voda* ('water') and focusing on its 'murkiness'.[3]

Thus, the expression signals that a participant uses a difficult situation to achieve a goal by taking part in various activities, usually at the expense of other participants. This is facilitated by the fact that what goes on in the container (the activities that the trajector is involved in) is hidden from view. This complex ICM is the basic cognitive mechanism motivating all the expressions. Thus, the first stable element of the conceptual core is the blend of the CONTAINER schema (indicated by various grammatical and lexical items expressing location) and the metaphor KNOWING IS SEEING, signalled by the component *mutan*.

The CONTAINER core provides a natural location for an element to appear as the trajector within it, as in the following example:

3 One of the reasons why the component *voda* may be easily left out in Croatian (whereas this is not so in Russian and in Polish) is the development of the word *loviti* ('hunt', v.) and its role in the motivation of the idiom (see below).

Schematic element in the location	Number of examples	%
Relation in the container	118	83%
Participant in the container	19	13%
Container as participant	6	4%
Total	143	100%

Table 1: Types of trajectors in the landmark *u mutnom*.

(4) Smatraju kako je ovo kraj djelovanja onih koji su lovili u mutnom i jedan od najvažnijih projekata vlade RH.
'They believe that this is the end of all the activities by those who have been fishing in troubled waters (lit. hunting in the murk) and one of the most important projects of the Croatian Government.'

In this example the CONTAINER *u mutnom* 'in the murk' is the landmark for the relation *loviti* 'hunt', which is the trajector. Our analysis shows that the container is the landmark for relations in a majority of cases, followed by things as the trajector. In only a handful of cases the container is a participant which does not serve as the location of the trajector. The results are presented in Table 1.

The results show that in the first group of expressions the trajector is a relation between participants located in the container. The relation is expressed by a verb or a deverbal noun, as in (4) above and (5) below:

(5) Gradonačelnik je, međutim, potvrdio da Belomanastirci opravdano sumnjaju na igre u mutnom…
'However, the Mayor has confirmed that the citizens of the town of Beli Manastir have grounds to suspect foul play (lit. games in the murk')'

The relation in the container is construed dynamically as it involves two or more participants that are engaged in an activity. This is reflected by the choice of constituents (e. g. *loviti* 'to hunt' in (4) and *igre* 'games' in (5).

In the second group of expressions the trajector is a participant located in the container. Let us have a look at the following examples:

(6) Turizam u mutnim vodama
'Tourism in troubled waters (lit. in murky waters)'

(7) Takva kaotična i nepregledna situacija idealna je za lovce u mutnom…
'Such a chaotic and unclear situation is ideal for those who fish in troubled waters (lit. for hunters in the murk)'

The example in (6) describes a relation between a theme (*turizam* 'tourism') and a location (*u mutnim vodama* 'in murky waters'), which is a static variant of a prototypical location

Model	Number of examples	%
HUNTING/FISHING	111	81%
STRUGGLE/COMPETITION	15	11%
SWIMMING	9	7%
BEING IN A DIFFICULT SITUATION IS BEING IN A CONTAINER	2	1%
Total	137	100%

Table 2: Extensions and elaborations of the underlying ICM in *loviti u mutnom*.

schema. In (7) the prepositional phrase is a postmodifier in the noun phrase *lovci u mutnom* ('hunters in the murk') and thus serves not only to locate, but also to qualify a participant.

In the third group of expressions the container has been promoted to one of the central participant roles, that of *theme*. This is demonstrated by the fact that the expressions profiling the container may have the syntactic function of subject or direct object:

(8) …prosvjednici nastavljaju štrajkati glađu sve dok ne dobiju čvrsta jamstva da će se mutna voda u njihovu poduzeću konačno razbistriti.
'…the protesters will continue their hunger strike until they are given a firm guarantee that the muddied waters (lit. the murky water) in their company will finally be cleared.'

(9) upravo takvo neraščišćeno (najblaže rečeno) stanje i mutnu vodu koriste potencijalni kandidati u pokušajima da se cijene obore i dovedu do donjih granica…
'…potential candidates are taking advantage of this (to put it mildly) confusing state of affairs and muddied waters (lit. murky water) in their attempts to bring down prices to the lowest level.'

These examples involve a non-human participant, *mutna voda* ('murky water'), as their theme. In (8) the agent is defocused and the container is seen as a participant undergoing a process, while the sentence in (9) illustrates a transfer of energy: a human participant (*potencijalni kandidati* 'potential candidates') consciously and volitionally acts upon the theme. In both constructions the non-participant role *place* is fused with the role *theme* (Radden/Dirven 2007: 288). As a result, the container is seen as being affected.

Our results have confirmed that the variant realizations leave the combination of the CONTAINER schema and the KNOWING IS SEEING metaphor unaltered, while the lexical components and syntactic structure of the idiom are varied systematically to reflect the underlying ICM.

Our next assumption is that variations are not unpredictable but constrained by the cognitive mechanism(s) motivating the expressions. Therefore, we looked into the expressions profiling a relation and a participant to check for any systematic variability of lexical components reflecting the elaborations and extensions of the underlying ICM and/or combinations with other models. The results are presented in Table 2.

The most numerous group of expressions profiles the metaphorical model PURPOSEFUL ACTIVITY IS A HUNT: a human participant, conceptualized as a predator, takes advantage of a complicated, unclear situation to achieve a goal. The chances for success are increased because the activity occurs in a context that is highly favourable for the predator – his actions cannot be seen. The HUNTING frame is reflected by the following lexical components: *lov* ('hunt', n.), *loviti* ('hunt', v.) *hvatati* ('catch', v.), which focus on the activity itself and *lovac* ('hunter', n.), which focuses on the predator as the initiator of activities.

In modern Croatian the verb *loviti* and the noun *lov* refer to catching wild animals (Anić 2004). However, if combined with *riba* ('fish'), *loviti* and *lov* may refer to fishing (e.g. *loviti ribu* ("to catch fish"), *ribolov* ('fishing'), *ribolovac* ('fisherman')), which indicates that they were wider in scope and used to include both catching wild animals and fish.[4] This is confirmed by older dictionaries of Croatian (e.g. *Rječnik hrvatskoga ili srpskoga jezika* (1880-1976)).[5]

The variant realizations in the second group are based on the COMPETITION model: they describe the involvement of a human participant in activities that allow him to compete against other participants. In this group the trajector is also conceptualized as the aggressor since he ensures a positive outcome of a struggle/competition by employing an unfair strategy, i.e. his actions are hidden from the opponents' view. This model is signalled by the following items: *bogatiti se* ('get rich'), *tražiti novce* ('look for money'), *zaraditi* ('earn money'), *dobiti poene* ('win points') and *profiteri* ('profiteers').

In the third group the container is conceptualized as a liquid, more specifically water, which is signalled by verbs such as *plivati* ('swim') and *ploviti* ('sail') and the verbal noun *plivanje* ('swimming'). Furthermore, in this group the item *voda* ('water') is explicitly expressed as *u mutnoj vodi* ('in murky water-LOC'), *u mutnim vodama* ('in murky waters-LOC'), *po mutnim vodama* ('on murky waters-LOC'), *u mutne vode* ('in murky waters-ACC'), *mutnim vodama* ('murky waters-INSTR'). The trajector is thus seen as being partly surrounded by the landmark, moving on its surface or towards its interior.

The fourth group profiles the location of the trajector relative to the landmark, which is signalled by the use of the copular verb *be* and the locative prepositional phrase. The container encloses the trajector from several sides and as a result the trajector is hidden from view.

5 Discussion

Our findings show that the Croatian expression *loviti u mutnom* has a relatively variable structure and that the variant realizations of this idiom reflect the underlying conceptual motivation in a systematic way. The idiomatic meaning is motivated by a complex ICM, which is a combination of the CONTAINER schema and the conceptual metaphor KNOWING IS SEEING. The container is essentially a non-transparent liquid, which enables the trajector,

4 I am grateful to the editors for pointing this out to me.

5 Interestingly, the Russian verb ловить and Polish *łowić* had a different history, and now refers primarily to fishing (see Zgółkowa 1994-2004; Boryś 2005; Ozhegov/Shvedova 2005).

prototypically a human participant, to engage in activities in which he is conceptualized as a predator. What all the expressions have in common is the model MUTNA VODA ('murky water'), which profiles a complicated situation regarded as a favourable context for secret, (mostly) illegal activities. This model constitutes the conceptual core which triggers the meaning of the expression as a whole and serves as the basis for variations. The conceptual core is subject to different construals, namely static and dynamic relations in the container, participant(s) in the container and the container as a location. In most examples, the conceptual core is expressed by the locative prepositional phrase *u mutnom*, but it may also be explicitly expressed by constructions containing the item *voda* ('water'). In addition to the "regular" use of idioms, the conceptual core is also preserved in creative modifications of the expression in the discourse. As the following example of a literal use shows, the modified idiom is compatible with the model MUTNA VODA:

(10) Da političari ne love samo u mutnom, potvrdio je njihov športski susret održan na rijeci Kupi u Sisku gdje su odmjerili snage i umijeće u ribolovu.
'Politicians do not fish only in troubled water (lit. hunt in the murk), which is evident from the fact that they took part in a fishing contest on the River Kupa in Sisak.'

In fact, the expression *ne love samo u mutnom* (lit. do not hunt only in the murk) in (10) does not explicitly mention the component *voda* 'water'. Still, it is successfully combined with the literal meaning of fishing, which shows that the modification in question is not unpredictable. When the idiom is creatively exploited, the underlying ICM determines the syntactic and semantic properties of the lexical components, i.e. they must be compatible with the conceptual core of the idiom.

Our results confirm a correlation between the lexical make-up and syntactic structure of variations and conceptual motivation. The trajector, the landmark and their relation, which is determined by the ICM MUTNA VODA, may be further elaborated or extended in combination with other models (HUNTING, SWIMMING, etc.) in accordance with the meaning of the idiom. This is reflected in the choice of constituents, which is constrained by and dependent on the underlying ICM and its combination with other models.

Further research, in particular a diachronic study, is necessary to track the process of the institutionalization of the PP *u mutnom* as the salient component of the idiom. In addition, the results need to be tested on native speakers using psycholinguistic methods in order to check their cognitive relevance.

6 Conclusion

Based on the data from the Croatian National Corpus we have established that the expression *loviti u mutnom* is a schematic idiom with lexically open slots and that various realizations reflect the underlying conceptual motivation in a systematic way. Furthermore, corpus evidence shows that, when it comes to conventionalized lexical realizations, the variation with the form of a NP (*lov u mutnom*) is on a par with the VP variation (*loviti u*

mutnom). This proves that the VP idiom which is registered in Menac et al. (2003) as the canonical form is merely one of the variations occurring within the schema.

The results have also shown that the ICM MUTNA VODA, based on the combination of the CONTAINER schema and the conceptual metaphor KNOWING IS SEEING, constitutes the conceptual core of the idiom which constrains the various lexical and syntactic realizations, even though the component *voda* ('water') occurs relatively infrequently in the corpus. Its status as the scene-setter is confirmed, on the one hand, by lexical substitutions which reflect various PURPOSEFUL ACTIVITY models (e. g. SWIMMING) and Matešić's dictionary citation-form on the other (*loviti (pecati) u mutnoj vodi*; lit. 'to hunt (fish) in murky water').

On a more general level, we have shown that in establishing the canonical form we need to perform the grammatical and conceptual analyses of an idiom using a corpus. Our findings thus highlight the importance of the use of large electronic corpora in the study of idioms and lexicography. In other words, conclusions about the variability of idioms and the forms they occur in, which includes the distinction between prototypical and less prototypical uses, have to be properly data-founded.

Bibliography

Dictionaries and corpora

Anić, V. (2004): *Veliki rječnik hrvatskoga jezika.* Zagreb.
Birikh, A./Mokienko, V./Stepanova, L. (2005): *Russkaja frazeologija: istoriko-etimologicheskij slovar.* Moskva.
Boryś, W. (2005): *Słownik etymologiczny języka polskiego.* Kraków.
Bąba, S./Liberek, J. (2002): *Słownik frazeologiczny współczesnej polszczyzny.* Warszawa.
CCDI = Collins Cobuild Dictionary of Idioms (1995). Glasgow, ²2002.
Croatian National Corpus. http://www.hnk.ffzg.hr
Matešić, J. (1982): *Frazeološki rječnik hrvatskoga ili srpskog jezika*. Zagreb.
Menac, A./Fink-Arsovski, Ž./Venturin, R. (2003): *Hrvatski frazeološki rječnik*. Zagreb.
Rječnik hrvatskoga ili srpskoga jezika (1880-1976). Zagreb.
Ozhegov, S./Shvedova, N. (2005): *Tolkovyj slovar russkogo jazyka.* Moskva.
Zgółkowa, H. (ed.) (1994-2004): *Praktyczny slownik współczesnej polszczyzny.* Poznań.

References

Andrijašević, M./Zergollern-Miletić, L. (eds.) (1997): *Tekst i diskurs*. Zagreb.
Brdar, M./Omazić, M./Pavičić Takač, V. (eds.) (2009): *Cognitive Approaches to English: Fundamental Interdisciplinary and Applied Aspects*. Newcastle.
Cacciari, C./Glucksberg, S. (1991): "Understanding Idiomatic Expressions: The Contribution of Word Meanings". In: Simpson (ed.) (1991); 217-240.
Cacciari, C./Tabossi, P. (eds.) (1993): *Idioms: Processing, Structure, and Interpretation.* Hillsdale, NJ.
Cignoni, L./Coffey, S./Moon, R. (2002): "Idiom Variation in English and Italian: two corpus-based studies", in: *Languages in Contrast, 2/2*; 279-300.
Cowie, A. P./Mackin, R./McCaig, I. R. (1985): *Oxford Dictionary of Current Idiomatic English. Volume 2.* Oxford.
Croft, W./Cruse, A. D. (2004): *Cognitive Linguistics.* Cambridge.

Fillmore, C. J./Kay, P./O'Connor, M. C. (1988): "Regularity and idiomaticity in grammatical constructions: The case of *let alone*", in: *Language, 64/3*; 501-538.
Fink, Ž. (1997): "Frazeološke igre u reklamama ili Misli li četkica za zube svojom glavom". In: Andrijašević/Zergollern-Miletić (eds.) (1997); 325-330.
Gibbs, R. W. (1990): "Psycholinguistic studies on the conceptual basis of idiomaticity", in: *Cognitive linguistics, 1/4*; 417-451.
Gibbs, R. W. (1994): *The Poetics of Mind: Figurative Thought, Language, and Understanding.* Cambridge/New York.
Gibbs, R. W./Nayak, N. P. (1989): "Psycholinguistic studies on the syntactic behaviour of idioms", in: *Cognitive Psychology, 21*; 100-138.
Gibbs, R. W./Nayak, N. P./Cutting, C. (1989): "How to Kick the Bucket and Not Decompose: Analyzability and Idiom Processing", in: *Journal of Memory and Language*, 28; 576-593.
Gibbs, R. W./ Nayak, N. P./Bolton, J. L./Keppel, M. E. (1989): "Speakers' assumptions about the lexical flexibility of idioms", in: *Memory and Cognition*, 17; 58-68.
Gibbs, R. W./O'Brien, J. (1990): "Idioms and mental imagery: the metaphorical motivation for idiomatic meaning", in: *Cognition, 36/1*; 35-68.
Glucksberg, S. (1993): "Idiom meanings and allusional content". In: Cacciari/Tabossi (eds.) (1993); 3-26.
Granić, J. (ed.) (2004): *Semantika prirodnog jezika i metajezik semantike*. Zagreb-Split.
Janda, L. (1993): *The Czech Dative and the Russian Instrumental*. Berlin/New York.
Johnson, M. (1987): *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* Chicago/London.
Kovačević, B./Mihaljević, M.(2004): "Frazemi u publicističkome funkcionalnom stilu". In: Granić (ed.) (2004); 394-404.
Kövecses, Z./Szabó, P. (1996): "Idioms: a view from cognitive semantics", in: *Applied Linguistics, 17/3*; 326-355.
Lakoff, G. (1987): *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* Chicago/London.
Langacker, R. W. (1987): *Foundations of Cognitive Grammar. Vol. 1.* Stanford.
Langlotz, A. (2006): *Idiomatic Creativity.* Amsterdam/Philadelphia.
Moon, R. (1998): *Fixed Expressions and Idioms in English. A Corpus-Based Approach.* Oxford.
Nunberg, G./Sag, I. A./Wasow, T. (1994): "Idioms", in: *Language, 70/3*; 491-539.
Omazić, M. (2003): *Modifications of Phraseological Units in English.* Unpublished PhD Thesis, University of Zagreb. Zagreb.
Radden, G./Dirven, R. (2007): *Cognitive English Grammar.* Amsterdam/Philadelphia.
Rudzka-Ostyn, B. (2003): *Word Power: Phrasal Verbs and Compounds: A Cognitive Approach.* Berlin/New York.
Silić, J./Pranjković, I. (2005): *Gramatika hrvatskoga jezika.* Zagreb.
Simpson, G. B. (ed.) (1991): *Understanding Word and Sentence.* Amsterdam, New York/Oxford/Tokyo.
Stanojević, M.-M./Parizoska, J./Banović, L. (2009): "Schematic idioms and cultural models". In: Brdar/Omazić/Pavičić Takač (eds.) (2009); 321-344.
Šarić, Lj. (2008): *Spatial Concepts in Slavic. A Cognitive Linguistic Study of Prepositions and Cases.* Wiesbaden.

The Canonical Form in Murky Waters

Jelena Parizoska
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10 000 Zagreb
Croatia
jparizo@ffzg.hr

II

Methodische Probleme und Tools in der computergestützten Phraseologie-Forschung

Semi-Automatic Retrieval of Phraseological Units in a Corpus of Modern Norwegian

Ruth Vatvedt Fjeld/Lars Nygaard/Eckhard Bick

Trotz des großen Interesses an der Erfassung und Beschreibung von Kollokationen in der Lexikographie der letzten 20 Jahre gibt es nur wenige Werkzeuge, die eine systematische, korpusbasierte Kollokationsforschung ermöglichen. In diesem Beitrag beschreiben wir das neue Verfahren DeepDict, mit dem relevante Mehrworteinheiten für ein norwegisches Kollokationswörterbuch systematisch unter Einbeziehung von syntaktischen und semantischen Kriterien im Korpus gefunden und für die lexikographische Arbeit aufbereitet werden können.

1 Introduction

The study and description of phraseological unit has received much attention in lexicography during the last 20 years (e. g. Kilgarriff/Tugwell 2002, Fontenelle 1994, Olsen 2000, Malmgren 2008). Corpus-based lexicography has shown clearly how some words cling to a limited set of other words. Of course this tendency has been known and given attention throughout the whole period of modern lexicography, but tools to do systematic investigations have been missing. Rules for multi-word constructions are known in modern lexicography as collocations, traditionally they have been called phrases, idioms or formulas. The term phraseological units is used in this article as a hypernym, and we shall not go into any detailed taxonomy of the different types of these units. Restrictions on phraseological units might be morphological, syntactic or semantic. Collocation restrictions are to be separated from restrictions in building free combinations. In fact there are no absolutely free combinations of words in a language; there are always some syntactic or pragmatic restrictions between words that might be combined, according to how the world in fact is, and to general grammatical rules. Selection restrictions might however be accounted for by means of general semantic rules or lexical properties (Cowie 1994/2008)

There are two main types of restrictions to which words fit into word-combinations: encyclopaedic and grammatical. In addition there are many unpredictable semantic and grammatical restrictions that are difficult to find any systematic rules for. Collocations are constructed from occasional rules that often seem to give arbitrary restrictions. Since these restrictions are, to a great extent, purely conventional, collocations often are language specific and hence problematic in translation, language learning etc.

Routine formulas, as a special type of phraseological units, are governed by strong collocation restrictions, known as collocability. Greetings and polite wishes are strongly conventionalised and different from language to language. It is difficult to find grammatical or semantic rules that could explain the difference between the English formula *Merry Christmas* and the Norwegian *God Jul* (literally "good Christmas"). Independent of the words in the collocations, the formula expresses a wish for joy and happiness and good feelings for a coming celebration. To wish a Norwegian *Lykkelig jul*, or an English-speaking person *Good Christmas* might lead to misunderstandings or indulgence from the receiver of the wish. The same difference is seen in Norwegian *God påske!* and Swedish *Glad påsk!* for *Happy Easter!* even though these languages are closely related. Even the two standards of Norwegian have their own collocability, as different dialects also might have.

The limitations of a collocation are arbitrary, especially when one of the constituents is used in a well-known figurative sense or as a dead metaphor (cf. Cowie 1994/2008: 165). In a free combination there might be a part-for-part-substitution of the single words in a phrase, which is not possible or very restricted without changing the meaning in a collocation, even making it meaningless.

2 Collocations in Norwegian Lexicography

Phraseological units in general and especially collocations have been given relatively little attention in Norwegian lexicography. There do not exist any phraseological dictionaries or dictionaries of collocations, not even a dictionary of valency for the Norwegian language, except for an extensive overview of phrases in the Bokmål standard which are not allowed in the Nynorsk standard, called *Med andre ord* ('In other words', Rommetveit 1993).

Although there is no Norwegian dictionary of collocations, in most lexicographical work a large amount of phrases are recorded. The Norwegian language has two standards, Dano-Norwegian standard called bokmål and the Neo-Norwegian standard called nynorsk. The majority standard is bokmål, used by over 90% of the population. There are two major monolingual dictionaries for Norwegian, *Norsk Riksmålsordbok* (1937-57), documenting the bokmål standard (under revision in *NAOB – Det Norske Akademis Store Bokmålsordbok*) and *Norsk Ordbok* (1966-) documenting nynorsk. In both dictionaries, multi-word constructions are accounted for as sublemmas under one-word lemmas. Bilingual dictionaries also have to describe phraseological differences between Norwegian and the second language.

Until recently there has been an obvious need for methods and theory concerning phraseological units in lexicography. With new tools and theory it is an interesting task to review the recording of multi-word phrases in the traditional dictionaries.

3 Semi-automatic retrieval of collocations

Computerized technology has provided the lexicographer with methods for finding the collocations in a language. Using large amounts of linguistic data in electronic text corpora it is possible to retrieve collocations automatically. The DeepDict Lexifier (Bick 2009) has been developed for this task, and has been adapted for several languages, including Norwegian.

This tool automatically creates context overviews for a given word, supporting the lexicographer in building complex dictionary entries. Word relations are based on Constraint Grammar dependency analysis and grammatical functions, not just co-occurrence. Relative and absolute frequency values are provided for each relation.

In its dynamic "lexicograms", DeepDict strives to identify the most characteristic and most semantically relevant relations, not just chains of words. Therefore real syntactical relations like subject-verb, verb-object, modifier-noun etc. are used instead of ordinary collocations (n-grams). The corpus material for a DeepDict dictionary is therefore subjected to a deep syntactic (and partially semantic) analysis using so-called Constraint Grammar dependency parsers. Relation pairs can then be extracted as "dep-grams" rather than "n-grams", linking the semantic head words of syntactic phrases independently of their distance in the sentence.

For a verb like 'eat', this would result in dep-grams like the following:

PROP_SUBJ → eat_V
cat_SUBJ → eat_V
apple_ACC → eat_V
mouse_ACC → eat_V

With little further processing, the result can be represented as a summary "entry" for *eat* in the following way:

{PROP, cat, <hum>, … } SUBJ → eat ← {apple, mouse, <fruit>, … } ACC

Obviously, the fields in such an entry would quickly be diluted by the wealth of corpus examples, and one has to distinguish between typical complements and co-occurrences on the one hand, and non-informative "noise" on the other. Therefore, a statistical measure for co-occurrence strength is introduced to filter out the relevant cases, normalizing the absolute count for a pair a → b against the product of the normal frequencies of a and b in the corpus as a whole:

$$C \log \frac{p(a \rightarrow b)^2}{p(a) * p(b)} \tag{3.1}$$

where p() are frequencies and C is a constant introduced to place measures of statistical significance in the single digit range.

The resulting database would then contain, for each dep-gram pair, both its absolute frequency, co-occurrence strength, as well as an index of relevant sentence ID's in the source corpus for concordancing example sentences for a given relation. Even for a single

corpus in a single language, parsing all corpus material and creating the databases, may take days or weeks, and the resulting datasets are so big (currently 90 GB for all DeepDicts together) that querying them in a straight-forward fashion would cause unacceptable delays to the devised.

From the point of view of a lexicographer, one advantage of DeepDict is that it shows all identified relations for a given word in one coherent graphical overview page, rather than the hundreds of individual concordances and statistics necessary to extract the same information from a classical corpus search interface. To further enhance this "overview effect", certain levels of abstraction were introduced. Such features are the use of the dummy-"words" PROP and NUM for names and numbers and the inclusion of pronouns, which help to abstract traces like +HUM (he/she), countability and mass. Finally, semantic prototypes were used for nouns, comprising categories like

<Hprof> (human professional, e.g. "carpenter"), <food> or <tool>.

4 DeepDict-retrieval of collocations compared to traditionally excerpted collocations

4.1 Traditionally excerpted collocations

Leksikalsk bokmålskorpus (LBK) is a 40 million word balanced corpus of modern Norwegian texts in the bokmål standard. The corpus texts are composed according to recorded reading of different text types among Norwegians (Time use survey 2000). The corpus is POS-tagged and each text is annotated with bibliographic and ethnographic information. The whole corpus is analysed using DeepDict, which gives an overview of potential collocations in the corpus. In this article we will account for a few examples of similarities and differences between the two methods.

Constructions from a delexicalised verb + noun + preposition is a special type of collocation. Typical examples are *gi plass til* (give space for), *ha mulighet til* (have the possibility to, be able to). Delexical verbs are known by lexicographers as "the large verbs", in the sense that they have a lot of restricted combination possibilities with lexicalised meaning.

One of these verbs is *å komme/koma* ('to come'), which might have several hundreds of sublemmas in large monolingual dictionaries. In *Norsk Ordbok* there are numbered 96 sublemmas, which all have a large amount of examples or subsublemmas each. All in all there are several hundred senses and subsenses described for this verb. To edit such a definition is a time consuming and difficult task.

The most extensive modern bilingual dictionary for Norwegian, *Engelsk-norsk stor ordbok* (*NEO*, Kunnskapsforlaget 2008), also has a long article for the verb *to come*, surprisingly also with 96 sublemmas or multi-word phrases defined. A closer comparison between the definitions of the verb *to come* in these two dictionaries show that only 33 of their sublemmas were concurrent. This discrepancy is noteworthy and might confirm the claim that there has been a lack of methodology in this field. We wanted to control these two quite different descriptions, using the DeepDict tool. Several sublemmas in NEO

seemed to be relevant also in a monolingual dictionary, as *komme bakpå* ('fall behind'), *komme igjen* ('come again') and *komme til bunns i* ('get to the bottom of'), but they were not lemmatised in *Norsk Ordbok*.

A comparison with a Swedish construction dictionary, *Svenskt språkbruk* (Svenska Språknämnden 2003) also showed a lot of multi-word phrases attached to the verb *to come*, that were not in *Norsk Ordbok*: *komma här* ('come here') *komma för sent/tidigt* ('bee too early/late'). The Swedish dictionary listed at least 23 phrases that were not to be found in *Norsk Ordbok*. A closer look at *Norsk Ordbok* shows that some of the missing phrases occur under the collocator or as a normal confirming example or attestation of the lemma (e.g. *let (aldri) nokon koma bakpå seg* under the lemma *bakpå*, with the meaning 'take somebody by surprise', or *koma til kort* in the meaning 'fall short' as an example under the adjective *kort* 'short').

The access structure of the dictionary is therefore an extra complication: Some of the collocations are listed as sublemmas, some as confirming examples of the main lemma. As sublemmas, there seems to be no rule to whether they are listed under their basis or their collocator. It is difficult to see any systematic approach, and the two dictionaries seems to have lemmatised multi-word phrases more or less at random.

The lack of direct equivalents of phraseological units in English shows that they are quite strongly lexicalised and these phrases also show that these phrases should be listed as sub-lemmas in a monolingual descriptive dictionary.

4.2 DeepDict retrieval of collocations

A DeepDict analysis of *å komme* 'to come' based on the 40 mill. word text corpus, gives a total of 500273 relations. The even more delexicalised verb *to have* has a total of 1105180 relations. To go through all these relations manually and sort out the candidates for sublemmas, would have been a difficult and time-consuming work.

DeepDict, however, shows several types of relations and frames. In this study we only investigate the adverb-verb-collocations, which in DeepDict are:

- Free temporal, locative and modal adverbs
- Valency dependent adverbial complements
- Verb-integrated particles.

Most relevant in this comparison between the traditional lexicographical descriptions are the verb-integrated particles. The result from DeepDict for the verb *to come* is:

> **Verbal particles:**
>
> 10.33:9 **hjem** · 9.29:9 **inn** · 9:9 **ut** · 8.28:9 **fram** · 7.74:9 **bort** · 7.69:9 **opp** · 7.55:9 **frem** · 6.68:9 **ned** · 7.12:8 **an** · 4.53:8 **sammen** · 6.08:6 **til** · 4.64:6 **rett** · 2.83:5 **med** · 2.34:5 **inne** · 1.91:4 **nedover** · 0.8:5 **for** · 1.98:3 stille · 1.98:2 hjemover · 0.45:2 hen

This means that the strongest relation between to come and an adverbial in the analyzed corpus is *komme* and *hjem* ('come home'), the second strongest relation between *komme* and *inn* and so forth.

DeepDict also provides the most frequent prepositional phrases for each verb. The analysis of å *komme/koma* ('to come') shows that *til* ('to') has the strongest collocational relation to the verb, followed by *med* ('with'), *fra* ('from') and *for* ('for').

Furthermore the complements are listed for each preposition, which shows that the strongest relation between *komme+til* ('come+to') is *syne* which gives the collocation *komme til syne* ('appear', "come in(to) view"), followed by *uttrykk* ('expression'), which gives the collocation *komme til uttrykk* ('be expressed'). None of these multi-word phrases are registered as sublemmas in *Norsk Ordbok*, and the same goes for several of the strongest relations shown by DeepDict, such as *komme til gode* ('give benefit'), *komme til unnsetning* ('come to someone's rescue/aid'), *komme til kort* ('fall short'), *komme til stykket* ('after all, actually'). This is partly due to the access structure in the dictionaries.

5 Conclusion

A preliminary study of multi-word phrases with one single verb suggests that the description of multi-word phrases in two Norwegian major dictionaries contain some shortcomings. In spite of an ample amount of phrases listed as sublemmas under the verb *å komme/koma* ('to come') in the two dictionaries, only one third of them are overlapping, which gives a hint that they might have been chosen at chance. This assumption is supported by a comparison of a Swedish dictionary of constructions, where they actually are lemmatised.

The descriptions of multi-word phrases with the delexicalised verb *komme/koma* in the two dictionaries are compared to the result of an automatic collocation retriever called DeepDict. DeepDict seems to be a valuable tool for investigating multi-word phrases and constructions in Norwegian. According to its statistics, it does not only show which words tends to appear together, it also shows which words are most likely to have a lexicalised use and meaning.

In this article we have only focused on delexicalised verbs with adverbial complements, and only for a single verb. In the future, we would like to do a more systematic study of a selection of words and grammatical phenomena, compared to a wider selection of dictionaries.

Bibliography

Bick, E. (2009): "DeepDict – A Corpus-based Dictionary of Word Relations". In: *Proceedings of Nodalida 2009, Odense, Denmark. NEAL proceedings series, Vol. 4.*

Corréard, M.-H. (ed.) (2002): *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Göteborg.

Cowie, A. P.(1994/2008): "Phraseology". In: Fontenelle (ed.), 163-168.

Fontenelle, Th. (ed.) (1994/2008): *Practical lexicography*. Oxford.
Kilgarriff, A./Tugwell, D. (2002): "Sketching Words". In: Corrérd (ed.), 125-135.
Malmgren, S.-G. (2008): "Collocations in Swedish Dictionaries and Dictionary Research". In: *Lexicographica* 24, 149-158.
Olsen, T. R. (2000): *Om fraseologiske enheter i norske ordbøker*. (Unpublished master thesis, INL 2000).

Dictionaries

Engelsk stor ordbok. Engelsk-norsk/Norsk-engelsk (2001). Oslo.
NAOB – Det Norske Akademis Store Bokmålsordbok (in prep.). Oslo.
Norsk Ordbok (1966 –). Oslo.
Norsk Riksmålsordbok (1937-1957). Oslo.
Rommetveit, M. (2007): *Med andre ord. Den store synonymordboka med omsetjingar til nynorsk.* 3. ed. Oslo.
Svenskt språkbruk. Ordbok över konstruktioner och fraser (2003). Ed. by Svenska Språknämnden, Stockholm.

Statistics Norway: *http://www.ssb.no/english/subjects/00/02/20/tidsbruk_en/*

Ruth Vatvedt Fjeld
University of Oslo
Institutt for lingvistiske og nordiske studier
Leksikografi og Målføregransking
Postboks 1021 Blindern
0315 Oslo
Norway
r.e.v.fjeld@iln.uio.no

Eckhard Bick
Institut for Sprog og Kommunikation
Syddansk Universitet Odense
Campusvej 55
5230 Odense
Denmark
eckhard.bick@mail.dk

Einsatz von Sketch Engine im Korpus – Vorteile und Mängel

Peter Ďurčo

The present paper deals with the usage of the statistical tool ‚sketch engine' by the compilation of collocation profiles for the first Slovak dictionary of collocations. We discuss shortly also the advantages and disadvantages of stochastic methods in corpus linguistic.

1 Einleitung

Das Slowakische Nationalkorpus (SNK)[1] enthält über eine halbe Milliarde Textwörter und verfügt über das Corpus Query System Word Sketch Engine[2], das die Analysen von Wortkookkurrenzen und von grammatischen Relationen in einem Programm integriert und Kookkurrenzdifferenzen sowie distributionelle Thesauri von sprachlichen Einheiten in lemmatisierten und morphologisch markierten Korpora generiert. Dieses System wird intensiv in verschiedenen lexikographischen Projekten des Instituts für Sprachwissenschaft der Slowakischen Akademie der Wissenschaften genutzt. Mit Hilfe dieses Instruments entsteht auch das erste Kollokationswörterbuch des Slowakischen.[3] SNK benutzt den Korpusmanager Manatee mit dem Benutzer-Interface Bonito (vgl. Rychlý 2007).[4] Dieser Korpusmanager ermöglicht statistische Berechnungen von Häufigkeiten und Kookkurrenzen nach verschiedenen statistischen Parametern wie absolute Frequenz, MI-score, T-score, MI3, log likelihood etc. Manatee wird auch vom integrierten statistischen Modul Sketch Engine benutzt.

Tabelle 1 zeigt, wie stark sich die Ergebnislisten von ermittelten Kookkurrenzen zu einer Basis voneinander unterscheiden können. Für das Lemma *pieseň* (Lied) haben wir die Identität der mithilfe verschiedener statistischen Maße gelieferten Kollokate innerhalb der ersten 50 Zeilen im SNK verglichen. Diese Tabelle zeigt, dass die größte Übereinstimmung zwischen den Maßen MI3 und salience besteht, die größte Differenz liegt dagegen in den Wortlisten zwischen T-score und MI-score.

Für die Korpuslinguistik und die korpusgestützte Lexikographie ergeben sich daraus zwei grundlegende Fragen: Erstens, welche sprachlichen Phänomene werden durch ver-

1 *http://korpus.juls.savba.sk/*: Stand Juni 2009.
2 *http://www.sketchengine.co.uk/*
3 *http://www.vronk.net/wicol/index.php/Main_Page*
4 *http://nlp.fi.muni.cz/raslan/2007/papers/12.pdf*

	Abs. Freq	T-score	MI	MI3	log likelihood	min. sensitivity	salience
Abs. Freq	–	73,91	5,38	5,38	54,94	34,41	36,56
T-score	73,91	–	5	53,61	70,1	62	44,44
MI	5,38	5	–	32,65	24,49	14	36
MI3	5,38	53,61	32,65	–	75	57,14	81,63
log likelihood	54,94	70,1	24,49	75	–	71,43	69,39
min. sensitivity	34,41	62	14	57,14	71,43	–	60
salience	36,56	44,44	36	81,63	69,39	60	–

Tabelle 1: Identität der gelieferten Kollokate zum Lemma *pieseň* (Lied) im SNK durch verschiedene statistische Maße innerhalb der ersten 50 Zeilen

schiedene statistische Maße eigentlich erfasst? Zweitens, wie kann man diese heterogenen Ergebnisse zu einem integrierten Konglomerat von linguistisch und lexikographisch verwertbaren Daten vereinen?

2 Sketch Engine

Bei der Software Sketch Engine handelt es sich um ein statistisches Instrument mit seinen Vor- und Nachteilen. Alle statistischen Methoden analysieren nämlich die Kookkurrenzen von Wortketten entweder als Nachbarschafts-, Satz- oder Fensterkombinationen, d. h. es können die linke und rechte Kontextbreite und Schwellenwerten für minimale Vorkommen der untersuchten Einheiten bestimmt werden. Dabei stellt sich die Frage, inwieweit es möglich ist, zuverlässige und lexikographisch verwertbare Informationen über das distributionelle Verhalten der Textwörter in Korpora durch die Kookkurrenzermittlungen zu gewinnen?

Die statistischen Methoden haben unbestreitbare Stärken (z. B. Vorstrukturierung von Massendaten, Erkennen von Usus-Phänomenen, die bei manueller Bearbeitung verschlossen bleiben), die linguistische Interpretationsleistung bleibt hingegen immer in den Händen der Linguisten oder Lexikographen.

Die grundlegenden Mängel der sprachstatistischen Verfahren liegen darin, dass verschiedene statistische Maße immer andere, diffus überlappende Teilmengen von Wortlisten mit unterschiedlichen Präferenzen liefern. Außerdem werden lange redundante Wortlisten mit so genannten statistisch unspezifischen Phänomenen generiert, die noch nichts über den Typ der Beziehung zwischen Wortformen, die ein Kookkurrenzpaar bilden, aussagen. Die Interpretation des Linguisten ist hier unverzichtbar. Dabei entsteht jedoch das Problem, dass statistisch nicht signifikant nicht zugleich linguistisch irrelevant bedeutet. Deswegen werden zum Optimieren und zur Interpretation der extrahierten Mengen hybride Verfahren

mit lexikalischen, kategoriellen und kookkurrierenden Filtern entwickelt (vgl. Heyer et al. 2006: 247-254).

Die Anwendung von statistischen Analysemethoden hat den Nachteil, dass individuelle Eingriffe in den vorprogrammierten Prozess der Datenerhebung in verschiedenen Stadien nur beschränkt oder überhaupt nicht möglich sind.

3 Die word-sketch Regeln

Im Unterschied zu anderen Methoden der statistischen Ermittlung von Textwörtern im Korpus liegt der größte Vorteil von Sketch Engine darin, dass wir Tabellen mit Kollokaten in allen vordefinierten grammatischen Relationen zur untersuchten Basis bekommen. Sketch Engine für Slowakisch arbeitet mit einem Set von 36 Regeln. Es handelt sich um vier Kategorien von Regeln: unäre, symmetrische, duale und trinäre Regeln:

Ein Beispiel für eine unäre Regel:

```
*UNARY
# dative
=datX
1:[tag="N...3.*"]
```

Diese Regel ermittelt alle Substantive im Dativ.

Die symmetrischen Regeln untersuchen Elemente im rechten und linken Kontext der Basis zugleich. Das vordefinierte Intervall des untersuchten Kontextes ist standardmäßig auf {0,2} eingestellt, d. h., das Programm analysiert bis zu zwei Elementen links und rechts von der untersuchten Basis. Nur bei Konjunktionen ist der untersuchte Kontext {0,1}.

Es folgt ein Beispiel für eine symmetrische Regel: Diese ermittelt den engsten linken/rechten Kontext bei Konjunktionen:

```
*SYMMETRIC
# any left context + conjunction/conjunction + any right context
=Y Cj X/X Cj Y
2:[tag!="Z.*"] [tag!="Z.*"]{0,1} [tag="J.*"|word=","]
  [tag!="Z.*"]{0,1} \
1:[tag!="Z.*"] & 1.gtag=2.gtag
```

Die nächste symmetrische Regel ermittelt im linken und rechten Kontext zur Basis X das Vorkommen von Verben im Intervall {0,2}:

```
# verb + X/X + verb
=Vb X/X Vb
2:"V.*" [tag!="[JRZ].*"]{0,2} 1:[tag!="Z.*"]
1:[tag!="Z.*"] [tag!="[JRZ].*"]{0,2} 2:"V.*"
```

Duale Regeln ermitteln heterogene Elemente im linken und rechten Kontext zur Basis X im vordefinierten Intervall. Die nachfolgende Regel ermittelt z. B. beliebige Elemente im linken/rechten Kontext zu X:

```
*DUAL
# almost any + X/X + almost any
=Y X/X Y
2:[tag!="[ADNVZ].*"] [tag!="[JRZ].*"]{0,2} 1:[tag!="[ADNVZ].*"]
```

Trinäre Regeln ermitteln zwei heterogene Elemente im Kontext von X, wobei der Kontext entweder bidirektional oder aber nur als rechter bzw. als linker Kontext definiert werden kann. Die folgende trinäre Regel ermittelt im linken Kontext zu X alle Präpositionen mit einem beliebigen Zwischenelement:

```
*TRINARY
# preposition + any + X
=%s Y X
3:[tag="[R].*"] [tag!="Z.*"]{0,1} 2:[tag!="Z.*"]
  [tag!="Z.*"]{0,1} 1:[tag!="Z.*"]
```

Das Problem von word sketch liegt darin, dass durch das Set von 36 Regeln auch redundante grammatische Relationen generiert werden können, z. B. die Regel `any + X/X + any` generiert Kombinationen des Typs `Adjektiv + Verb`, oder `Adverb + Substantiv`, was eher Kolligationen (also grammatisch zusammenhängende aber nicht begrifflich gebundene Wortkombinationen) darstellt und keine Kollokationen belegt. Das zweite Problem liegt darin, dass word sketch isolierte Lemmata in der jeweiligen grammatischen Relation angibt und nicht die reale syntaktische Struktur. Dieses Problem wird jedoch durch den Umschaltemodus zwischen den abgebildeten Daten in Tabellenform von word sketch zu KWIC-Form, also zu Textteilen in Zeilenlänge, eliminiert.

4 Anwendung

Bei der Erstellung eines Kollokationsprofils wird für die untersuchte Basis ein word sketch erstellt. Um alle relevanten Kollokate möglichst vollständig zu erfassen, werden die Schwellenwerte für minimale Häufigkeiten der Kollokate standardmäßig niedrig eingestellt (normalerweise wird das minimale Vorkommen im gesamten Korpus auf 4 Vorkommen eingestellt). Die Zahl der Kollokate innerhalb einer grammatischen Relation wird gegenüber der Standardeinstellung von 32 auf 150 eingestellt, damit auch die statistisch wenig signifikanten, jedoch linguistisch und lexikographisch wichtigen Daten (z. B. seltene lexikalisierte Wortverbindungen, die aus niedrig frequenten Einheiten bestehen) erfasst werden.[5]

5 Zum verwendeten statistischen Modell in Sketch Engine s. *http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/ske-stat.pdf?format=raw*

Vb X/X Vb	272 036	0,9	Aj X	160 060	2,6	Sb X	63 172	0,3	X Sb	108 885	0,4
tráviť	4128	8,61	krátky	13 435	9,38	odstup	2135	9,1	uzávierka	642	7,08
nastať	2548	7,49	voľný	10 754	9,31	dostatok	1712	8,37	trvanie	513	6,71
stráviť	2078	7,31	dlhý	16 975	9,05	trávenie	568	7,93	socializmus	428	6,38
nadísť	1230	7,15	hrací	2302	8,47	plynutie	435	7,7	čakanie	336	6,27
venovať	2911	6,98	vysielací	1588	8,13	skrátenie	341	7,11	vojna	1752	6,23
plynúť	1213	6,97	dohľadný	1426	8,05	krátkosť	313	7,1	sláva	477	6,23
nemať	7903	6,69	riadny	2001	8,03	uplynutie	328	6,82	vznik	932	6,15
strácať	1321	6,67	pracovný	5865	7,63	centrum	2085	6,63	kríza	632	6,03
potrebovať	2022	6,3	dávny	1295	7,51	využívanie	404	6,32	ríša	391	5,89
ukázať	1821	6,28	posledný	12 408	7,5	väčšina	1602	6,2	pôsobenie	535	5,67
uplynúť	747	6,26	určitý	3858	7,44	nedostatok	778	6,14	obed	309	5,57
nestrácať	663	6,22	istý	7478	7,23	strata	875	6,06	dovolenka	330	5,21
dozrieť	597	6,08	oddychový	781	7,21	hľadanie	411	6,05	príchod	377	5,14
nájsť	2967	6,07	rovnaký	3249	7,06	otázka	2659	6,04	návšteva	482	4,92
krátiť	570	6,03	blízky	3664	6,96	postup	994	5,85	odchod	394	4,9
prísť	4019	5,98	celý	9018	6,73	využitie	368	5,79	príprava	539	4,85
ušetriť	642	5,96	reálny	1282	6,62	zub	340	5,78	konanie	404	4,67
skrátiť	527	5,85	týždenný	548	6,54	dĺžka	353	5,72	smrť	335	4,01
vyžadovať	869	5,76	miestny	1612	6,49	priebeh	564	4,99	vláda	660	3,53
meniť	942	5,76	dnešný	1701	6,45	hodina	919	4,99	zmluva	309	3,45

Tabelle 2: Word sketch zum Lexem *čas* (Zeit) im SNK

Im folgenden Beispiel (siehe Tabelle 2) wurde aus Platzgründen die Einstellung von Elementen in einer grammatischen Relation auf 20 Elemente und die minimale Häufigkeit im Korpus auf 300 reduziert. Word sketch zeigt die Relationen zum Lemma *čas* (Zeit) im SNK. Aus Platzgründen werden nur die Relationen zu vier Regeln angezeigt: 1. `Verb + X/X + Verb`, 2. `Adjektiv + X`, 3. `Substantiv + X`, 4. `X + Substantiv`. Die erste Spalte zeigt Kollokate des vordefinierten Typs, die zweite Spalte die absolute Häufigkeit der Kollokate im Korpus und die dritte Spalte die Signifikanz nach dem statistischen Maß.

5 Beispiel

Im Slowakischen Kollokationswörterbuch werden im ersten Schritt Kollokationsprofile für die ca. 500 häufigsten slowakischen Substantive erstellt. Die durch word sketch ermittelten Kollokate werden mit Daten aus Lexika und aus dem Web (über die Suchmaschine Google) ergänzt, weil dadurch weitere lexikographisch relevante Kollokationen gefunden werden können, die durch statistische Methoden nicht ermittelbar sind. Angeordnet werden die Kollokate in einer Schablone, die eine universelle Matrix von binären Relationen für alle auch nur theoretisch denkbaren Wortformenkombinationen der Basis und des Kollokators darstellen (vgl. Ďurčo 2007). Nichtexistente Wortformenkombinationen für das konkrete Lemma bleiben unbesetzt. Mehrgliedrige Kollokationen werden auf binäre Strukturen zurückgeführt und zu entsprechenden Typen zugeordnet. Die Bedeutungen eines polysemen Lexems werden am Anfang des Eintrages separat angegeben und in der Matrix nicht berücksichtigt, es sei denn, eine konkrete Kollokation erweist sich als polysem, dann wird dies durch Ziffern 1. 2. … angegeben. Das gesamte Wörterbuch wird in einer Media-Wiki Anwendung konzipiert und editiert.

Der Wörterbucheintrag hat folgende Struktur:[6]

1. Lemma
2. Bedeutungen
3. Kollokationen
4. Links zum Lemma in digitalen Lexika
5. Links zum Lemma in Suchmaschinen
6. Verweis zur grammatischen Kategorie in der wiki-Datenbank

Die nachfolgende Abbildung des Wörterbucheintrages zum Lemma *slovo* ('Wort') veranschaulicht diese Mikrostruktur:

6 Zum Gesamtkonzept s. Ďurčo (2007).

Slovo

From Wicol

Významy

1. základná jazyková jednotka s ustálenou formou i významom
2. jazykový prejav; hovorenie
3. sľub, prísľub, uistenie

Kolokácie

Singulár

Atr + Sub1Nom

abstraktné | básnické | biblické | božie | cudzie | čarovné | čestné | definitívne | duchovné | expresívne | frekventované | heslové | hľadané | hlavné | hovorené | hovorové | hrubé | chlapské | jasné | jediné | jednoduché | jednoslabičné | kľúčové | konečné | krásne | kritické | kvetnaté | láskavé | láskyplné | magické | milé | mrzké | nárečové | nelichotivé | nepekné | neslušné | nespisovné | neznáme | nezrozumiteľné | nežné | nové | odvodené | oplzlé | opytovacie | ostré | pánovo | patetické | pekné | písané | plané | plnovýznamové | poetické | pochvalné | posledné | používané | pravdivé | prázdne | prevzaté | príkre | priliehavé | prisilné | prosté | rozhodujúce | rozumné | silné | sladké | slangové | smelé | sprievodné | sprosté | strašné | škaredé | tlačené | trpké | tvrdé | upokojujúce | úprimné | urážlivé | úvodné | vhodné | vľúdne | vulgárne | vyrieknuté | výstižné | zastarané | záverečné | zbytočné | zdomácnené | zložené | znejúce | zrozumiteľné | živé |

Sub1Nom + Sub2

|

Sub1Nom + Verb

padlo zásadné / dôležité / rozhodujúce / … slovo | slovo dalo slovo | slovo označuje niečo | slovo pochádza z nejakého jazyka | slovo sa spája s niečím | slovo znamená niečo |

Atr + Sub1Gen

v dobrom slova zmysle | v pravom slova zmysle | v širokom / širšom / najširšom slova zmysle | v úzkom / užšom / najužšom slova zmysle |

Sub2 + Sub1Gen

ekvivalent slova | etymológia slova | hlásanie (Božieho) slova | koreň slova | liturgia slova | majster slova | odvodzovanie slova | ohýbanie slova | opakovanie slova | písanie slova | použitie slova | používanie slova | pôvod slova | pravopis slova | sémantika slova | skloňovanie slova | skratka slova | slabika slova | sloboda slova | synonymum slova | tvar slova | váha niekoho slova | vážnosť niekoho slova | výklad slova | výskyt slova | vyslovenie slova | vývin slova | význam slova | zaobalenie slova | zdôraznenie slova | zmysel slova |

Verb + Sub1Gen

byť neschopný slova | ujať sa slova | byť odvodený od slova | odísť bez slova |

Atr + Sub1Dat

|

Sub2 + Sub1Dat

|

Verb + Sub1Dat

dostať sa k slovu | hlásiť sa k slovu | nepustiť niekoho k slovu |

Atr + Sub1Aku

|

Sub2 + Sub1Aku

|

Verb + Sub1Aku

brať slovo späť | byť skúpy na slovo | dať dôraz na slovo | dať niekomu slovo | Dávam ti svoje slovo. | dodržať slovo | hľadať správne / vhodné / výstižné / … slovo | hlásať Božie slovo | hlásiť sa o slovo | hltať každé slovo niekoho | chápať slovo | chytať niekoho za slovo | mať hlavné / rozhodujúce slovo | Máš moje slovo. | nahradiť slovo (iným slovom) | napísať slovo | nebolo mu slovo rozumieť | Nechcem počuť ani slovo! | Nemôzem povedať (jediné) krivé slovo na neho. | Nepočujem ani slovo! | nepovedať ani slovo | Nerozumiem ani slovo! | Neverím ti ani slovo! | nezmôcť sa na slovo | opakovať nejaké slovo | padlo slovo o niekom, o niečom | počúvať / čítať / … slovo za slovom | počúvať slovo Božie | poslúchať niekoho na slovo | použiť nejaké slovo | prečítať slovo | predniesť úvodné slovo | prehodiť slovo s niekým | prekladať / čítať / … slovo po slove | skloňovať nejaké slovo | splniť svoje slovo | spoliehať sa na niekoho slovo | vážiť každé slovo | vyriecť slovo | vysloviť slovo | vyslovovať nejaké slovo | zachovávať Božie slovo | zachytiť nejaké slovo | zdôrazniť nejaké slovo | zvažovať každé slovo |

Atr + Sub1Lok

|

Sub2 + Sub1Lok

|

Verb + Sub1Lok

|

Atr + Sub1Ins

|

Sub2 + Sub1Ins

|

Verb + Sub1Ins

nezmieniť sa o niekom, o niečom ani slovom | slovom sa označuje niečo | stáť si za slovom |

Plurál

Atr + Sub1Nom

citované slová | ďakovné slová | dojímavé slová | chlácholivé slová | krásne slová | láskavé slová | lichotivé slová | matkine slová | milé slová | múdre slová | nesúvislé slová | neuvážené slová | nevyberané slová | ostré slová | otcove slová | pochvalné slová | posledné slová | povzbudivé slová | pravdivé slová | prázdne slová | prorocké slová | silné slová | sladké slová | srdečné slová | tvrdé slová | uznanlivé slová | varovné slová | vl'údne slová | vrúcne slová | vyberané slová | vybrané slová | výstižné slová | vzletné slová | záverečné slová |

Sub1Nom + Sub2

slová evanjelia | slová hovorcu | slová piesne | slová primátora | slová proroka | slová útechy | slová uznania | slová vďaky |

Sub1Nom + Verb

niekoho slová sa naplnili | niekoho slová sa potvrdili | niekoho slová zneli / zazneli | padli silné / kritické / … slová | padli slová o niekom, o niečom | slová mu uviazli v hrdle | slová odzneli | slová vystihujú niečo | slová vyzneli zaujímavo / naliehavo / ... | slová zaznievajú |

Atr + Sub1Gen

podl'a slov iných | podl'a vlastných slov |

Sub2 + Sub1Gen

ozvena slov | pravdivosť niekoho slov | preberanie slov | príval slov | prúd slov | skladanie slov | spojenie slov | spŕška slov | tvorenie slov | útržok niekoho slov | vodopád slov | výber slov | vyhl'adávanie podl'a slov |

Verb + Sub1Gen

povedať pár / zopár slov | rozumieť si bez slov | vymeniť si s niekým pár slov | zhrnúť niečo do pár / niekol'kých slov |

Atr + Sub1Dat

|

Sub2 + Sub1Dat

|

Verb + Sub1Dat

načúvať niekoho slovám | rozumieť niekoho slovám | uveriť niekoho slovám |

Atr + Sub1Aku

|

Sub2 + Sub1Aku

|

Verb + Sub1Aku

adresovať niekomu slová | brať niekoho slová vážne | brať si niekoho slová k srdcu | citovať niekoho slová | dať si pozor na slová | doplniť slová niekoho | hľadať (správne / vhodné / výstižné / …) slová | hltať niekoho slová | komentovať niekoho slová | nájsť vhodné slová | nenachádzať (vhodné / správne / výstižné / …) slová | odvolávať sa na niekoho slová | počúvať niekoho slová | podčiarknuť niekoho slová | pochopiť niekoho slová | potvrdiť niekoho slová | pripomenúť si niekoho slová | reagovať na niekoho slová | spomenúť si na niekoho slová | šepkať niekomu nežné / krásne / … slová | tlmočiť niekoho slová | utrúsiť nejaké slová | vážiť slová | vkladať niekomu do úst slová | voliť opatrné / diplomatické / … slová | vyberať slová | vychrliť zo seba nejaké slová | vypočuť si niekoho slová | vypustiť nepekné / neslušné / … slová z úst | vyriecť slová | vysloviť (prorocké / varovné / …) slová | začuť niekoho slová | zachytiť slová | zopakovať niekoho slová | zvažovať slová |

Atr + Sub1Lok

|

Sub2 + Sub1Lok

|

Verb + Sub1Lok

nezostať len pri slovách | pochybovať o niekoho slovách |

Atr + Sub1Ins

inými slovami | povedané inými slovami | povedané slovami niekoho |

Sub2 + Sub1Ins

medzera medzi slovami | príbuznosť medzi slovami | rozdiel medzi slovami | vzťah medzi slovami |

Verb + Sub1Ins

charakterizovať niečo slovami | inými slovami to znamená | končiť nejakými slovami tému / rozhovor / … | nešetriť slovami chvály / uznania / … na niekoho, niečo | nešetriť slovami chvály / uznania / … na adresu niekoho | opísať niekoho, niečo slovami | pohrávať sa so slovami | reagovať slovami | stotožniť sa so slovami niekoho | súhlasiť so slovami niekoho | uvažovať nad niekoho slovami | vyjadriť niečo slovami | vyjadriť sa nejakými slovami | vysvetliť niečo slovami | za niekoho slovami sa skrýva niečo | zachytiť slovami niečo | zareagovať slovami na niečo | zhodnotiť niečo slovami | zhrnúť niečo (niekol'kými) slovami |

Externé odkazy

slovo ↗ v slovníkoch JÚĽŠ Paradigma slova slovo ↗

- slovo v SNK ↗
- slovo v Google ↗
- slovo v Ask ↗
- slovo v cuil ↗
- slovo v Yahoo ↗

Category: Slovak Nouns

6 Fazit

Die Korpuslexikographie verfügt heute über ein robustes und leistungsfähiges Instrument – die Sketch Engine. In der korpusbasierten Kollokationsforschung und -lexikographie hat dieses Werkzeug eine rasante Steigerung der Effizienz bei der Erstellung der Kollokationsprofile von Wörtern sowie bei der automatischen Sortierung der Wortkombinationen nach ihren grammatischen Relationen bewirkt.

Die Nachteile dieses Tools sind zweierlei Ursprungs. Das erste Problem liegt in der Ungenauigkeit der morphologischen Markierung und der Lemmatisierung der Korpusdaten, wodurch auch fehlerhafte Daten geliefert werden. Das zweite Problem besteht in der Unvollkommenheit der stochastischen Methoden, die für bestimmte Zwecke der empirischen Sprachforschung ungeeignet sind, weil sie primär mit Häufigkeiten und Wahrscheinlichkeitswerten, mit bestimmten vorprogrammierten Schwellenwerten und Abstandsintervallen operieren. Zudem erfassen sie immer nur einen Aspekt, z. B. die Quantität, wobei das lexikographisch und qualitativ Relevante oft verborgen und unentdeckt bleibt.

Der Ausweg aus dem ewigen Dilemma liegt im alten und bewährten Prinzip des Heranziehens aller verfügbaren Quellen, was letztendlich zu dem gewünschten Ziel, also zu einem besseren lexikographischen Produkt führt, jedoch die Arbeit des Lexikographen trotz aller modernen Methoden und Instrumenten nicht leichter und einfacher macht.

Literaturverzeichnis

Ďurčo, P. (2007): *Zásady spracovania slovníka kolokácií slovenského jazyka. http://www.vronk.net/wicol/images/Zasady.pdf*

Ďurčo, P. (2008): „Zum Konzept eines zweisprachigen Kollokationswörterbuchs. Prinzipien der Erstellung am Beispiel Deutsch – Slowakisch". In: Hausmann (Hrsg.) (2008); 69-89.

Hausmann, F. J. (Hrsg.) (2008): *Collocations in European lexicography and dictionary research*. Tübingen.

Heyer, G./Quasthoff, U./Wittig, T. (2006): *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* Bochum.

Kilgarriff, A./Rychly, P./Smrz, P./Tugwell, D. (2004): „The Sketch Engine". In: *Proceedings of Euralex 2004*. Lorient; 105-116. *http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw*

Rychlý, P. (2007): „Manatee/Bonito – A Modular Corpus Manager." In: *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*. Brno; 97-102. *http://nlp.fi.muni.cz/raslan/2007/papers/12.pdf*

Links

Slovenský národný korpus: *http://korpus.juls.savba.sk/*
Morfologická anotácia textov Slovenského národného korpusu: *http://korpus.juls.savba.sk/files/tagset-www.pdf*
Tu vzniká kolokačný slovník: *http://www.vronk.net/wicol/index.php/Main_Page*

Prof. Dr. Peter Ďurčo
Katedra germanistiky
Filozofická fakulta
Univerzita sv. Cyrila a Metoda v Trnave
Námestie J. Herdu 2
917 01 Trnava
Slovakia
durco@vronk.net

Computer-Mediated Discourse vs. Traditional Text Corpora as a Data Source for Idiom Variation Research in Finnish

Oksana Petrova

Dieser Artikel befasst sich mit Usenet Newsgroups als einer alternativen Datenquelle, die neben den traditionellen Textkorpora in der Untersuchung von phraseologischer Variation genutzt werden kann. Die Vor- und Nachteile der beiden Quellen werden diskutiert und am Beispiel des finnischen Phraseologismus *heittää helmiä sioille* ‚Perlen vor die Säue werfen' veranschaulicht. Im ersten Abschnitt werden allgemeine Probleme der Suche nach Idiom-Modifikationen in einem Korpus untersucht. Dabei zeigt sich, dass die Ergebnisse der finnischen Sprachdatenbank nicht genügen, um Entscheidungen über die Muster der Nutzung dieser phraseologischen Einheit zu treffen. Abschnitt II beschreibt die Nutzung des Web und der Usenet Newsgroups als Korpus. Die erweiterte Suche in Google Groups nach den Lexemen *helmi* ‚Perle' und *sika* ‚Schweine', die Suchergebnisse und die Methoden der Datenanalyse werden in den Abschnitten III und IV präsentiert. Anschließend wird das Problem der Relevanz der Variationsvorkommnisse diskutiert, dabei wird argumentiert, dass einige Grenzfälle getrennt behandelt werden müssen. Die Verwendungsmuster der phraseologischen Einheiten werden in Form von verschiedenen syntaktischen Konstruktionen präsentiert. Obwohl die transitive verbale Konstruktion (Vtr-$N1_{OBJ}$-$N2_{TERM}$) in den Usenet-Daten eindeutig am häufigsten gebraucht wird, scheint die verblose Konstruktion ($N1$-$N2_{ALL}$) in der modernen finnischen Sprache mit ihr zu konkurrieren.

1 Introduction

The present paper approaches some problematic aspects of extraction of idiom variants from two different data sources, the first one being the traditional text corpus represented by the Language Bank of Finland (Kielipankki) and the second one – a type of asynchronous computer-mediated communication medium represented by Usenet discussion groups (Google Groups). By idiom variation here I mean all kind of discrepancies that can exist between the default morphosyntactic and phonological form of the target unit and the actual tokens of its occurrence in the data source. Some of this variation can be regarded as more or less regular and unrestricted, e.g. it is normally the case that verbs in predicate idioms inflect,

although there may be some restrictions here as well. By relying on intuition or a smaller result set derived from a corpus, some types of variation (modification, transformation, substitution) can be perceived as more occasional, ad hoc manipulations with the default form, other – as non-available for a particular unit. Still, examination of a larger result set can often lead to the discovery of unexpected variation phenomena and even shake our preconceived notions of variation and the default form. The natural way to obtain a larger result set is to look for a larger corpus, although there cannot be an unequivocal answer to the question about how large the corpus should be in order to be considered "big enough". The most obvious reason for this is different frequencies of occurrence for different idioms within the same corpus, which can be big enough for a more frequent idiom (e.g. the Finnish idiom *kantaa kortensa kekoon* 'lit. to carry one's straw to the stack, id. to do one's bit' with 276 hits returned by the query `[bf='korsi'][]{0,2}[bf='keko']`) and is not for a less frequent one (e.g. the entire result set for *heittää helmiä sioille* 'cast pearls to swine' contains only 16 tokens). Thus, Usenet-texts are expected to provide a more substantial evidence of variation for less frequent idioms than traditional corpora do, due to the archive's extensive size and creative language use characteristic for this type of discourse.

2 Traditional text corpora

By a "traditional" text corpus we mean a large and structured set of texts, which are electronically stored, processed and often completed with linguistic annotation (e.g. British National corpus, Russian National corpus, the Language Bank of Finland etc.). One of the main arguments in favor of using traditional corpora is that they "are carefully compiled in order to be used as a representative sample of language" (Hoffmann 2007: 151) and therefore can be used as a support for generalizations concerning language use. Another argument is that, they do not change so rapidly and the amount of words can be determined at any time. It allows reliably replicating search results and counting normalized frequencies.

Moon, one of the first linguists to systematically use corpus analysis in the study of idiom variation, claims that "effective and robust descriptions of any kind of lexical item must be based on evidence, not intuition" (Moon 1998: 44). She points out, that studies of the variation potential of idioms "are marred by a lack of authentic data or detailed examination of data" (Moon 1998: 105). According to Moon, corpora would provide us with evidence of a suitable type. However, Moon herself admits that her corpus (an 18 million-word Oxford Hector Pilot corpus) is too small to give conclusive information about certain variations, and that variations have to be investigated more fully with much larger corpora (Moon 1998: 105). Given that our objective is a usage-based study of idiom variation in Finnish, the best available traditional corpus would be the Language Bank of Finland (Kielipankki) maintained at the Finnish IT center for science (CSC[1]). It is Finland's largest electronic corpus with approximately 130 million running words of Finnish texts,

1 *http://www.csc.fi*

most of them are periodicals from 1990-2000. Kielipankki is definitely larger than OHPC used by Moon. But does it provide with enough examples of idiom variation?

According to Moon (1998: 51), finding idiom variations is the hardest part of corpus-based investigations. Moon emphasizes that the success of corpus investigation is entirely based on the effectiveness of the corpus tools (Moon 1998: 49). Traditional corpora offer a possibility of using powerful and linguistically oriented search syntax, and Kielipankki with its advanced search syntax is not an exception. However, even the most delicate and flexible corpus tools do not resolve all problems. One of these problems is that investigator bias can hardly be avoided: searches for idiom variants are doomed to be deterministic and only report what has been sought, not what should or could have been looked for. Intuition is necessary, otherwise variations will not at all be found (Moon 198: 49). On the other hand, the construction of search queries preferably should not be affected by preconceptions about non-variability of the investigated item, since it can possibly result in leaving some unexpected tokens of variation outside the search results. A similar point of view is expressed by Herold (2007: 61), who argues that one of the major principles for creating corpus queries is expecting all possible modifications. Thus, our assumptions about idiom variation include at least the following points (here and henceforward the Finnish idiom *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' will serve as an example):

1. Lexical constituents of an idiom can appear in other syntactic constructions than the construction of the default form (e.g. the verbless construction *helmiä sioille* 'pearls/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' instead of the transitive verbal construction *heittää helmiä sioille* 'cast pearls/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$')

2. The linear order in which lexical components of an idiom appear can differ from the word order of the default form (e.g. *heittää sioille helmiä* 'cast swine/$_{\text{PL ALL}}$ pearls/$_{\text{PL PTV}}$' instead of *heittää helmiä sioille* 'cast pearls/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$'). In addition, the sequence of the default constituents can be interrupted e.g. by a modifier (*heittää helmiä saastaisille sioille* 'cast pearls/$_{\text{PL PTV}}$ filthy/$_{\text{PL ALL}}$ swine/$_{\text{PL ALL}}$')

3. Any lexical component of an idiom can appear in a morphological form that differs from the default form (e.g. *heittää helmiä sialle* 'cast pearls/$_{\text{PL PTV}}$ swine/$_{\text{SG ALL}}$')

4. Phonological/orthographic form of any of the idiom's lexical components can differ from those of the default form (e.g. *heittää helemiä sioille* 'cast pearls/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$', where the second *e* in *helemiä* is a dialectal schwa-vowel)

5. Any default lexical component can be substituted with other lexical items (e.g. *syöttää helmiä sioille* 'feed pearls/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$').

It is important to mention, that the above assumptions are not to be considered as some special variation classes: they can both manifest themselves separately and freely combine within the same token, e.g. substituted components can appear in a non-default syntactic construction and also have non-default morphological forms.

When studying idiom variation, the task of the corpus search would be locating possible idiom variants with high accuracy. Accuracy is usually characterized by the two aspects – precision (i.e., the query has to be composed in such manner, that it does not return too many irrelevant hits) and recall (i.e., the search does not miss too many relevant tokens). It is well known that maximizing recall typically leads to low precision, i.e. the least features a query specifies, the more relevant tokens it is likely to include, but at the same time the larger number of irrelevant hits will be returned, which in its turn will demand manual analysis of a very large amount of data.

Herold (2007: 61) remarks that lexical substitution is one of the major modifications that need to be taken into consideration during the query design. According to Herold "we need to assume lexical substitution to be possible for every constituent" (ibid.). The same possibility for each lexical constituent to undergo substitution is expressed in the above assumption (5). Thus, tokens where all three constituents of the Finnish idiom *heittää helmiä sioille* 'cast pearl/PL PTV swine/PL ALL' are substituted are theoretically possible, but apparently they can be recognized as idiom variants only in cases of substitution by lexical units bearing a close semantic relation (synonymy, hypernymy, hyponymy etc.) to the default ones (e.g. *nakata jalokiviä porsaille* 'toss precious stones to piglets'), or if the context contains the default constituents which cohere with the variant, e.g.:

(1) *ennakkovaikutelmat tulevat koetun pohjalta, enkä ole vakuuttunut vielä japanilaisen sarjakuvan/animaation ihanuudesta, mutta jos voit* ***heittää*** *mielestäsi joitakin* ***helmiä****, niin* ***nakkaappa*** *tälle* ***karjulle jokunen****...*[2]
'preconceived impressions come on the basis of experience, and I am not yet convinced of the beauty of Japanese comics/animation, but if you think that you can **throw** some **pearls**, then **toss a few to** this **boar**…'

Since Kielipankki does not support queries based on semantic criteria, the query which is able to match variants with a triple substitution has to be based on morphosyntactic criteria only: `[pos="Verb"][]{0,2}[pos="Noun" case="Part" number="PL"][]{0,2}[pos="Noun" case="All" number="PL"]`[3]. In Kielipankki such query returns 1299 hits. Possible relevant tokens can only be excerpted from this search result manually.

Tokens where any two of the three lexical constituents are substituted can be located by running a separate query for each constituent (the verb *heittää* 'to cast, throw' and the nouns *helmi* 'pearl' and *sika* 'pig, swine'). The search query `[bf="heittää" pos="Verb"]`, which looks up tokens matching only the base form feature of the verb *heittää* 'to cast,

2 *http://groups.google.fi/group/sfnet.harrastus.kulttuuri.sarjakuvat/msg/fe50541c78c3f5ed?hl=fi*

3 In Kielipankki's advanced search syntax query expressions, a search parameter is denoted by square brackets `[ ]`. A required feature in a search parameter is denoted by an equals sign (=). The feature's name is given to the left of the sign and the required value to the right of the sign within citation marks: `[key="value"]`. The empty search parameter `[ ]` matches any token whatsoever. The keys *bf* and *pos* are abbreviations for 'base form' and 'part of speech' respectively.

throw' with no restrictions on the morphological form whatsoever, returns 17292 hits. The query `[bf="helmi" pos="Noun"]` returns 3187 hits, and the query `[bf="sika" pos="Noun"]` returns 2705 hits. Again, the only way to find all relevant hits is to scroll through all results and extract relevant hits manually, which is obviously a very labor-intensive and time-consuming task. A very common solution used by corpus linguists in order to decrease the amount of data is to look at a randomly selected subset. However, this solution is hardly applicable to the analysis of idiom variation due to the low idiom frequencies: e.g. Moon (1998: 60) observes that over 70% of fixed expressions and idioms in her database have frequencies of less than 1 per million tokens[4]. There is a high probability that randomly selected hits will not match any tokens of idiom use at all, apart from tokens of its variation.

Since Moon remarks, that searches are most successful when the query consists of two lexical words fairly close together (1998: 51), several queries, each consisting of two lexical components of the idiom *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$', were made. The query `[bf='helmi'][]{0,5}[bf='sika']` finds all matches of the lemma *helmi* 'pearl' with the lemma *sika* 'swine' occurring within a window of between zero and five arbitrary tokens. This query returns 16 hits with the 100% precision, i.e. all of the hits are relevant. The reversed order query `[bf='sika'][]{0,5}[bf='helmi']` matched 1 hit, which, however, was not a relevant hit. The query `[bf='heittää'][]{0,5}[bf='helmi']` returned 3 hits, one of them was relevant, but the same token was already matched by the `[bf='helmi'][]{0,5}[bf='sika']` query. The query `[bf='helmi'][]{0,5}[bf='heittää']` returned 1 hit, which was not relevant. The query `[bf='heittää'][]{0,5}[bf='sika']` returned 3 hits, 1 of them was relevant, 0 new. The query `[bf='sika'][]{0,5}[bf='heittää']` returned no hits. Thus, all the above queries matched altogether 16 tokens of the idiom in Kielipankki. These queries could find tokens of variation, where constituents are used in other constructions than the "base form", tokens with constituents in morphological forms other than the "canonical" forms and tokens where constituents are in a reversed order. They could also detect tokens where one of the three lexical constituents is substituted. However, they could not match tokens with two substituted constituents. In fact, the Kielipankki search results matched zero tokens with substituted noun constituents and 3 tokens with verbal substitution: 2 transitive (*tarjota* 'offer', *heitellä* 'fling') and one intransitive (*kadota* 'disappear').

The main advantages of traditional corpora are their representativeness, ability to count normalized frequencies and flexible linguistically oriented search tools. It has been pointed out that the latter two do not provide any substantial help in idiom variation analysis. As for the representativeness, one must be careful making generalizations on the basis of data obtained from the Kielipankki, since its texts are representative to the large extent only of a

4 Another problem is that idiom frequencies can be difficult to assess in the first place. Corpora are quantified in terms of individual words, but word-based frequency counts are not ideal for fixed-expressions and idioms that are multi-word units (Moon 1998: 57).

single type of discourse – newspaper articles[5]. Although Kielipankki is larger than OHPC used by Moon (1998), the scarce number of tokens which we could retrieve for the idiom *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' is obviously not enough to identify generalities (patterns) of its use. Thus, based only on 16 hits it is virtually impossible to determine the default form of this idiom. E.g., while dictionaries would normally list the transitive verbal construction *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' as the default form, 50% of the total of 16 tokens obtained from Kielipankki are represented by the verbless construction *helmiä sioille* 'pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$'. Finally, although corpus linguistics should be based on observation rather than introspection, finding tokens of idiom variation is inevitably a matter of serendipity (Moon 1998: 51).

3 Google Groups and Usenet

The Web as corpus has its advantages: it is machine readable, free and easily accessible and what is more important for the study of low-frequency phenomena - it is exceptionally large. For comparison, the query *"helmiä sioille" group:sfnet.** which searches for the exact string *helmiä sioille* 'pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' in all *sfnet* groups, performed in Google Groups on 05.03.2009 returned 287 hits, i.e. about 18 times more than the above-mentioned more flexible searches in Kielipankki did. This particular search has 100% precision, however, its recall is low, since it cannot retrieve any case and number variants or lexical substitutes of the noun components *helmi* 'pearl' and *sika* 'swine'. Thus, the search indicates that the quantity of data for this idiom that can be obtained via Google Groups is substantially larger than that of the Kielipankki.

But if we consider the whole of the World Wide Web as a single corpus that can be accessed via commercial search engines such as Google, we will face a number of serious limitations (Hoffman 2007: 151). The first one concerns reproducibility: in contrast to the traditional corpora, the World Wide Web is of indeterminable size and moreover is constantly growing, i.e. no normalized frequencies can be counted. Search results are very unstable and replicability of linguistic findings is virtually impossible in the World Wide Web. The second limitation has to do with search flexibility. Commercial search engines, like Google, Yahoo, AltaVista etc, do not allow search algorithms available in traditional corpus tools. It makes the retrieval of data for linguistic purposes a far more difficult and time-consuming enterprise. Only specially designed linguistic pre-/post-processing search engines like Webcorp, KWiCFinder or Linguist's Search Engine are able to present examples of word usage from the Web in a form somewhat suitable for linguistic analysis.

Hoffman (2007) presents a number of solutions, which could make Internet more suitable for linguistic investigations. One of them is to restrict the object of study to a clearly defined subsection of the World Wide Web. Another solution is to create a local copy of data by downloading relevant Web pages, post-process them and search with corpus tools. According to Hoffmann, using smaller and tailor-made Web-derived corpora allows to

5 Nenonen (2007: 215) remarks, that the absence of equal corpora of modern colloquial Finnish at present makes the World Wide Web the best source for this kind of language.

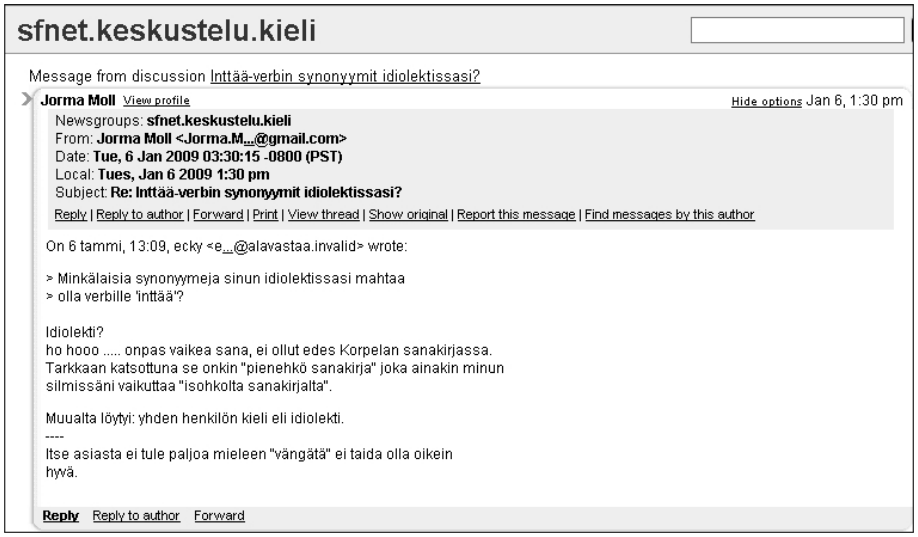

Figure 1: Screenshot of the Usenet message posted 6. 1. 2009 on *sfnet.keskustelu.kieli* obtained from *http://groups.google.com/group/sfnet.keskustelu.kieli/msg/00736490b46c901a* (seen 21. 2. 2009).

expand the range of available data without compromising on the application of the standard corpus methodology. Hoffmann himself creates such a specialized Internet-derived corpus from a selection of Usenet newsgroup messages.

Usenet (USEr NETwork) is a global, decentralized computer network communications system. It was conceived in 1979 and by 1990-s it has developed into the largest system of discussion groups (often called newsgroups) on the Internet. It consists of thousands of discussion groups – hierarchically and thematically organized forums that allow people to share their thoughts and opinions on just about every imaginable subject and comment on the postings of others (Hoffmann 2007). Names of discussion groups indicate the topics that are discussed, e.g. a group *sfnet.keskustelu.foreigners* from the Finnish Usenet hierarchy *sfnet* is intended for foreigners living in Finland or visiting the country (*keskustelu* is the Finnish for 'discussion'). Figure 1 shows an example of a Usenet message posted on the newsgroup *sfnet.keskustelu.kieli* - 'sfnet.discussion.language'. Here, the author replies to a question about the synonyms of the verb *inttää* 'argue, insist' in his idiolect. The question from the topic-starting message appears before the answer as a quotation, marked by an angle bracket at the beginning of each line of the quoted text.

Usenet newsgroups do not require participants to be online simultaneously which puts them into the category of asynchronous computer-mediated communication (CMC).

However, even though messages can be replied to with a considerable lapse of time, the nature of Usenet discussions is clearly interactive: participants often quote passages from previous posts as part of their replies, which greatly facilitates the establishment of topical coherence. Usenet discussions can thus be regarded as a hybrid form of communication which combines features of face-to-face talk with those of written texts (Hoffmann 2007). From the text-linguistic point of view, Usenet messages are a more homogenous data source than the World Wide Web. The interactive character of Usenet texts makes them an excellent data source for the study discourse-pragmatic aspects of idiom use and variation. The nature of the medium also stimulates its writers to use their language in an expressive, creative way, thus producing interesting tokens of occasional idiom variation.

4 Obtaining idiom variation data from Google Groups Usenet archive

The Usenet-corpus-compiling procedure described by Hoffmann (2007) requires programming skills, UNIX system administration skills for setting up a news server, as well as access to a commercial newsfeed. The process of downloading the entire contents of all selected newsgroups onto the local hard disk can result in the transfer of enormous amounts of data and thus requires adequate hardware and network bandwidth. Unfortunately, the present study could not meet these requirements. Instead, we have chosen to obtain data via advanced searches in Google Groups[6] which serve as the Web's most comprehensive archive of and interface to Usenet newsgroup postings dating back to 1981[7]. This method of data retrieval has its problems and limitations. However, taking into the consideration the general problematical character of the application of corpus methodology for the study of idiom variation discussed in the previous section, I will, in what follows, try to demonstrate that it can be justified, as long as we do not concern ourselves with normalized frequencies but concentrate primarily on the qualitative aspects of idiom variation.

In previous corpus-based idiom variation studies (Moon 1998, Sköldberg 2004, Fellbaum 2007) data was gathered for a set of different expressions: Moon looks at a set of 6776 English fixed expressions and idioms, participants in the Wolfgang Paul-Preis Project whose results are presented in Fellbaum (2007) investigate some 1000 pre-selected German multi-word units, while Sköldberg restricts her set to 36 Swedish idioms. Due to the low frequencies of idioms in traditional corpora, the amount of tokens for the majority of idioms is rather low. E.g., in Moon's data, 72% of fixed expressions and idioms have 0-17 tokens. Sköldberg, who deliberately chooses to look at idioms which occur in her 33 million word corpus with frequencies of more than one token per million words (i.e. not less than 33 tokens for a single idiom), reports that 32 in her set of 36 idioms are represented by less than 100 tokens, while the most frequent idiom is represented by 177 tokens. On the other hand, the above mentioned test query *"helmiä sioille" group:sfnet.** which returned 287 hits in Google Groups indicates, that by running recall-maximizing queries the total

6 *http://groups.google.com*

7 Initially Usenet discussions were archived by DejaNews and the archive was acquired by Google in 2001

number of tokens that could be obtained for the Finnish idiom *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' can amount to several hundred. Such quantity of data allows us to perform a radically different kind of idiom variation analysis: a thorough study of variation patterns in a single idiom.

Thus, my goal is to gather variation data for *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$'. Advanced search in Google Groups[8] allows restricting search queries according to several parameters: exact wording or phrase, language, site or domain, message date, group, subject and author. Since Google search engine does not offer a possibility of searching for different word forms using wildcard truncation the search queries have to be carried out for all possible surface forms of the leximes *helmi* 'pearl' and *sika* 'swine'. Searching for each and every lexical constituent separately in all possible forms is a very labor-intensive and time-consuming task. The situation is complicated by the fact, that Finnish is a morphologically rich language (Karlsson 1983). Consequently, even if we restrict our analysis to the case-number-possessive suffix paradigm, the noun components *helmi* 'pearl' and *sika* 'swine' will still have 312 different morphological forms altogether. If we add clitics (*-kin, -kAAn, -hAn, pA(s)*), possible dialectal or slang variants the number of word forms will increase considerably, so I had to exclude the latter from the search queries. Therefore, the total of 312 advanced search queries (159 for the leximes *helmi* 'pearl' and 153 for the leximes *sika* 'swine') were performed. The task at hand could have been made less complicated, by using an inflection generator, on the one hand, and by creating custom-built software specifically fitting our research problem, on the other. The latter kind of software would e.g. use Google SOAP Search API[9] to automatically search for and analyze different predefined inflection forms of a given word. However, this option was not available in the course of the current research, since it requires programming skills that are beyond my capabilities at the present time, although it is possible that designing such software can become a part of a future research project. Queries corresponded to the following parameters:

- Search for exact wording or phrase (e.g. *"helmi"*, *"helmeni"*, *"helmesi"* etc.)
- Messages posted between 1 Jan 1990 and 31 Dec 2006
- Messages from the group at this location: *sfnet.**

Language restriction was unnecessary, since sfnet.* is by default a Finnish language hierarchy. In addition, search results were requested to be sorted by date and to be displayed 100 results per page. The search results page presents results in the following form:

Karalahti pillittää taas
sfnet.keskustelu.vitsit – 74 posts – 16 authors – Last post: Apr 30, 2006
...reader1.news.jippii.netsfnet.keskustelu.vitsit:178905 Jupehan se siellä tykittää taas.

8 *http://groups.google.com/advanced_search*
9 *http://code.google.com/apis/soapsearch/*

Tiesikö edes jupe tuota. Rietas naisenkuva on vain **helmi** sioille?
http://groups.google.com/g/8fb7ff09/t/f448990d1507ea9a/d/bbe546cbddfcabaf

Here, the first row is the title of the discussion thread (Karalahti pillittää taas - 'Karalahti blubs again'). The second row contains: name of the discussion group (*sfnet.keskustelu.vitsit* - 'sfnet.discussion.jokes'), number of posts in the thread (74), number of authors (16) and date of the last post (Apr 30, 2006). The following two rows constitute the snippet, which is Google's algorithmic attempt to extract the part of the discussion thread most relevant to the search query. Normally it is an excerpt from the message which contains the searched item in boldface type (**helmi** 'pearl' in the above example). In most cases the snippet is enough for determining whether the hit is relevant or not, thus there is no need to open every thread returned by the search. The last row contains the thread's URL address.

Making sense of Google search results is actually quite a difficult problem to tackle. For instance, the search for the exact form *"helmi"* on sfnet.* between 1 Jan 1990 and 31 Dec 2006 returns about 5,380 results. First of all, this is not by any means an exact number. For the sake of efficiency, Google estimates the number of results and this estimate of the total number of results is rather unreliable. Secondly, Google would never display more than about 400-700 (presumably, randomly selected) search results. My personal solution to this problem is to run for the most common morphological forms[10] separate searches for each year or a couple of years at a time and then sum up the number of results. Thirdly, search results only show the number of discussion threads where the searched item occurs, not the actual number of occurrences. Thus, since the same item could possibly occur several times within the same discussion thread, the actual number of occurrences could be much higher. Overall, having only Google search engine at hand, counting the exact number of occurrences for each word form seems to be a difficult, if not impossible task. On the other hand, it is unnecessary, since our goal is limited to the retrieval of idiom variants. The only point of our concern should be the reproducibility of search results. As it has been pointed out earlier, search results obtained from the Web are unstable and therefore not reproducible. Web content is constantly changing, web pages can disappear. In this sense, Google Groups has a clear advantage: old messages in the archive can still be retrieved and running advanced searches with restricted message dates could in theory be reproducible, but unfortunately this is not exactly the case. Searches for one year at a time significantly improve reproducibility, but even they can return slightly different results on different occasions (usually about ± 0-5 hits) for reasons which remain largely unknown, since Google search algorithms are Google's trade secret.

5 Search results and data analysis

Only 47 out of 159 word forms for the lemma *helmi* 'pearl' have returned any results and among these only 20 word forms have returned results, which contained altogether 496

10 This kind of queries were performed for the following word forms: *"helmi", "helmet", "helmiä", "sika", "siat", "sian", "sikaa", "sikoja", "sioille".*

Date	Author	Group	Example
16.12.1997	Korhonen Tommi	sfnet.keskustelu.seksi	Mutta ajattelinkin että joku *poimisi helmet/*
16.12.1997	Korhonen Tommi	sfnet.keskustelu.seksi	/ ja *tarjoaisi sioille*, mina vaan röhnöttäisin sillä välin (kyl-jelläni) ja röhkisin.

Table 1: Lexical constituents *helmi* 'pearl' and *sika* 'swine' distributed between two autonomous syntactic constructions

relevant word form tokens. For the component *sika* 'swine', 42 out of 153 word forms have returned any results and only 14 word forms have returned results, containing altogether 470 relevant word form tokens. By a relevant word form token we mean a single token of the searched word form, occurring in a sentence which can, according to the semantic criteria, be regarded as a context of the *heittää helmiä sioille* 'cast pearl/$_{\text{PL PTV}}$ swine/$_{\text{PL ALL}}$' idiom use. Query results were then exported into Microsoft Access database application, where each record corresponds to a single token of idiom use within an autonomous syntactic construction. For instance, the following example (2) was recorded as two separate entries (table 1): one containing the verb *poimia* 'pick' and the noun component *helmi* 'pearl' and the other including the verb *tarjota* 'offer' and the noun component *sika* 'swine'.

(2) *Mutta ajattelinkin etta joku **poimisi helmet/** ja **tarjoaisi sioille**, mina vaan röhnöttäisin sillä välin (kyljelläni) ja röhkisin.*[11]
'But I thought that someone would **pick pearls/** and **offer to swine**, I would just loll about on my side and grunt'.

In the database different types of data relating to the organizational and formal aspects of individual tokens of idiom use were recorded: type of construction, morphological form of each noun constituent (case, number and possessive suffix), negation and modality, lexical substitution, word order, modifiers, appellatives and evidentials. Organizational fields recorded message ID, date, year, author and the name of discussion group. The entire database contains 588 tokens of idiom use occurring in 521 different newsgroup messages written by authors with 343 different nicknames[12] on 97 different sfnet.* groups. The question of what should count as a relevant token is not a trivial one, when the object of study is idiom variation. The "default form" of idioms is usually defined as a form which simultaneously meets several different criteria:

– phonological, i.e. the presence of certain lexical items in the same structure;

– syntactic, i.e. a particular structure in which lexical items appear;

11 *http://groups.google.com/group/sfnet.keskustelu.seksi/msg/937ee650b2db8fa9*

12 It is difficult to trace whether the same author is actually writing under several different nicknames.

– semantic, i.e. a particular conceptual structure associated with the phonological and syntactic structures;

– criterion of institutionalization, i.e. the string being recognized and accepted as a phraseological unit of the language, or in corpus terms the frequency of the string (Moon 1998).

The last criterion is neither necessary nor sufficient for an idiom variant (although some variation classifications distinguish between "usual" vs. "occasional" variation). As for the first three (phonological, syntactic and conceptual structures) – the most difficult problem is to determine which combination of these would be necessary and sufficient for a variant to be considered as a relevant token. Borderline cases are inevitable and one has to decide whether or not to include them into database. My solution was to record them as well, but by adding a field "UNCLEAR" enable their filtering from more clear-cut results. This was mainly done for the sake of the quantitative morphosyntactic analysis, which allowed distributing clear tokens between different constructions (table 3 on page 148) without completely discarding interesting but less clear tokens of idiom variation. Thus, 85 borderline tokens (marked as "UNCLEAR" in the database) have been recorded, which include:

– Quotations of the original biblical passage (including slightly inexact ones, as in the following example where *helmiä* 'pearl/$_{\text{PL PTV}}$' lacks a possessive suffix *-nne* '2PL' which is present in the biblical source):

(3) *Sillä tiedäthän, että "**Älkää heittäkö helmiä sikojen eteen**, ja sitä mikä on pyhää, koirille, etteivät ne kääntyisi ja repisi teitä".*[13] "For you know, that '**Do not cast pearls before swine** and what is holy to the dogs, lest they turn and tear you in pieces'"

Although such quotations meet both the phonological requirement and the semantic requirement, they cannot be considered as tokens of the idiom in a strict sense. As Dobrovol'skij and Piirainen (2005: 231) remark, there are many text fragments of that were initially used as citations before they gradually developed into conventional-figurative units. In the above example citation marks and inclusion of the "holy to the dogs"-passage both indicate that we are rather dealing with a biblical quotation, which is considered to be a source for this particular idiom, than with the idiom itself.

– Cases where both lexical components *helmi* 'pearl' and *sika* 'pig, swine' are "hosted" by another syntactic construction, which belongs to a different construction family, e.g. in the following example (4) the host is the conventionalized Finnish construction NPsubj {X} Vtr {*etsiä* 'search'/ *seuloa* 'sieve'/ *poimia* 'pick'/ *tonkia* 'dig'/ *löytää* 'find'} NPobj {N ***helmi*** 'pearl'} NP_{ELA} {*roska* 'garbage'/ *paska* 'shit'/ *romu* 'junk'} where NP_{ELA} is substituted by *sika* 'pig, swine', and in example (5)

13 *http://groups.google.fi/group/sfnet.keskustelu.ihmissuhteet/msg/ee5db18c263f52d8?hl=fi*

the host is another Finnish construction NPsubj {X} Vcogn {*ymmärtää* 'understand'/ *tietää* 'know'} NP_{ELA} {Y} *yhtä paljon kuin* NP {N ***sika*** 'pig'} $NP_{ELA/PTV}$ {*hopealusikka* 'silver spoon'} 'X understands/knows about Y as much as a pig about a silver spoon' where NP_{ELA} is substituted by *helmi* 'pearl'. Boldfaced lexical items in both of the host constructions are shared with the *helmiä sioille*-construction family and are preserved in the resulting blends together with original syntactic structure of the hosts. Since both *helmi* 'pearl' and *sika* 'pig, swine' appear in the blend, one could assume that (4) and (5) are tokens of *heittää helmiä sioille* 'cast $pearl/_{PL\ PTV}$ $swine/_{PL\ ALL}$'. However, according to the syntactic and the semantic criteria these are rather tokens of the host constructions.

(4) *Joudut siis **noukkimaan helmiä sikojen joukosta**.*[14]
'So you have to **pick pearls among pigs**'

(5) *Tiedätte epilepsiasta yhtä paljoa **kuin sika helmistä!***[15]
'You know about epilepsy as much **as a pig about pearls**'

On the other hand, tokens where one of the recurrent *helmiä sioille*-constructions itself functions as a "host" for a lexical component from another construction family are not labeled as UNCLEAR and are therefore counted together with other similar constructions in the database. E.g. in the following example the component *helmi* 'pearl' is substituted by the lexical unit *hopealusikka* 'silver spoon', which is "borrowed" from the construction NPsubj {X} Vcogn {*ymmärtää* 'understand'/ *tietää* 'know'} NP_{ELA} {Y} *yhtä paljon kuin* NP {N *sika* 'pig'} $NP_{ELA/PTV}$ {N *hopealusikka* 'silver spoon'} 'X understands/knows about Y as much as a pig about a silver spoon':

(6) *Enpä taida enään herra Burmaniin soveltaa ironiaa, sehän on kuin **hopealusikoita sioille**.*[16]
'I am not likely to apply irony to Mr Burman anymore, it is like **silver spoons to swine**.'

– Isolated lexical components i.e. components of an idiom that do not occur together with other lexical components of the same idiom within the syntactic structure of the same clause (Petrova 2007), e.g.:

(7) *Eikä **sioissakaan** mitään vikaa ole. (Hengellisiä) **helmiä** kun on loputtomasti tarjolla. Toisaalta **siat** haluaisivat ennemmin ruokaa, koska siitä on heille enemmän hyötyä. Mutta **helmet** voivat ne tappaa joutuessaan henkireikään.*
'There is nothing wrong with **swine** either. 'Cause (spiritual) **pearls** are in

14 *http://groups.google.fi/group/sfnet.viestinta.nyyssit/msg/d6425463f4112183?hl=fi.*
15 *http://groups.google.fi/group/sfnet.keskustelu.varaventtiili/msg/8a6c3be4b14334d5?hl=fi.*
16 *http://groups.google.fi/group/sfnet.atk.sodat/msg/4d5ef2ebb5e7a858?hl=fi.*

> endless supply. On the other hand **swine** would like more food, because it is of more benefit to them. But **pearls** they can kill by getting into the windpipe.'

> Isolated lexical items of the type presented in (7) are not appearing as constituents of any construction of the "*helmiä sioille*"-construction family (table 3 on page 148) and therefore could not be counted in the database.

As for the 20 cases where NPs *helmi* 'pearl' and *sika* 'swine' are distributed between the main clause and the relative clause as in table 2 on the facing page, although the main clauses containing modified NPs *helmi* 'pearl' and *sika* 'swine' (labeled as MODIFIED N1 or MODIFIED N2) have been recorded as separate 10 tokens, they were excluded from the final calculation of construction patterns, where only 10 tokens of the relative clause constructions have been counted according to the construction, which they represent.

The remaining 494 tokens are represented by the morphosyntactic constructions in table 3 on page 148. The category OTHER (5% of the total) contains several different constructions, each of which occurs in the data with less than 4 tokens. Constructions are ordered according the number of tokens, the transitive verbal construction (Vtr-$N1_{OBJ}$-$N2_{TERM}$) is represented by the majority of tokens (190), followed by the verbless construction (N1-$N2_{ALL}$). Thus, based on 494 tokens we can examine syntactic patterns of the idiom's use with more liability, than in the scarce data obtained from the traditional text corpus like Kielipankki, where the most frequent construction for the examined idiom was the verbless construction (N1-$N2_{ALL}$).

6 Summary

The article presented Usenet newsgroups as an alternative data source for phraseological variation study compared to the traditional text corpus. Advantages and disadvantages of both sources were discussed and exemplified by the case study of the Finnish phraseological unit *heittää helmiä sioille* 'cast pearls before swine'.

The in-depth empirical study of this particular idiom based on the result set consisting of 588 tokens manually extracted from Google Groups has several goals, all of which unfortunately can not be fully explicated in the course of this single paper. Together they constitute a research project which has been conducted by me at Åbo Akademi University (Turku, Finland) and which will hopefully result in a PhD dissertation published in the nearest future. Although the results of this usage-based study can also have lexicographic implementations, its primary goals rather lie within the scope of theoretical methodology: it is an attempt to adapt the theoretical model and formal descriptive tools of Conceptual Semantics (Jackendoff, Nikanne) for the purpose of integrated analysis of the idiom's internal structure and variation. Some of the contextual and discourse-pragmatic aspects of the idiom's use can also be studied in relation to its structure, e.g. the formal description of the idiom's structure can be applied to the analysis of textual cohesion (see Petrova 2007 for some preliminary results).

ID	Date	Author	Group	Example	CONSTR
76	27.6.1998	Patrick Uotinen	sfnet.keskustelu.uskonto	*Gregorius, joka siis itse tallensi omat sanansa, jatkoi vielä letkautuksella* ***helmistä****,* / 'Gregorius, who himself recorded his words, continued with a quip about pearls'	MODIFIED N1
76	27.6.1998	Patrick Uotinen	sfnet.keskustelu.uskonto	***joita*** *hänen ei tarvitse* ***heitellä*** *saastaisen* ***sian eteen****!*[a] / 'that he does not need to throw before a filthy swine!'	Vtr-N1$_{OBJ}$-N2$_{TERM}$

Table 2: Idiom constituents in the main clause and the relative clause.

a http://groups.google.fi/group/sfnet.keskustelu.uskonto/msg/c6ab0b13214142c4?hl=fi

Construction	Tokens	%	Example
Vtr-N1$_{OBJ}$-N2$_{TERM}$	190	38	*Minä en viitsi* ***heitellä helmiä sioille.*** / 'I do not bother to **throw pearls to swine**'
N1-N2$_{ALL}$	152	31	***Helmiä sioille****; ei se kuitenkaan ymmärrä.* /'**Pearls to swine**; it does not understand anyway.'
NEG-N1$_{PTV}$-N2$_{ALL}$	52	11	***Ei helmiä sioille****, kuten sanonta kuuluu.* /'**No pearls to swine**, as the saying goes.'
Vintr-N1$_{SUBJ}$-N2$_{TERM}$	32	6	***Menee*** *kyllä* ***helmet sioille.*** /'**Pearls** do **go to swine**'
N1$_{GEN}$-DVtrN-N2$_{TERM}$	31	6	Inkan euroviisuehdokas oli **helmien heittämistä sioille**. /'Inka's Eurovision candidate was **throwing of pearls to swine**'
N1-&-N2	5	1	*Aivan, taide ei ole kaikkia varten. Miten se menikään se juttu* ***helmistä ja sioista****....* /'Exactly, the art is not for everyone. How did it go this thing about **pearls and swine**…'
N1-N2$_{ILL}$	5	1	*Luepa se viimevuoden keväällä tähän ryhmään lähetetty "****Helmii sikoihin****" keskustelun alku.* / "Read the beginning of this '**Pearls into swine**' discussion posted in this group last spring."
N2$_{ADE}$-PCP-N1	4	1	***Possulle heitetty helmi*** *tämäkin viesti, mutta kyllä sellaisia on.* /'**A pearl thrown to a piggy** [is] this message as well, but it is just like that.'
OTHER	23	5	

Table 3: Morphosyntactic constructions of the "*helmiä sioille*"-construction family

In the course of the present paper I have mainly concentrated on the methods of data retrieval and analysis. Since I do not master any programming skills and thus was not able to use advanced computational-linguistic tools, my methods may appear rather unsophisticated and probably inapplicable for a study of a larger set of idioms, but I still find them quite sufficient for the tasks of the current research. When it comes to the detection of idiom variants, the main point of my concern was to formulate queries that would meet the following principles: 1) all possible modifications have to be expected and 2) search accuracy has to be maximized so that all, or at least most of the possible modifications can be retrieved. The development and implementation of custom-made software aimed at automatic extraction of idiom variants is definitely a problem for future research, but it is not a problem, which I was trying to approach here. To my knowledge, the problem still remains largely unsolved in corpus linguistics, e.g. Herold (2007: 54) remarks: "Developing queries is essentially a manual task. We do not use techniques for automatic identification and extraction of target idioms or any other expressions." He also points out that: "So far there is no sufficiently robust automatic process known to us that would permit corpus-driven extractions of idiomatic expressions" (2007: 56). There are indeed automatic processes that allow the extraction of statistically significant co-occurrences of certain lemmas, but I do not see how it can help us when it comes e.g. to the extraction of occasional lexical variants, that cannot be predicted beforehand. As regarding the application of part-of-speech annotation to the Usenet texts, which constitute the major data source for the project, it appears to be technically impossible due to the tremendous size of this archive. It could be achieved only by downloading a small part of it (as Hoffmann 2007 does), which again would require programming skills that I do not possess.

Apropos the variation analysis, due to the paper size limitations, I could not cover all of its aspects and therefore have concentrated primarily on the morphosyntactic patterns of the target unit's use, which were presented in the form of various constructions. Together these are assumed to constitute this unit's construction family. The scarce number of tokens which I could retrieve for the target idiom in Kielipankki was obviously not enough to identify generalities (patterns) of its use. Thus, based only on 16 hits it is virtually impossible to determine the default form of this unit: 50% of the total of 16 tokens obtained from Kielipankki are represented by the verbless construction (N1-N2/$_{\text{ALL}}$). On the basis of the data obtained from Usenet, it has been observed that, although the transitive verbal construction (Vtr-N1obj-N2$_{\text{TERM}}$) is clearly the most frequent one for the examined phraseological unit, the verbless construction (N1-N2$_{\text{ALL}}$) seems to compete with it for the default status in the modern Finnish.

List of Abbreviations

N1 the first noun component, i.e. *helmi* 'pearl' or its substitute

N2 the second noun component, i.e. *sika* 'pig, swine' or its substitute

Vtr transitive verb, i.e. *heittää* 'to cast, throw' or its substitute

Vintr intransitive verb, e.g. *mennä* 'to go'
DVtrN deverbal noun, e.g. *heittäminen* 'throwing'
PCP participle, e.g. *heitetty* 'thrown'
NEG negative word, e.g. *ei* 'no'
ALL allative case[17]
ADE adessive case
ELA elative case
GEN genitive case
ILL illative case
PTV partitive case
TERM terminative local case
OBJ object case
SUBJ subject case

Bibliography

Dobrovol'skij, D./Piirainen, E. (2005): *Figurative Language: Cross-Cultural and Cross-Linguistic Perspectives.* Amsterdam.
Fellbaum, C. (ed.) (2007): *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies.* London.
Herold, A. (2007): "Corpus queries". In: Fellbaum (ed.) (2007); 54-63.
Hoffmann, S. (2007): "Processing Internet-derived Text—Creating a corpus of Usenet Messages", in: *Literary and Linguistic Computing 22/2*; 151-165.
Karlsson, F. (1983): *Suomen kielen äänne- ja muotorakenne.* Porvoo.
Moon, R. (1998): *Fixed Expressions & Idioms In English. A corpus-Based Approach.* Oxford.
Nenonen, M. (2007): "Unique, but not cranberries: idiomatic isolates in Finnish". In: Nenonen/Niemi (eds.) (2007); 213-226.
Nenonen, M./Niemi, S. (eds.) (2007): *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, May 19-20, 2006.* Joensuu.
Petrova, O. (2007): "Phraseological component isolation in computer-mediated discourse: a formal approach", in: Nenonen/Niemi (eds.) (2007); 269-281.
Sköldberg, E. (2004): *Korten på bordet. Innehålls- och uttrycksmässig variation hos svenska idiom.* Göteborg.

Oksana Petrova
Department of Finnish Language
Åbo Akademi University
Fabriksgatan 2
20500 Turku
Finland
opetrova@abo.fi

17 Case and number are marked as a subscript attached to the noun component, e.g. $N2_{TERM}$ means "the second noun component (i.e. *sika* 'pig, swine' or its substitute) in a terminative local case".

Methoden und Tools zur Erstellung eines korpusbasierten Kollokationswörterbuchs (am Beispiel des Kroatischen)

Melita Aleksa Varga

The present paper discusses the preliminary work needed to be conducted when compiling corpus-based dictionaries, with a special emphasis on a highly inflectional language, the Croatian. The collocations in the scope of this paper are regarded as word combinations of two or more lexical units, whereas there is a high probability that these lexemes, when considering the language use, will always be placed near each other. In opposition to the traditional methods of looking up collocations in order to compile a collocations dictionary, there are automatic ways presented here, i.e. the automatic search in corpora using statistical freeware. The programs (NSP, Collocations Extract and KWIC Concordance for Windows) were tested on a test-corpus and the results were compared. Due to the fact that the project of compiling a Croatian corpus-based collocations dictionary is still in its beginning phase, there is the needed Croatian preliminary work discussed, as well as language-specific problems, such as lemmatising of texts written in Slavic languages and the language policy.

1 Einführung

Für die Erstellung eines korpusbasierten Wörterbuches sind eine Reihe von Vorarbeiten notwendig. Der vorliegende Beitrag beschreibt diese am Beispiel des Kroatischen als einer flektierenden Sprache und schildert die damit zusammenhängenden Probleme sowohl hinsichtlich der maschinellen Bearbeitung des Korpus, als auch hinsichtlich der sprachpolitischen Situation. Da manche Probleme verallgemeinert werden können, bietet dieser Beitrag einige nützliche Hinweise für andere flektierende Sprachen.

Ehe man sich mit der eigentlichen Problematik beschäftigt, ist der Begriff *Kollokation* näher zu erläutern. Das Duden Universalwörterbuch (1997) definiert eine Kollokation als a) „inhaltliche Kombinierbarkeit sprachlicher Einheiten“ (z.B. *dick* + *Buch*, aber nicht: *dick* + *Haus*); b) „Zusammenfall verschiedener Inhalte in einer lexikalischen Einheit.“ Das Große Wörterbuch der kroatischen Sprache (Anić 2004) definiert eine Kollokation als eine „verbindliche oder übliche Wortverbindung“, die nicht durch die Grammatik bestimmt

wird[1]. Weitere Definitionen des Begriffs Kollokation aus phraseologischer Perspektive finden sich in Fleischer (1997: 251-253).

In diesem Beitrag werden Kollokationen aus korpuslinguistischer Sicht betrachtet und als Wortverbindungen von zwei oder mehreren lexikalischen Einheiten definiert, wobei die Wahrscheinlichkeit, dass deren Bestandteile im tatsächlichen Sprachgebrauch immer nebeneinander auftreten, sehr groß ist. Diese Auffassung stammt von Lemnitzer und Zinsmeister, die unter dem Begriff Kollokation das gemeinsame Vorkommen (‚Kookkurenz', ‚Kovorkommen') sprachlicher Elemente verstehen (Lemnitzer/Zinsmeister 2006:16). Diese Definition basiert auf Bensons Definition der Kollokationen als „arbitrary and recurrent word combinations" (zitiert nach Seretan et al. 2004), die auf der direkten Beobachtung von Sprachdaten und nicht auf quantitativen Daten beruht (vgl. Lemnitzer/Zinsmeister 2006:16). Die vorliegende Arbeit versucht einen Überblick über die Schwierigkeiten zu geben, die in der Anfangsphase der Erstellung eines korpusbasierten kroatischen Kollokationswörterbuchs auftraten und mit denen auch andere slawische Sprachen zurechtkommen müssen. Zuerst werden die sprachpolitischen Probleme erwähnt, die bei der Zusammenstellung des nötigen kroatischen Korpus aufkamen, danach wird die computergestützte Bearbeitung des Korpus sowie die Problematik der Auswahl des passenden statistischen Verfahrens erläutert.

2 Braucht man ein korpusbasiertes kroatisches Kollokationswörterbuch?

Wenn man das Curriculum[2] des Kroatischunterrichts im Ausland näher untersucht, stellt man fest, dass der Erwerb von Kollokationen im traditionellen Sinne einen wichtigen Bestandteil des Spracherwerbs ausmacht und dass diese bei der Wortschatzarbeit eine große Rolle spielen. Ihr explizites Lehren ist daher im Curriculum des Kroatistikstudiums (Kroatisch als Zweit- oder Fremdsprache) an der Philosophischen Fakultät in Zagreb vorgesehen. Kollokationen werden im Rahmen von Sprach- und Konversationsübungen[3] gezielt unterrichtet.

Aus den bisherigen Ergebnissen der kroatischen Forschung kann man schließen, dass Untersuchungen überwiegend mit Bezug auf englische Kollokationen (vgl. Špiranec 2005) oder korpusbasierte Kollokationssuche durchgeführt wurden, mit dem Ziel, Übersetzungsäquivalente im Englischen und Kroatischen zu kontrastieren. In den Vorarbeiten zur Erstellung eines Kollokationswörterbuchs wird hauptsächlich der Aufbau der Wörterbucheinträge diskutiert (vgl. Bergovec 2007). Trotz einiger Vorarbeiten besteht derzeit Bedarf nach einem kroatischen Kollokationswörterbuch und zugleich ein Mangel an korpuslinguistischen Daten zum tatsächlichen Sprachgebrauch.

---

1 Übersetzung hier und im Folgenden: MA. Der kroatische Text lautet: „obvezatna ili uobičajena veza riječi koja nije određena gramatikom (*podnijeti molbu ali uložiti žalbu*)".

2 *Kurikulum hrvatske nastave u inozemstvu*: *http://public.mzos.hr/Default.aspx?sec=2116*, gesehen am 6. Juni 2008, weiter im Text: Curriculum.

3 *http://web.archive.org/web/*/www.ffzg.hr/kroat/Files/o_Hrvatski%20kao%20strani%20jezik%203.doc* (10. Juni, 2007), gesehen am 30. Mai 2009.

3 Zur Erstellung eines korpusbasierten Kollokationswörterbuchs

3.1 Analyse der Struktur

Bevor man sich über die Struktur eines korpusbasierten kroatischen Kollokationswörterbuchs Gedanken machen kann, sind einige korpusbasierte Kollokationswörterbücher und Erkenntnisse aus anderen Sprachen näher zu betrachten, bei denen ähnliche Probleme im Bereich der maschinellen Sprachbearbeitung gelöst werden mussten. Im Folgenden werden daher Ergebnisse aus dem polnischen korpusbasierten Kollokationswörterbuchs präsentiert, nämlich Bańkos ‚Wörterbuchs des Guten Stils' (‚Słownik dobrego stylu', im Folgenden: SDS, vgl. Bańko 2007), und mit der möglichen kroatischen Lösung verglichen. Das zu analysierende Wörterbuch wurde aus zwei Gründen gewählt. Zum einen eignet sich das Polnische durch seinen Flexionsreichtum – im Gegensatz zum Englischen – gut für einen Vergleich mit dem Kroatischen. Zum zweiten musste man bei der Erstellung eines polnischen korpusbasierten Kollokationswörterbuchs ähnliche Probleme bei der Lemmatisierung und Tokenisierung behandeln wie im Kroatischen, was verwertbare Lösungsansätze für ein kroatisches Kollokationswörterbuch erwarten lässt. Meines Wissens gibt es kein weiteres korpusbasiertes Kollokationswörterbuch einer südslawischen Sprache, die keine kyrillische Schrift verwendet.

Das SDS ist, wie schon erwähnt, ein auf der Basis von Korpusanalysen zusammengestelltes Kollokationswörterbuch. Im Vergleich mit dem ‚Oxford Collocations Dictionary' (im Folgenden: OCD) ist auffällig, dass das SDS die Lemmata ohne grammatische Informationen anführt. Im Unterschied zu anderen Kollokationswörterbüchern (z.B. Kozłowska 1993), aber ähnlich wie beim ‚BBI Dictionary' (Benson/Benson/Ilson 1990) gibt es in diesem Wörterbuch keine Einteilung der Token nach Wortarten. Das könnte auf das Problem des POS-Tagging[4] bei der Korpusanalyse zurückgeführt werden. Wenn man diese Angaben mit dem kroatischen Beispiel vergleicht, stellt sich die Frage, ob man eine ähnliche Lösung anstreben könnte, oder ob die grammatischen Informationen bei der Erstellung eines kroatischen Kollokationswörterbuchs unverzichtbar wären. Diese Frage ließe sich durch eine Pilotstudie beantworten.

Zudem ist, anders als beim OCD, der veränderbare Teil des Lemmas im Wörterbuch mit einem Querstrich vom Rest des Lemmas getrennt. Diese Lösung beruht wahrscheinlich auf der computergestützten Bearbeitung der flektierenden Sprachen und der automatischen morphologischen Analyse, wodurch Lemmata in Stems (d.h. Bestandteile der Lemmata, die bei der Flexion unverändert bleiben) und Terms (Bestandteile der Lemmata, die sich bei der Flexion verändern und mit grammatischen Markierungen versehen sind) eingeteilt werden[5]. Dem Fremdsprachenlerner hilft das bei der Bearbeitung der Flexion bei den

4 Unter POS-tagging (Part-of-speech Tagging, oder grammatical tagging, word-category disambiguation) versteht man die Zuordnung von Wörtern eines Textes (Korpus) zu Wortarten (engl.: part of speech), wobei sowohl die Definition des Wortes als auch der Kontext berücksichtigt wird.

5 Die traditionellen morphologischen Kategorien werden bei der computergestützten morphologischen Analyse der Sprache (Parsing) absichtlich nicht benutzt, da die im Parser definierten Begriffe mit diesen nicht immer übereinstimmen. Allgemein gesagt, umfasst die

Kollokationen nur dann, wenn er die benötigten Informationen zur entsprechenden Term-artigen Flexionsendung in der Sekundärliteratur finden kann (vgl. Aleksa 2008).

Allgemein gesehen bietet das SDS dem Leser nur sehr knappe Informationen. Für einen Fremdsprachenlerner bedeutet dies, dass er auf andere Lehrwerke zugreifen muss, um den Gebrauch von Kollokationen korrekt zu lernen. Im Fall des Kroatischen stellt sich die Frage, ob in einem zukünftigen kroatischen Kollokationswörterbuch dieses Minimum ausreichend wäre.

3.2 Suche nach Informationen

Wie bekannt, lassen sich Kollokationswörterbücher auf mindestens dreierlei Weise erstellen. Die erste Methode ist ein manuelles Kompilieren von Wörterbucheinträgen, das i. d. R. auf der Intuition des Linguisten beruht. Die zweite Möglichkeit wäre ein halbautomatisches Verfahren, bei dem man nur die vorbestimmten Kollokationen mithilfe verschiedener Web-Suchmaschinen wie z.B. Google untersucht. Das dritte Verfahren ist die Recherche in einem Korpus und die automatische Extrahierung der Kollokationen mithilfe verschiedener statistischer Tools.

Wenn man beispielsweise das zweite Verfahren für das Zusammenstellen eines Kollokationswörterbuchs gewählt hat, stößt man auf zahlreiche Schwierigkeiten. Man möchte nämlich nicht bereits vor der Suche exakt die Wortpaare oder Wortgruppen bestimmen, die untersucht werden sollen, sondern systematisch alle möglichen Wortverknüpfungen aufdecken. Zudem geben die verbreiteten Suchmaschinen keine Auskunft darüber, wie groß die Anzahl der untersuchten Texte ist und wie umfangreich die Texte selbst sind. Daher muss die Relevanz der Ergebnisse in Frage gestellt werden.

Bei der Erstellung eines Kollokationswörterbuchs wäre also die dritte Methode empfehlenswert, da dadurch mit statistischen Verfahren alle relevanten Wortverknüpfungen gefunden werden. Mithilfe verschiedener statistischer Tools können Angaben ermittelt werden, die auf dem tatsächlichen Sprachgebrauch beruhen. Kollokationen und ihre Token können beispielsweise nach der Häufigkeit des gemeinsamen und einzelnen Vorkommens im Korpus sowie nach der Wahrscheinlichkeit ihres Vorkommens im Sprachgebrauch angeordnet werden, was eine große Hilfe bei der Erstellung des Wörterbuchs bedeutet.

Die wichtigsten Fragen, die hier angeschnitten werden, betreffen die Wahl eines geeigneten Verfahrens für die Korpusanalyse, die Lemmatisierung sowie die Herstellung eines relevanten Korpus. Diese werden im folgenden Absatz näher besprochen.

Kategorie von Stems jene Wortteile, die im Laufe der Flexion unverändert bleiben, während der veränderliche Rest des Lemmas als Term bezeichnet werden kann. Wichtig zu betonen ist auch die Tatsache, dass, ähnlich wie bei der traditionellen Morphologie, ein Wort nicht aus einem Ø–Stem, aber aus einem Ø–Term bestehen kann. Am Beispiel des kroatischen Adjektivs *hrvatski* ‚kroatisch' ist zu sehen, dass (im Gegensatz zu den traditionellen Kategorien) *hrvatsk-* den Stem darstellt, gefolgt vom Term *-i*. Bei der Flexion bleibt nämlich *hrvatsk-* unverändert; die anderen Adjektivformen des Paradigmas sind beispielweise *hrvatska*, *hrvatskoga*, *hrvatskih…* Aus diesem Beispiel ist zu sehen, dass die zum Stem *hrvatsk-* hinzugefügten Terms u.a. *-a*, *-oga* und *-ih* sind (vgl. Aleksa 2008).

3.3 Sprachspezifische Problematik

A. Zusammenstellung und Bearbeitung eines Korpus

Das erste Problem im Zusammenhang mit der Erstellung eines kroatischen korpusbasierten Kollokationswörterbuchs betrifft die Zusammenstellung eines relevanten Korpus. Da im Kroatischen infolge politischer Prozesse einige sprachpolitische Probleme noch ungelöst geblieben sind, bedeutet das Zusammenstellen eines relevanten Korpus eine zusätzliche Herausforderung (ausführlicher dazu vgl. Aleksa 2006).

In der Zeit der neuen Eigenständigkeit des Kroatischen nach 1990 entstanden sprachpuristische Tendenzen, welche, unterstützt durch die Massenmedien, zu zahlreichen Versuchen führten, das serbische Wortgut aus dem gegenwärtigen kroatischen Sprachgebrauch zu entfernen.[6] Es ist jedoch nachgewiesen worden, dass serbische Lexeme weiterhin im modernen Kroatischen existieren. Damit stellt sich die Frage, ob man ein zusammengestelltes Korpus eigentlich als ‚kroatisches' Korpus bezeichnen darf, und inwieweit ein Korpus als repräsentativ bezeichnet werden könnte. Die sprachpolitische Problematik und die Problematik der Repräsentativität von Korpora werden im Rahmen dieser Arbeit nicht ausführlicher erörtert.

Da das kroatische Nationalkorpus nicht frei zugänglich ist, wurde für die Erstellung eines kroatischen Kollokationswörterbuchs in dieser Projektphase ein Testkorpus von 5 Millionen Wörtern kompiliert, das nach dem Muster des Kroatischen Nationalkorpus aus 5 Subkorpora besteht: einem Subkorpus mit literarischen Texten, einem Korpus mit Texten aus den schriftlichen Medien, ein Korpus mit Texten aus Lehr- und Fachbüchern und einem mit Texten aus dem Internet (Webseiten und Blogs), die eine Art Kodifizierung der gesprochenen Sprache beinhalten (vgl. Aleksa 2006). Das Nationalkorpus wird ständig erweitert.

Für die Durchsuchung der Korpora steht eine Menge statistischer Software zur Verfügung. Zuvor muss das Korpus jedoch vorbereitet, d.h. von Elementen, die sich für die Suche nach Kollokationen als unwichtig erwiesen haben, gereinigt werden. Zu diesen Elementen gehören z.B. falsch gedruckte Zeichen, Striche, Klammern etc., d.h. alles, was sich für die N-gram-Bestimmungen als irrelevant erwiesen hat (mehr zu N-grams im Abschnitt 3.4.2). Danach sollten die Texte tokenisiert und lemmatisiert werden. Viele statistische Programme (wie z.B. NSP) haben einen großen Nachteil: Sie sind für die im morphologischen Sinne flexionsarme englische Sprache entwickelt worden, daher hat sich die Bearbeitung einer flektierenden Sprache wie des Kroatischen (und auch der deutschsprachigen Texte) als schwierig erwiesen. Die damit zusammenhängenden Probleme ließen sich für das Kroatische durch den von Aleksa und Wołosz in Zusammenarbeit mit der Firma

6 Dass das nicht erfolgreich war, hat die Analyse des von Melita Aleksa und Robert Wołosz zusammengestellten schriftlichen Subkorpus der gegenwärtigen kroatischen Sprache gezeigt. Einige serbische Lexeme kommen noch immer nicht nur in kroatischen Texten, sondern auch in den Medien sehr häufig vor (z. B. *ponekad* statt *katkad* ‚manchmal', oder *stepenice* statt *stube* ‚Treppe') (vgl. Aleksa 2006).

MorphoLogic entwickelten morphologischen Analysator HUMOR lösen[7] (vgl. Aleksa 2006).

B. Die Wahl des geeigneten Verfahrens

Die verfügbaren statistischen Verfahren und Programme bieten eine Vielfalt von Möglichkeiten für das Auffinden von Kollokationen in einem Korpus. Die Software-Pakete dienen in erster Linie der raschen Durchsuchung des Korpus nach Kollokationen, d.h. nach Kombinationen aus zwei, drei oder theoretisch auch unbegrenzt vielen Wörtern, auch wenn diese durch eines oder mehrere Wörter bzw. durch Satzzeichen getrennt sind. Wichtig ist, dass die zu untersuchenden Einheiten oder Token nicht nur aus einem Wort bestehen müssen, sondern auch mehrere Wörter enthalten können. Auf diese Weise können auch Relationen zwischen zwei (Bigrams), drei (Trigrams) und theoretisch unbegrenzt vielen Wörtern (N-grams), aber auch Relationen zwischen einer Wortgruppe als einer zu untersuchenden Einheit oder einem Token und anderen Wörtern bzw. Wortgruppen untersucht werden. Viele Programme (wie z.B. NSP) bieten dem Benutzer die Möglichkeit, zwischen den links- und rechtspositionierten Token zu unterscheiden, was bei der Kompilation eines Wörterbuchs eine große Rolle spielt (vgl. Aleksa/Wołosz 2009).

Im Folgenden wird auf die Vorarbeiten sowie auf die Ergebnisse eines Testvergleichs verschiedener statistischer Tools eingegangen. Es wurden vier populäre, frei erhältliche Programme, ‚Ngram Statistics Package' (im Folgenden: NSP), ‚Kwic Concordance for Windows' (im Folgenden: KWIC) und ‚Collocation Extract' (im Folgenden: COLLEX) verglichen und an einem aus 45.150 Token bestehenden, nicht lemmatisierten kroatischen Testkorpus getestet. Der Test beschränkte sich auf die Suche nach Bigrams. Obwohl die Programme COLLEX und KWIC bei umfangreicheren Korpora in einem Pilottest gescheitert sind (vgl. Aleksa 2009), wurden sie hier noch einmal getestet. Sie erwiesen sich auch bei einigen Aufgaben an diesem kleinen Testkorpus als unstabil. Erst im zweiten oder dritten Anlauf gelang der Versuch.

NSP und COLLEX wurden für Sprachen ohne Sonderzeichen (z. B. Acut, Caron, Querstrich oder Umlaut) entwickelt. Man muss diese daher entweder in der Tokendefinition berücksichtigen oder die Texte vor der tatsächlichen Analyse bearbeiten und die Sonderzeichen durch intern ausgearbeitete Codes ersetzen, was man später automatisch rückgängig machen kann. Wenn man die Programme näher betrachtet, stellt man fest, dass für die Arbeit mit NSP gewisse Vorkenntnisse der Programmiersprache Perl erforderlich sind, was bei COLLEX und KWIC nicht der Fall war. Während COLLEX vier (Chi-square,

7 HUMOR, oder High-Speed Unification Morphology ist ein von MorphoLogic entwickelter unifikationsbasierter morphologischer Parser, der in erster Linie der morphologischen Analyse von Sprachen dient und als Grundlage vieler maschineller Übersetzungsprogramme fungiert. Das Programm wird unter anderem als Basis für Übersetzungssysteme wie MobiMouse, MobiDic und MetaMorpho gebraucht. Bisher wurde HUMOR sowohl für das System der agglutinierenden Sprachen, als auch der flektierenden Sprachen implementiert, was auch die unterschiedlichen sprachlichen Versionen der oben genannten Übersetzungsprogramme und Übersetzungstools erklärt. Seine Vorteile liegen u. a. in der hohen Verarbeitungsgeschwindigkeit, aber auch in der Möglichkeit, den Parser allen Sprachsystemen anzupassen.

Mutual Information, Direct Likelihood und Raw Likelihood) und NPS dreizehn statistische Verfahren zur Kollokationssuche (Dice Coefficient, Fishers exact test - left sided, Fishers exact test - right sided, Fishers twotailed test - right sided, Jaccard Coefficient, Log-likelihood ratio, Mutual Information, Odds Ratio, Pointwise Mutual Information, Phi Coefficient, Poisson Stirling Measure, T-score) bietet, ermöglicht KWIC nur eine Konkordanzsuche und eine gezielte Suche nach eigentlichen Kollokationen und ihren links- oder rechtspositionierten Token.

Wenn man mithilfe von KWIC nach Kollokationen sucht, bekommt man eine Liste mit der Vorkommenshäufigkeit einzelner Wörter und eine Berechnung der Anzahl der Token, der unterschiedlichen Lexeme und das Verhältnis von Lexemen und Token (‚type/token ratio') (Abbildung 1 auf der nächsten Seite).

KWIC erlaubt nur eine gezielte Suche nach Kollokationen mit vorgegebenen Bestandteilen, wobei das Programm eine Tabelle mit beispielsweise fünf Wörtern (vom angegebenen Token jeweils links und rechts positioniert) präsentiert (Tabelle 1 auf der nächsten Seite).

Eine Berechnung des Wahrscheinlichkeitsgrades, der Vorkommenshäufigkeit der Kollokationen oder ihre Rangierung ist mit KWIC nicht möglich.

Wenn man sich die Optionen der anderen zwei Tools im Hinblick auf die Ziele dieser Arbeit anschaut, stellt man fest, dass sie unterschiedliche Möglichkeiten und statistische Verfahren bieten. COLLEX ermöglicht die manuelle Eingabe vorgegebener Einheiten bei der Suche nach Kollokationen auf der linken, auf der rechten oder auf beiden Seiten; des weiteren werden Mindestfrequenzen sowie die Signifikanzniveaus (p-Wert) angegeben. Zudem ermöglicht das Programm auch die Suche nach Kollokationen für eine Anzahl N Token, die sogar durch mehrere Wörter oder Zeichen voneinander getrennt sein können. Wie es die nachfolgenden Abbildungen zeigen, können die Kollokationen, dem Verfahren entsprechend, nach der Signifikanz (Tabelle 2) oder nach der Frequenz (Tabelle 3) angeordnet sein.

NSP bietet zwar auch alle diese Möglichkeiten, erfordert aber, wie schon erwähnt, Perl-Kenntnisse (vgl. Aleksa/Wołosz 2009). In Tabelle 4 ist ein Auszug aus den Ergebnissen der T-Score Bigramanalyse des Testkorpus mithilfe von NSP zu sehen, wobei die Zahlen den Rang des Bigrams, das Ergebnis des statistischen Tests, die Frequenz des Bigrams und die Frequenzen jedes einzelnen Tokens im Bigram angeben.

Je nach Test bringt die Berechnung unterschiedliche Ergebnisse. Daraus kann man schlussfolgern, dass bei der Zusammenstellung eines Kollokationswörterbuchs zielgerichtete Tests durchgeführt werden sollten, mit deren Hilfe man zuverlässige Ergebnisse veröffentlichen kann (vgl. Grzybek 2007: 196-202). Da NSP die Möglichkeit bietet, die Tests selbst zu gestalten, wurde dieses Programm schließlich für die Zusammenstellung eines korpusbasierten kroatischen Kollokationswörterbuchs gewählt. Unabhängig von der Wahl des Testverfahrens und des Programms ist eine weitere, manuelle Bearbeitung der Ergebnisse bezüglich der Auswahl repräsentativer Kollokationen in dem Prozess der Lemmaselektion auf jeden Fall nötig.

661 bolnici. ... na vježbi. ... u banci. ...(0 ""n u knjižnici.ivan je dopodne išao u park
569 u. koji tramvaj vozi u centar? tramvaj broj. 1. hvala lijepa, vrlo ste ljubazni! dovidenja!g
575 dom vozi tramvaj br. 2 (dva), a ne br. 1! ali jedan građanin je rekao da ovaj tramvaj
708 građanin je odgovorio da tamo vozi tramvaj broj 1. tramvaj je brzo stigao, günther i helga n
710 je studentski dom vozi tramvaj br. 2, a ne br. 1. oni su se ipak vozili u centar, a zatim su
1647 (prijatelj) ... žena vrlo je bolesna. (otac)1... bolest nije teška. i (šoljan) ...
2147 se rado se vozim brodom. rado bih se vozio 1 kupati se rado se kupam u moru. ...sun
2251 ki ured. ta se ulica ne nalazi u ovom d ij e 1 u grada. od koga su g. i h. to saznali?od
2445 . mjesta)... prometni znak nisam primijetio; 1. ne nalazi se na vidljivom mjestu. 2.
2446 vidljivom mjestu. 2. automobili su se sudarili; 1. vozač nije poštivao prednost drugoga. 2. v
2447 prednost drugoga. 2. vozači su ozlijeđeni; 1. sudar automobila bio je vrlo jak. 2. vozač je
2448 bio je vrlo jak. 2. vozač je platio kaznu; 1. parkirao je kola na zabranjenom mjestu. 2.
2449 zabranjenom mjestu. 2. promet je zaustavljen; 1. dogodila se prometna nezgoda. 2. nije s
2450 se prometna nezgoda. 2. nije smio voziti auto; 1. nije imao vozačku dozvolu. 2. günther i h

Abbildung 1: Auszug aus den Ergebnissen einer Konkordanzsuche mit KWIC am Testkorpus

Keyword: jesam												
Word	Total	-5	-4	-3	-2	-1	0	1	2	3	4	5
a	1	0	0	0	0	0	-	1	0	0	0	0
anje	1	1	0	0	0	0	-	0	0	0	0	0
anu	1	0	0	0	0	1	-	0	0	0	0	0
ao	1	0	0	0	0	0	-	0	0	0	1	0
bio	6	0	1	2	0	0	-	0	2	1	0	0
bolestan	1	0	1	0	0	0	-	0	0	0	0	0
cijelu	1	0	0	0	0	0	-	0	0	1	0	0
da	3	0	0	0	0	0	-	1	0	0	2	0
danas	1	0	1	0	0	0	-	0	0	0	0	0

Tabelle 1: Auszug aus den Ergebnissen der Analyse des Testkorpus mit KWIC

Word1	**Freq1**	**Word2**	**Freq2**	**Freq12**	**ll**
čitali	6	odlomak	13	6	100,742 03
hvala	23	dobro	104	11	102,250 22
su	713	razgovarali	46	20	102,923 31
mala	8	vlati	7	6	103,947 96
ti	124	si	117	17	105,257 47
bit	11	će	155	10	106,741 37
u	1053	zagreb	24	18	107,626 53
prošle	8	godine	54	8	108,341 65
narodne	8	nošnje	6	6	109,680 78

Tabelle 2: Auszug aus den Ergebnissen der Analyse des Testkorpus mit Collocation Extract (Test: Raw Frequency)

Word1	Freq1	Word2	Freq2	Freq12	Freq
su	713	se	992	100	100
da	467	je	1909	85	85
je	1909	li	336	77	77
kod	154	kuće	91	75	75
što	267	je	1909	70	70
to	398	je	1909	68	68
gdje	127	je	1909	59	59
günther	127	i	1139	55	55

Tabelle 3: Auszug aus den Ergebnissen der Analyse des Testkorpus mit Collocation Extract (Test: Raw Frequency)

Bigram		Rang	T-Score	Freq12	Freq1	Freq2
kod	kuće	1	8,5047	73	154	89
su	se	2	8,2650	100	713	992
u	seminaru	3	7,0248	52	1053	52
je	li	4	6,9828	77	1908	336
günther	i	5	6,9423	55	126	1137
i	helga	6	6,9310	54	1137	110
gdje	je	7	6,9073	59	127	1908
što	je	8	6,8730	70	267	1908
da	je	9	6,8488	85	467	1908
ja	sam	10	6,4296	44	171	322

Tabelle 4: Auszug aus den Ergebnissen der Analyse des Testkorpus mit NSP (Test: T-score)

4 Fazit und Ausblick

Wenn man die Problematik bei der Erstellung eines korpusbasierten Kollokationswörterbuchs des Kroatischen zusammenfasst, ist festzuhalten, dass hierbei viele Entscheidungen getroffen werden müssen, bevor die tatsächliche lexikographische Arbeit beginnen kann. Zunächst ist das Problem der Erstellung eines „guten" Korpus und der Lemmatisierung zu lösen; erst im nächsten Schritt kann die statistische Analyse des Korpus in Angriff genommen werden. Eine ausführliche Untersuchung der statistischen Möglichkeiten sowie die Berücksichtigung der Ergebnisse bisheriger Forschung (vgl. Jurisch 2003, 2005; Grzybek 2007) schaffen eine solide Grundlage für die in naher Zukunft bevorstehende Erstellung eines kroatischen korpusbasierten Kollokationswörterbuchs.

Literaturverzeichnis

Aleksa, M. (2006): „Automatic Morphological Analysis of the Croatian Language: The Verbal, Adjectival and Nominal Inflections within the Morphological Parser HUMOR". In: Gyuris (Hrsg.) (2006). *http://www.nytud.hu/cescl/proceedings/Melita_Aleksa_CESCL.pdf*, gesehen am 14. Mai 2009.

Aleksa, M. (2008): „HUMOR als Basis für maschinelle Übersetzungsprogramme: Morphologische Analyse der kroatischen und deutschen Adjektivformen". In. Karabalić/Omazić (Hrsg.) (2008); 219-232.

Aleksa, M./Wołosz, R. (2009): „Automatische Suche nach phraseologischen Einheiten in kroatischen und polnischen Texten". In: Földes (Hrsg.) (2009) (im Druck).

Anić, V. (2004): *Rječnik hrvatskoga jezika.* Zagreb.

Bańko, M. (2007): *Słownik dobrego stylu.* Warszawa.

Bergovec, M. (2007): „Leksički pristup u nastavi stranih jezika s posebnim osvrtom na hrvatski", in: *LAHOR: Časopis za hrvatski kao materinski, drugi i strani jezik, 3*; 53-66.

Benson, E./Benson, M./Ilson, R. (1990): *Kombinatornyj slovar' anglijskogo jazyka.* Moskva.

Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N.R. (Hrsg.) (2007): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin/New York.

Duden (1997): *Deutsches Universalwörterbuch.* Mannheim.

Fleischer, W. (1997): *Phraseologie der deutschen Gegenwartssprache.* 2. durchgesehene und ergänzte Auflage. Tübingen.

Földes, C. (2009): *Phraseologie disziplinär und interdisziplinär.* Tübingen.

Grzybek, P. (2007): „Semiotik und Phraseologie". In: Burger et al. (Hrsg.) (2007); 188-208.

Gyuris, B. (Hrsg.) (2006): *CESCL1, Proceedings of the First Central European Student Conference in Linguistics*. Budapest.

Jurish, B. (2005): „Hybrid Syntactic Category Induction". Vortrag an der CPALA, Split, Kroatien, 25-27 Juli, 2005. *http://www.ling.uni-potsdam.de/~moocow/pubs/talk-split-slides.pdf*, gesehen am 27. März 2009.

Jurish, B. (2003): „Part-of-Speech Tagging with Finite-State Morphology". Poster an der Konferenz Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives , Berlin, 18.-20. September, 2003. *http://www.ling.uni-potsdam.de/~moocow/pubs/kollok2003.pdf*, gesehen am 27. März 2009.

Karabalić, V./Omazić, M. (Hrsg.) (2008): *Istraživanja, izazovi i promjene u teoriji i praksi prevođenja. Explorations, challenges and chances in translation theory and practice. Theorie und Praxis des Übersetzens*. Osijek.
Kozłowska, C.D. (1993): *Selected English collocations.* Warszawa.
Lemnitzer, L./Zinsmeister, H. (2006): *Korpuslinguistik. Eine Einführung.* Tübingen.
Oxford Collocations Dictionary for Students of English (2003). Oxford.
Seretan, V./Nerima, L./Wehrli, E. (2004): „A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora". In: Williams/Vessier (Hrsg.) (2004); 755-766. *http://www.latl.unige.ch/personal/vseretan/publ/EURALEX2004_VS_LN_EW.pdf*, gesehen am 20. Januar 2008.
Špiranec, I. (2005): „Priroda i upotreba kolokacija: Primjeri iz tehničkoga engleskog jezika", in: *Strani Jezici, 34/3*; 219-227.
Williams, G./Vessier, S. (Hrsg.) (2004): *Proceedings of the XI EURALEX International Congress.* Lorient.

Software

Collocation Extract: *http://pioneer.chula.ac.th/~awirote/colloc/*, gesehen am 10. Juni 2006.
Kwic Concordance For Windows: *http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html*, gesehen am 12. Juni 2006.
Ngram Statistics Package: *http://www.d.umn.edu/~tpederse/nsp.html*, gesehen am 12. Juni 2006.

Melita Aleksa Varga
Abteilung für deutsche Sprachwissenschaft und angewandte Linguistik
University of Osijek
L. Jaegera 9
31000 Osijek
Croatia
maleksa@ffos.hr

III

Korpora und Web in der phraseographischen Praxis

Online-Datenerhebungen im Dienste der Phraseographie

Marcel Dräger/Britta Juska-Bacher

Many questions that are the focus of modern phraseography cannot be answered by methods used in traditional phraseography. Corpus linguistics has advanced our understanding in some aspects, in particular the representation of meaning and variants of frequent phrasemes, but the geographical distribution of phrasemes and the representation of meaning and variants of rare phrasemes still present a challenge. In this article we shed light on the potential of other digital methods of data collection, in particular online surveys and online dictionaries as interfaces for phraseological data collection. Online surveys with large numbers and wide geographical distribution of informants offer relevant current data on the representation of meaning and variants, areal aspects and acquaintance of (even rarely used) phrasemes. A major problem of phraseography is the immense expenditure of time needed for collecting and editing phraseological data. To optimize the process we suggest to establish and maintain an online dictionary that offers phraseological information, while at the same time incorporating users' feedback. In this way, one can collect – with little additional expenditure – data on the representation of meaning and variants of phrasemes, their remotivation and geographical distribution.

1 Einleitung

Betrachtet man die phraseologischen Einträge in allgemeinsprachlichen wie auch in phraseologischen Wörterbüchern genauer, so lassen sich vier, nicht immer ganz offensichtliche Quellen ausmachen, aus welchen sich die phraseographischen Informationen speisen (Dräger eingereicht-b). Grundlage älterer Nachschlagewerke war vorrangig das, was der Autor aus seiner sprachlichen Intuition und seinem direkten sprachlichen Umfeld ableiten bzw. aus dem Lateinischen übertragen konnte (vgl. bspw. Agricola [1534] und Frisch [1741]). Mit den monumentalen Wörterbuchprojekten von Adelung (1793-1801), Campe (1807) und Grimm (1854-1960) gewinnen die Mitarbeiter und Korrespondenten als Quelle lexikographischer Information an Bedeutung, denn von einer Person ließen sich Werke in diesem Umfang nicht mehr bewerkstelligen. Gängige Praxis war daher, dass Belege und Erläuterungen von Beiträgern eingereicht und – oft wohl kaum nachgeprüft – von den Autoren in die Wörterbücher übernommen wurden (besonders bei Wander 1867-1880).[1]

1 Zur lexikographischen Praxis des 18. und 19. Jahrhunderts vgl. Haß-Zumkehr (2001).

Mit einem zunehmenden Bestand an Wörterbüchern wuchs auch die dritte bis heute wichtige Quelle lexikographischer Information an. Unbestreitbar und mehrfach nachgewiesen zählt das ‚Recycling' der Vorgängerwerke über viele Jahrzehnte zur nicht offiziellen, aber gängigen lexikographischen Praxis. Ausschlaggebend für diese mehrfache Wiederverwertung lexikographischer Inhalte sind verfahrensökonomische Gründe, denn weitere Quellen wie Fach- und Sekundärliteratur sind nur sehr schwer und mit hohem Zeitaufwand erschließbar und empirische Daten fehlen größtenteils noch. Erst der Einsatz von Korpora (vierte Quelle) erleichtert die lexikographische Arbeit um ein Vielfaches und bietet neue, von den wiederverwerteten Inhalten unabhängige Erkenntnisse. Zwar gilt es gerade in der Phraseologie noch einige verfahrenstechnische Fragen zu lösen (Dräger eingereicht-a), doch zeigt sich schon jetzt am Beispiel der Kollokationsanalyse der enorme inhaltliche und zeitliche Gewinn des korpusbasierten Arbeitens. Das Korpus ersetzt damit den Zettelkasten historischer Werke, es erlaubt einen Zugriff auf umfassende und nicht nur die exzerpierten Sprachdaten, Suchergebnisse lassen sich in Zweifelsfällen leichter überprüfen und ergänzen, und das gewonnene Material liegt bereits digital zur Weiterverarbeitung bereit.

Doch Korpora – so beeindruckend ihr Erfolg in der Linguistik gerade ist – können nicht alle Fragen beantworten und haben, je nach Zusammensetzung und Ausrichtung, auch und gerade für die Phraseographie einige Nachteile. In diesem Aufsatz wollen wir daher auf der Basis einer kurzen Bedarfsanalyse für die einsprachig ausgerichtete Phraseographie die Frage stellen, welches Potential in alternativen Methoden der digitalen Datenerhebung steckt. Als solche Methoden, die wir hier unter dem Sammelbegriff „Online-Datenerhebungen" zusammenfassen, wollen wir in den folgenden Abschnitten auf Online-Befragungen und einen interaktiven Ansatz eines Online-Lexikons als Sammelstelle für indirekte phraseologische Informationen eingehen.

Den einschlägigen Veröffentlichungen zum Thema folgend (u. a. Kempcke 1986, Pilz 2002, Schemann 1989, Steffens 1989) sollte ein solider phraseographischer Artikel folgende Informationen zu einem Phrasem liefern: Angabe einer gebräuchlichen Nennform und von Varianten, Bedeutungsangabe, Hinweise zu möglichen Restriktionen sowie pragmatische Angaben. Je nach Fachausrichtung und Interessenlage können hierzu etymologische Erklärungen (Mokienko 2002), Angaben zur arealen Verteilung (Piirainen 2001, Schmidlin 2001), zum Bekanntheitsgrad respektive der Verwendungshäufigkeit (wichtig u. a. für die Phraseodidaktik, siehe Arbeiten zur Erstellung sog. phraseologischer Minima oder Optima von Grzybek 1991, Ďurčo 2001, Hallsteinsdóttir et al. 2006), zu Textsortenpräferenzen (Kühn 1994) und einige andere Detailinformationen kommen.

Die oben genannte dritte phraseographische Quelle, also die bestehenden Wörterbücher, sind bei vielen dieser Fragen überfordert. Am ehesten lassen sich dort Nennformen, Varianten wie auch (historische) Bedeutungserläuterungen finden. Einzelne Werke, wie beispielsweise das *Deutsche Sprichwörterlexikon*, liefern einige Hinweise zu arealen Aspekten, während die Bereiche Pragmatik und Restriktionen vollständig fehlen. Des Weiteren haftet nahezu allen Nachschlagewerken aufgrund einer weitestgehend fehlenden wissenschaftlichen Dokumentation der Quellen (die gibt es lediglich für die Belege) der Mangel nicht nachprüfbarer Erkenntnisse an. Für die aktuelle Lexikographie muss man also die Frage stellen, wie viel Energie und Zeit man in die Überprüfung des historischen Materials steckt

und was man einfach – im Idealfall mit Verweis auf die Quelle – übernimmt. Beispielhaft mag man sich vorstellen, welcher Aufwand hinter einer fachlich fundierten Überprüfung der in Röhrichs *Lexikon der sprichwörtlichen Redensarten* (2002) teilweise bis zu zehn (s. v. „jemanden ins Bockshorn jagen") unbewertet nebeneinander aufgeführten Etymologien steckt.

Mithilfe größerer Korpora lassen sich einige dieser Defizite beseitigen. Allen voran bietet sich die Suche nach im Sprachgebrauch[2] existierenden Varianten an, wobei hier aufgrund der spezifischen Problematik bei der Suche nach Phrasemen (Dräger eingereicht-b, Heid 2007, Rothkegel 2007, Steyer 2003) – abgesehen von der schon weiterreichenden Kollokationssuche – unterm Strich nur finden lässt, wozu man schon mehr oder weniger Anhaltspunkte hat. Ein zweiter – und vielleicht für die künftige Phraseographie noch bedeutenderer – Aspekt ist die Ermittlung einer Bedeutung aus dem Kontext der Phraseme. Hier bietet sich vor allem die Möglichkeit, in anderen Wörterbüchern angeführte Bedeutungen auf ihre tatsächliche Existenz in Ausschnitten der Sprachverwendung zu überprüfen. Des Weiteren lassen sich recht verlässlich mögliche Textsortenpräferenzen und statistische Aussagen zur Verwendungshäufigkeit eines Phrasems eruieren, wobei hier stets bedacht werden muss, dass Phraseologismen in Korpora relativ selten vorkommen (vgl. Colson 2003: 50), selbst die größten Korpora nur Sprachausschnitte darstellen und daher weite Teile der nicht schriftlich dokumentierten Sprachverwendung auch nicht per Korpusrecherche zugänglich sind.

Weiterer Informationsbedarf, der in Wörterbüchern häufig der Experten-Intuition des Bearbeiters überlassen bleibt (vgl. Rothkegel 2007: 1028), besteht also hinsichtlich der folgenden Punkte: Bedeutungsparaphrasen müssen auf ihre gegenwärtige wie historische Gültigkeit und Genauigkeit hin geprüft werden und sollten in der Regel auch spezifischer, gegebenenfalls durch Ansetzen mehrerer verschiedener Bedeutungen, ausgeführt werden. Konnotationen lassen sich in der Regel noch schwieriger aus dem Kontext eines Phrasems ableiten, hierfür wäre es notwendig, die Einstellung des Sprechers und Rezipienten zu kennen. Anhand von Korpora lässt sich nur feststellen, welche Varianten eines Phrasems realisiert wurden, welchen potentiellen Restriktionen die Verwendung allerdings unterworfen ist, lässt sich nicht klären. Für eine umfassende Aufarbeitung arealer Aspekte der Phraseologie fehlen bislang die notwendigen Korpora. Auch bleibt fraglich, ob das dann anzutreffende Material in seiner Fülle noch sinnvoll zu gliedern wäre. Auch bei pragmatischen Angaben erlauben Korpora lediglich den Rückschluss von der Textsorte oder vom Kontext auf die pragmatische Verwendungsmöglichkeit eines Phrasems. Mit Korpora ließen sich, wie oben erwähnt, statistische Erkenntnisse zur Verwendungshäufigkeit eines Phrasems gewinnen, Aussagen über dessen Bekanntheitsgrad lassen sich daraus aber nicht eins zu eins ableiten. All diese Informationen erfordern empirische Daten der Sprachbenutzer, was allerdings (fast) nur zeitgenössisch möglich ist und gerade für die ältere historische Phraseographie aufgrund der fehlenden sprachlichen Kompetenz ein enormes Problem darstellt (Filatkina 2007: 226).

---

2 Der Sprachgebrauch wird natürlich nur in dem Ausschnitt erfasst, in welchem ihn das Korpus abbildet.

Im Folgenden werden wir in einem ersten exkursorischen Versuch ausloten, inwiefern sich spezifisch für die Phraseographie computergestützt Daten erheben lassen, welche die oben angedeuteten Defizite weiter minimieren und damit die Qualität phraseographischer Produktionen verbessern.

2 Online-Befragungen

Eine potentielle phraseographische Informationsquelle, die bisher nur punktuell genutzt wurde, steht mittels einer Befragung von (möglichst vielen) Sprachbenutzern zur Verfügung. Die Zurückhaltung beim Einbezug dieser fünften möglichen Quelle hängt mit dem vergleichsweise großen Aufwand der empirischen Methode Befragung zusammen. Das Erstellen, Testen, der postalische Versand, das Retournieren, das Erfassen und Auswerten eines Fragebogens ist zeit- wie kostenintensiv und die Kapazität der Probanden bleibt auf eine relativ geringe Zahl von Phrasemen begrenzt. Die Möglichkeit, Befragungen online durchzuführen, ist mit einer Reihe von Vorteilen verbunden. Bevor wir auf diese näher eingehen, möchten wir den Begriff „Online-Befragung" für diese Arbeit definieren.

Unter „Online-Befragung" werden im Allgemeinen mindestens drei Realisierungsmöglichkeiten elektronischer Befragungen zusammengefasst:

1. Die potentiellen Teilnehmenden erhalten den Fragebogen per E-Mail und senden ihn ausgefüllt auf demselben Wege zurück.

2. Der Fragebogen wird auf einem Server abgelegt, von dort heruntergeladen und per E-Mail zurückgesandt.

3. Der auf dem Server abgelegte Fragebogen wird von den Probanden online ausgefüllt.

Wir beschränken den Begriff „Online-Befragung" hier auf die dritte Möglichkeit, d.h. das Online-Ausfüllen eines auf dem Server abgelegten Fragebogens. Die Daten der Probanden werden elektronisch zum Server übermittelt, dort abgespeichert und können jederzeit vom Untersuchenden abgerufen werden beziehungsweise sofort statistisch weiterverarbeitet oder ausgewertet werden.

Online-Befragungen werden in der Linguistik besonders im Bereich der Dialektologie (hier wiederum bevorzugt im Wortschatzbereich) genutzt (vgl. z.B. Elspaß/Möller 2006, Juska-Bacher im Druck), in der Phraseologie im Allgemeinen, in der Phraseographie im Besonderen ist man beim Einsatz dieser Methode deutlich zurückhaltender (vgl. Juska-Bacher 2009: 74–76).

Im Vergleich zum postalisch versandten Fragebogen reduziert die Online-Befragung deutlich den Zeit- und Kostenaufwand (bei der Datenerhebung wie bei der -erfassung), die Fragebögen können attraktiver gestaltet, mit erklärenden Informationen versehen und – unabhängig vom Standort des Forschenden – praktisch weltweit verbreitet werden. Auf diese Weise können mit gegebenem Aufwand wesentlich mehr Probanden rekrutiert und damit phraseographisch relevante Informationen erhoben werden. Der zentrale Kritikpunkt an

dieser Methode, der auf die passive Rekrutierung[3] der Probanden sowie damit verbunden eine mangelnde Repräsentativität der Ergebnisse für die gesamte Sprachgemeinschaft[4] zielt, hat für phraseographische Ansätze nur dort Bedeutung, wo quantifizierende Aussagen gemacht werden sollen, beispielsweise bei der Ermittlung des Bekanntheits- oder Verwendungsgrads (s.u.). Aufgrund der großen und breit gestreuten Probanden(zahl) kann eine Online-Befragung zuverlässige Angaben liefern, die über die Möglichkeiten der bisher für die Phraseographie genutzten Quellen hinausgehen und einige der eingangs formulierten Informationslücken füllen. Die folgenden Beispiele entstammen einer Online-Befragung rund 850 deutschsprachiger Probanden (siehe Juska-Bacher 2009).

Die auf der Basis von Experten-Intuition, älteren Wörterbüchern oder Korpusanalysen formulierten Bedeutungsparaphrasen können anhand einer soliden Datenbasis bezüglich ihrer aktuellen Gültigkeit bestätigt, differenziert, gewichtet oder auch korrigiert werden. Mit Hilfe von Korpusanalysen können Bedeutungsbestimmungen über den Kontext in der Regel nur für Phraseme vorgenommen werden, die in der schriftlichen Sprachverwendung häufig vorkommen. Durch Online-Befragungen einer großen Zahl von Muttersprachlern hingegen können gezielt Bedeutungen und ihre Nuancierungen auch für seltenere Phraseme und für den mündlichen Sprachgebrauch erfragt werden. Diese Abfrage setzt allerdings einen hohen Reflexionsgrad und hohes metasprachliches Bewusstsein der Probanden voraus. Dass ihnen diese anspruchsvolle Aufgabe durchaus zuzutrauen ist, zeigt die oben genannte Befragung, in der die Probanden für das Verbalphrasem *jmdn. an der Nase herumführen* die im Fragebogen vorgegebene Bedeutung „jmdn. täuschen, irreführen" (vgl. auch Duden 11) durch die Angabe „(jmdn.) anführen", „beschwindeln", „leichter Betrug", „veralbern", „veräppeln", „verulken" oder „zum Narren halten" mehrfach explizit abschwächten.

Auch Angaben zu gebräuchlichen Varianten können bei entsprechender Gestaltung der Fragen (s.u.) gewonnen werden. Aufgrund einer großen Zahl von Muttersprachlern ist eine deutlich zuverlässigere Ermittlung allgemein geläufiger Varianten möglich. Beispielsweise konnten mit Hilfe der erwähnten Online-Befragung verschiedene Varianten des Nominal- bzw. Verbalphrasems *zwei Fliegen auf einen Schlag*, *zwei Fliegen mit einer Klappe*, *zwei Fliegen mit einem Streich ((er)schlagen/erledigen)* ermittelt werden, während

3 Die Untersuchenden haben keine umfassende Kontrolle darüber, wer den Online-Fragebogen im Internet findet und wer tatsächlich an der Befragung teilnimmt, sondern potentielle Probanden entscheiden selbst, ob sie aktiv werden möchten.

4 Gemäß Bandilla (1999: 15) weichen Teilnehmende einer Online-Befragung hinsichtlich dreier wichtiger soziodemographischer Faktoren deutlich von der Normalverteilung der Bevölkerung ab: Erstens sind die Teilnehmer deutlich jünger als der tatsächliche Altersdurchschnitt der Bevölkerung. Zweitens ist der Anteil weiblicher Teilnehmerinnen geringer als der männlicher (lt. Bandilla ca. 30% gegenüber 70%). Und drittens sind Teilnehmer mit höherem Bildungsabschluss überrepräsentiert. Diese Tendenzen wurden in linguistischen Online-Befragungen nur teilweise bestätigt. So weisen eine phraseologische Untersuchung von Juska-Bacher (2009: 85f) sowie zwei dialektologische Online-Befragungen von Elspaß/Möller (2006) und Juska-Bacher (im Druck) zwar ebenfalls eine linksgipflige Altersstruktur, hingegen aber einen ausgeglichenen Anteil von Frauen und Männern auf. Die Vermutung liegt nahe, dass sich seit Entstehen von Bandillas Arbeit Ende der 1990er Jahre das Internet-Nutzungsverhalten von Frauen dem der Männer angeglichen hat und/oder das Interesse an sprachlichen Fragen und damit die Teilnahmebereitschaft bei Frauen größer ist.

in Duden 11 einzig die Variante *zwei Fliegen mit einer Klappe schlagen* angeführt ist. Da Online-Befragungen die Möglichkeit bieten, große Probandenzahlen inkl. ihrer arealen Zugehörigkeit zu erfassen, stellt diese Erhebungsmethode eine außerordentlich gute Möglichkeit dar, areale Aspekte der Phraseologie zu berücksichtigen, die relevante Aspekte für phraseologische Nachschlagewerke (die Bekanntheit oder Verwendung von Phrasemen im Sprachraum, die Verteilung von Varianten) darstellen. So ergeben die Daten der oben genannten Online-Befragung die in Abbildung 2 kartographisch dargestellte Verteilung der Varianten *zwei Fliegen mit einer Klappe*, *zwei Fliegen auf einen Schlag* und *zwei Fliegen auf einen Streich*.

Zu Abbildung 1 ist es wichtig zu erwähnen, dass im Fragebogen die Variante *zwei Fliegen auf einen Schlag* vorgegeben war. Die Probanden wurden gebeten, ihnen geläufige Varianten zu nennen, wenn die Vorgabe nicht bekannt war. Damit hatte die vorgegebene Variante *zwei Fliegen auf einen Schlag* einen deutlichen Selektionsvorteil und wurde wahrscheinlich deutlich häufiger genannt, als wenn die Probanden frei, d.h. ohne Vorgabe, hätten wählen können (die Ermittlung der arealen Verteilung von Varianten war nicht Ziel dieser Studie und die ungleiche Verteilung der Probanden auf die Verwaltungseinheiten erlaubt in diesem Fall keine abschließende Auswertung dieser Daten). Dennoch lässt sich aus der Karte grob die Tendenz ablesen, dass im nördlichen und mittleren Osten und im mittleren Westen des deutschen Sprachraums deutlich häufiger die Variante *zwei Fliegen mit einer Klappe* angegeben wurde als im Süden (in der Schweiz und Österreich). Die Variante *zwei Fliegen auf einen Streich* wurde viel seltener und verteilt über den Sprachraum genannt. Diese arealen Tendenzen ließen sich in einer gezielten Online-Befragung konkretisieren. Diese könnte entweder mit Bedeutungsparaphrasen und offenen oder halboffenen (diese drei Varianten und eine offene Kategorie als gleichberechtigte Alternativen) Antwortkategorien arbeiten. Denkbar ist auch eine Teiltextpräsentation[5], in der die Probanden auf der Grundlage einer Bedeutungsparaphrase des Phrasems den festen Kern der Einheit (*zwei Fliegen*) vorgegeben bekommen und gebeten werden, den variablen Teil zu ergänzen. Auf diese Weise können durch Online-Befragungen verlässliche phraseographische Zusatzinformationen generiert werden.

Auch wenn die Testgruppe der Probanden streng genommen nicht repräsentativ für die Grundgesamtheit der Sprechergemeinschaft ist, liefert sie doch einen recht zuverlässigen ersten Eindruck vom Bekanntheits- und/oder Verwendungsgrad von (auch weniger geläufigen) Phrasemen. Tendenziell ist davon auszugehen, dass die per Online-Befragung ermittelten Werte höher liegen als die tatsächlichen Werte, da einerseits das Interesse der Probanden (nur Interessierte nehmen teil) einen positiven Effekt auf die Beurteilung der eigenen Kenntnis bzw. Verwendung von Phrasemen hat und Probanden andererseits dazu neigen, ihre Kenntnis aufgrund der Verstehbarkeit eines Phrasems oder als Effekt sozialer Erwünschtheit eher höher einschätzen (zur möglichen Überschätzung von Probanden vgl. Juska-Bacher 2009: 140f). Um letzteres zu vermeiden, könnte auch hier mit einer Teiltextpräsentation gearbeitet werden, so dass Phraseographen aufgrund der Ergänzungen der Probanden per Fremdeinschätzung eine Bewertung als „(un)bekannt" bzw. „(nicht) verwendet" vornehmen können und den Bekanntheits- und Verwendungsgrad objektiv bestimmen

5 Vgl. Grzybek (1991) zur Bekanntheit von Sprichwörtern.

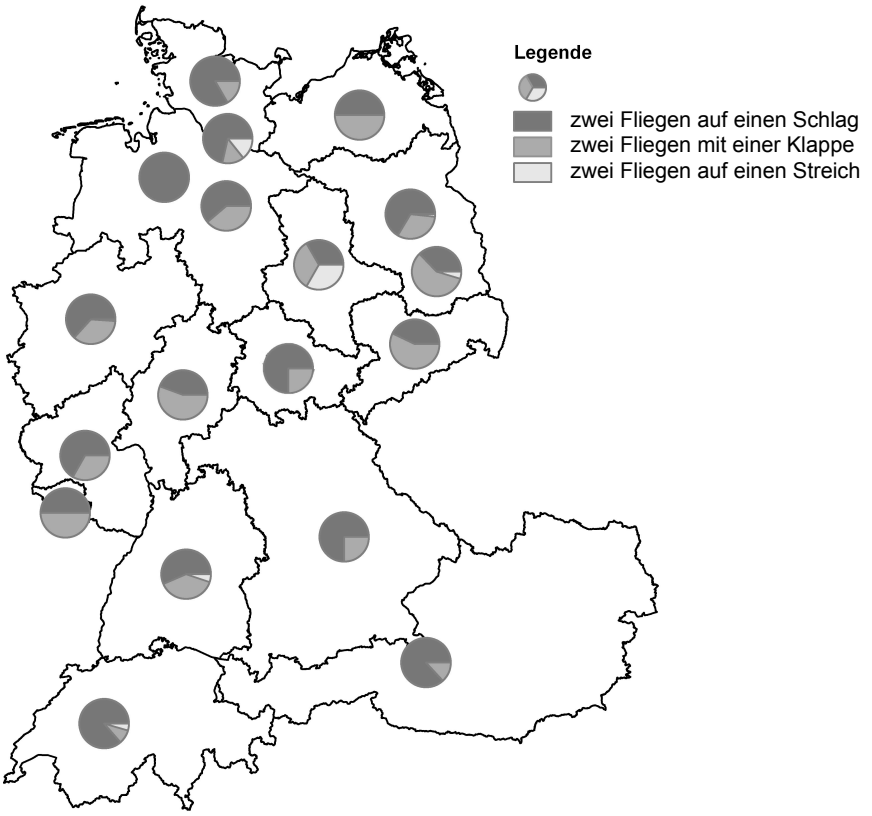

Abbildung 1: Areale Verteilung der Varianten *zwei Fliegen auf einen Schlag*, *zwei Fliegen mit einer Klappe* und *zwei Fliegen auf einen Streich* in Deutschland (nach Bundesländern), der Schweiz und Österreich. (Für die Erstellung der Karte geht unser Dank an Stefan Meier, Universität Basel.)

können. Diese sind besonders für die Sprachlerner-Phraseographie von Interesse, können aber auch als Aufnahmekriterium für aktuelle phraseologische Wörterbücher herangezogen werden. In der bereits mehrfach erwähnten Befragung (Juska-Bacher 2009) ergab sich für das Sprichwort *An den Federn erkennt man den Vogel.* ein Bekanntheitsgrad von über 30%, in Duden 11 jedoch fehlt es (hingegen ist das in derselben Befragung deutlich weniger bekannte Verbalphrasem *der Katze die Schelle umhängen* (15%) in Duden 11 enthalten).

Im Gegensatz zu den bereits genannten Bereichen relativ aufwändig ist auch im Rahmen einer Online-Befragung eine gezielte Erfassung von Konnotationen, deren Fehlen in Wörterbüchern oft angemahnt wird (Steffens 1989: 83) und die korpusanalytisch meist schwierig zu bestimmen sind. Dies ist beispielsweise durch Lückentexte, in denen unterschiedliche Kontexte vorgegeben sind, denkbar, doch auf Seiten der Probanden erfordern Angaben zu Konnotationen einen hohen Reflexionsgrad. In der oben genannten Online-Befragung wurden jedoch zu im Fragebogen vorgegebenen Bedeutungen immer wieder Hinweise wie „normalerweise negativ", „nicht unbedingt negativ gemeint" oder „es muss keine negative Bedeutung haben" angemerkt, obwohl Konnotationen nicht explizit erfragt wurden. Die Bündelung dieser oft breiter als im obigen Beispiel gestreuten Angaben und die Formulierung einer allgemein gültigen Aussage stellt unter Umständen erhebliche Ansprüche an die Phraseographen.

Ähnlich schwierig ist auch eine Abfrage von Restriktionen. Um Beschränkungen in Tempus, Numerus, Modus etc. ermitteln zu können, muss gezielt nach solchen gefragt werden. Auch hier gilt wie bei der korpusanalytischen Herangehensweise, dass sich nur das finden oder bestätigen lässt, wozu bereits bei der Untersuchungsanlage Anhaltspunkte vorhanden sind. Für eine Ermittlung pragmatischer Faktoren ist beispielsweise denkbar, dass Probanden gebeten werden, Phraseme bestimmten Verwendungssituationen zuzuordnen beziehungsweise solche auszuschließen.

3 Interaktive Komponenten im Online-Lexikon

Es zeigt sich, dass mit der Korpusrecherche und der Online-Befragung die derzeit in der phraseologischen Forschung populären empirischen Herangehensweisen im Vergleich zu den eingangs erwähnten drei „klassischen" Quellen (Autorenintuition, Beiträger, Vorgängerwerke) zwar bereits einen deutlichen Mehrwert bei der genauen Bestimmung und Erläuterung einzelner Phraseme bieten, aber immer noch zahlreiche Leerstellen offen lassen. Einerseits hat das methodologische Ursachen (bspw. die begrenzten Möglichkeiten des Textretrievals oder die Abhängigkeit der Antwort von der Formulierung einer Frage), andererseits spielen auch die spezifischen Eigenschaften von Phrasemen, wie die erhöhte Expressivität, die große Variabilität, die verknüpfte Kulturgeschichte etc. eine Rolle. Insgesamt kostet daher selbst unter oder gerade durch die Zuhilfenahme von Korpora und Online-Befragungen die phraseographische Aufarbeitung eines Phrasems um ein Vielfaches mehr Zeit und Aufwand als das bei einem Einwortlexem der Fall ist (Fellbaum et al. 2004: 43, Schmidlin 2001: 108, Scholze-Stubenrecht 1988: 286). Schon allein deshalb ist es legitim zu überlegen, welche Wege der Datenerhebung sich noch beschreiten lassen. Wege, die vor

allem auszeichnet, dass sie bei der phraseographischen Datengewinnung möglichst wenig zusätzlichen Aufwand bereiten.

Stellvertretend für viele – teilweise allerdings noch sehr spekulative Methoden – präsentieren wir hier zwei Möglichkeiten, welche uns durchaus praktikabel erscheinen. Ausgangspunkt der Überlegungen ist ein – noch zu erstellendes – phraseologisches Online-Lexikon, in welchem quasi durch die Benutzung weitere Datenbestände generiert werden[6].

In einem Online-Lexikon ist es beispielsweise leicht realisierbar – und das Ergebnis dürfte exakter sein als jede Ableitung aus einer Befragung zum Bekanntheitsgrad beziehungsweise zur Verwendungshäufigkeit –, herauszufinden, welche Phraseme wirklich nachgeschlagen werden. Hierfür müssen lediglich die Suchanfragen in Kombination mit den anschließend ausgewählten Artikeln ausgewertet werden. Rein von den Suchausdrücken lässt sich in vielen Fällen nicht auf das vom Benutzer gesuchte Phrasem schließen, dessen Auswahl eines Artikels hingegen ist eine eindeutige Information. Nebenbei gäbe dieses Verfahren Aufschluss darüber, welche Suchstrategien die Benutzer anwenden, also ob sie beispielsweise nach Grundformen, nach flektierten Formen oder nach einzelnen Elementen suchen. Diese Erkenntnisse wiederum sind hilfreich für die Verbesserung des Retrievals und der Ausgabestruktur des Online-Lexikons. Aufgrund von Suchstatistiken ließen sich dann Prioritäten für die Erarbeitung weiterer Artikel erstellen. Damit wird das Nachschlagewerk stärker auf die Nutzungssituation ausgerichtet (siehe unten) und unabhängig(er) von den mutmaßlichen Rückschlüssen von Verwendungshäufigkeit, Repräsentativität in Texten etc. Wir gehen davon aus, dass jene Phraseme, die wirklich in Wörterbüchern nachgeschlagen werden, nicht die häufigsten und nicht die seltensten einer Sprache sind. Erstere dürften den Sprechern zu bekannt sein, als dass sie sie nachschlagen, letztere geben zu wenig Anlass, nachgeschlagen zu werden, da sie zu selten in der Sprache vorkommen.

Insgesamt ist der Wörterbuchbenutzer für den Wörterbuchautoren bisher eine theoretische Institution, die Artikel werden nach gemutmaßten Annahmen erstellt. Die Lexikographie – und noch weniger die Phraseographie – hat sich bislang wenig um die Benutzer gekümmert (Hallsteinsdóttir 2006: 96; Scholze-Stubenrecht 2004: 348), obwohl diese nicht nur als Zielgruppe näher zu spezifizieren sind, sondern auch in einigen Fällen sinnvoll zur Verbesserung phraseographischer Artikel beitragen können. Im Folgenden möchten wir denkbare Beispiele vorstellen, wie der Benutzer eines Online-Lexikons zur Verbesserung der Artikel beitragen kann – beispielsweise bezüglich der Belege und der Belegauswahl. Der Phraseograph wählt in der Regel einen sprechenden, erklärenden Beleg aus und achtet dabei nach Möglichkeit auf Eindeutigkeit und Kürze. Gibt es sprachspielerische oder besonders bekannte Textpassagen als Alternative, kommt er schon in einen Zwiespalt. Welcher Beleg dem Benutzer am hilfreichsten ist, kann er bestenfalls abschätzen. In einer vom Benutzer durch einfaches Anklicken abgegebenen Bewertung der Belege – beispielsweise durch eine einfache Skala von eins bis vier – könnte somit eine Wertigkeit der erfassten Belege erstellt werden. Nach einer gewissen Schwelle an abgegebenen Bewertungen pro Belegsammlung zu einem Phrasem, wird dann automatisch der als am nützlichsten bewertete Beleg nach oben sortiert. In gleicher Weise ließen sich Bedeutungsangaben bewerten, wobei hier weni-

6 Die Nutzung von Log-Files als phraseographische Quelle über das Nutzungsverhalten der Wörterbuchnutzer wie es Bergenholtz/Johnsen (2007) vorstellen, ist ein solcher Ansatz.

ger die per se sinnlose Frage zu stellen ist, welche Bedeutungsangabe die nützlichere ist, sondern vielmehr abgefragt werden könnte, welche Bedeutungen ein Benutzer für aktuell hält. Hinsichtlich der Varianten von Phrasemen könnte man den Benutzern ermöglichen, die Varianten, die in ihrer Heimatvarietät gesprochen werden, zu erfassen und dazu auch die entsprechende Postleitzahl des Verbreitungsgebietes des Phrasems einzutragen. Bei ausreichender Datenlage ließen sich diese Angaben dann auf Karten übertragen und es würde ein Bild der geographischen Verteilung einzelner Varianten eines Phrasems entstehen. Entscheidend bei allen diesen Ergänzungen zur phraseographischen Arbeit ist, dass sie weitestgehend automatisch ablaufen und erst ab einer gewissen erfassten Menge an Daten veröffentlicht werden (auch diese Freischaltung lässt sich automatisieren). Strukturell sollten solche Angaben von Benutzern klar nachvollziehbar von den anderen Angaben getrennt werden, was – sofern man es von Anfang an bedenkt – technisch keine Probleme bereitet. Ein Vorteil des Einbezugs von Benutzern ist auf jeden Fall auch, dass die Erfassung von Daten (Belegen, Varianten, geographischen Angaben etc.) extern übernommen wird und dadurch den Phraseographen nicht in Anspruch nimmt. Somit können einige Daten, die auch auf anderen Wegen zu gewinnen sind (Online-Befragung, Korpusrecherche), zeitökonomischer erfasst werden.

Die Formulierung von Bedeutungsparaphrasen und Konnotationen halten wir für zu komplex, um sie in irgendeiner Weise von Benutzern vornehmen zu lassen. Hier ist weiterhin die Sach- und Fachkenntnis eines ausgebildeten Phraseographen gefragt, auch um eine gewisse Einheitlichkeit in der Formulierung der Bedeutungsparaphrasen zu gewährleisten. Im Bereich der Motivierung und Remotivierung hingegen – beides spielt schließlich bei der Bedeutungszuschreibung zu einem Phrasem eine große Rolle –, kann die Meinung des einzelnen Benutzers durchaus von Interesse sein. Denn gerade Remotivierungsprozesse bilden einen Nährboden für den phraseologischen Bedeutungswandel und sollten in einem Nachschlagewerk nicht unterbewertet werden. Zudem ist zu fragen, ob eine Motivierung, welche eine Mehrzahl von Benutzern für „richtig" erachtet, nicht auch in einem Nachschlagewerk thematisiert werden sollte. Diese Diskussion würde hier zu weit führen, Anhaltspunkte dafür ließen sich aber gewinnen, fragte man solche (Re-)Motivierungen der Phraseme von den Benutzern eines Online-Lexikons ab. Als Beispiel sei auf einige rege mit phraseologisch-etymologischen Fragen konsultierte Internetplattformen (z.B. *wer-weiss-was.de* oder auch *wikipedia.de*) verwiesen, auf welchen zahlreiche Diskussionen zum Ursprung und zur Motivierung verschiedenster Phraseme nachzulesen sind. Entscheidend hierbei ist, dass – im Gegensatz zur Befragung – die Diskussion völlig ungesteuert abläuft. Inwiefern sich solche Diskussionen provozieren lassen, sei dahingestellt, doch es spräche nichts dagegen, in einem Online-Lexikon zu den entsprechenden Artikeln auch eine Diskussionsplattform einzurichten. Diese Diskussionen nach ein paar Jahren zumindest teilweise auszuwerten, könnte durchaus neue Erkenntnisse und Anregungen für die weitere phraseographische Bearbeitung eines Artikels liefern. Letztendlich würde eine solche Diskussionsplattform nicht nur interessierten Laien eine Möglichkeit zum Austausch bieten, sondern auch Fachpublikum könnte auf diesem Weg kritisch-konstruktiv die entstandenen Artikel kommentieren.

4 Zusammenfassung und Ausblick

Ein Blick auf die Anforderungen, die gegenwärtig an die Phraseographie gestellt werden, hat gezeigt, dass die traditionell genutzten Quellen von der Autorenintention bis zur Korpusrecherche gewisse Informationslücken, beispielsweise in Bezug auf aktuelle Bedeutungen, areale Aspekte, pragmatische Angaben, offen lassen. Mit Hilfe von Korpusanalysen können einige Aspekte (z.B. aktuelle Bedeutungen) zumindest für häufig verwendete Phraseme ergänzt werden, für andere Aspekte (z.B. areale Verbreitung) liefert diese Methode wenig oder kein Material. Vor diesem Hintergrund haben wir den Versuch unternommen, das Potential der digitalen Datenerhebungsmethoden für die Ergänzung phraseographisch relevanter Informationen auszuloten.

Wir haben zunächst aufgezeigt, welche Möglichkeiten die phraseographisch bislang nicht systematisch nutzbar gemachte Online-Befragung bietet. Sie ist im Vergleich zu postalisch versandten Fragebögen sehr viel zeit- und kostenökonomischer und es lässt sich mit vergleichbarem Aufwand eine deutlich größere Zahl von Probanden rekrutieren. So können zusätzliche Angaben im Bereich von Bedeutungsbestimmungen, Varianten, arealen Aspekten sowie Bekanntheits- und Verwendungsgrad gewonnen werden. Diese Methode zeichnet sich dadurch aus, dass die Daten aktuell (besonders im Vergleich zur Wiederverwertung von Angaben aus älteren Wörterbüchern), authentisch und objektiv sind (besonders im Vergleich zur Expertenmeinung) und auch für seltene Phraseme erhoben werden können (im Vergleich zur Korpusanalyse). Ein Problem dieser Methode ist und bleibt allerdings einerseits die Begrenzung der Kapazität der Teilnehmenden, die für Anschlussuntersuchungen motiviert oder neu rekrutiert werden müssen. Andererseits besteht auch bei der relativ unpersönlichen Online-Befragung immer noch eine Beeinflussung durch Fragenformulierung, Antwortvorgaben (Echoformen) oder einfach durch die Erhebungssituation (soziale Erwünschtheit), die eine große Sensibilität bei der Gestaltung der Fragen wie auch bei der Deutung der Ergebnisse erfordert.

Aufgrund der aufgezeigten Möglichkeiten, bestehende Defizite in der Datenerhebung für die Phraseographie weiter zu minimieren, scheint es uns erstrebenswert, empirisch-digitale Methoden für die Wörterbucherstellung stärker als bisher einzusetzen, um dem Phraseographen neben Wörterbuchmaterial, Primärquellen und Expertenmeinung eine breite Grundlage von Sprachbenutzerdaten zur Verfügung zu stellen.

Als zweite Methode der Online-Datenerhebung haben wir den interaktiven Ansatz eines Online-Lexikons als Sammelstelle für indirekte phraseologische Informationen vorgestellt. Mehrfach wurde bereits darauf hingewiesen, dass die lexikographische Aufarbeitung von Phrasemen um ein Vielfaches umfassender und aufwändiger ist als bei Einwortlexemen. Da sich bei der Erläuterung der Phraseme kaum an Fachpersonal sparen lässt, scheint uns die Datenerhebung und deren wo möglich automatische Auswertung jener Punkt zu sein, an welchem sich der phraseographische Prozess deutlich optimieren lässt.

Hier bieten sich für ein entsprechend konzipiertes Online-Nachschlagewerk zahlreiche Perspektiven, die den Benutzer nicht nur als Informationskonsumenten betrachten, sondern ihn auch als Datenlieferanten einbinden. Wie solche Lösungen im Detail aussehen können, und ob sie technisch umsetzbar sowie wissenschaftlich ausreichend fundiert sind,

muss die Praxis zeigen. Entscheidend ist, dass stets nachvollziehbar bleibt, welche Daten von einem ausgebildeten Phraseographen stammen und welche von einer undefinierbaren Menge an Benutzern zusammengetragen wurden. Sofern dieser Aspekt gewährleistet ist, kann je nach Qualität der Daten jederzeit für oder gegen eine wissenschaftliche Nutzung entschieden werden. Der Aufwand – und das sollte allen alternativen Methoden der Datenerhebung eigen sein – beschränkt sich anfänglich auf die technische Realisierung entsprechender Module eines Online-Nachschlagewerks zur Einbindung der Benutzer. Erst wenn man sich nach einer qualitativen Kontrolle der Beiträge entscheidet, die Daten weiterzuverwerten, zu interpretieren oder anderweitig einzubinden, entsteht ein Mehraufwand. Deutlich hervorheben möchten wir zum Schluss noch einmal, dass aufgrund der erwähnten Ausmaße phraseologischer Wörterbuchartikel und auch aufgrund mancher – vorübergehend – nicht lösbaren Frage hinsichtlich der Geschichte oder Motivierung von Phrasemen nur ein digitales (Online-)Nachschlagewerk diesen sprachlichen Einheiten zwischen Wort und Satz gerecht werden kann. Und wahrscheinlich ist auch dieser Gedanke noch zu kurz formuliert, denn der von der Wörterbuchtradition geprägte Begriff „Werk" impliziert etwas Fertiges, Abgeschlossenes. Aufgrund ihrer Vielschichtigkeit und ihrer großen Affinität sich strukturell und semantisch beständig zu wandeln, können Phraseme jedoch faktisch nie fertig und abgeschlossen bearbeitet werden. Sie benötigen also nicht nur ein Online-Nachschlagewerk, sondern ein solches, das technisch die Möglichkeit bietet, bei veränderter Datenlage Ergänzungen und Korrekturen zu erfassen, und das trotzdem eine wissenschaftliche Verlässlichkeit und Zitierfähigkeit gewährleistet. Einen geeigneteren Indikator für Veränderungen im Sprachusus als die Sprecher einer Sprache und damit die Benutzer ihrer Nachschlagewerke können wir uns nicht vorstellen.

Literaturverzeichnis

Adelung, J. C. (1793–1801): *Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart. Mit beständiger Vergleichung der übrigen Mundarten, besonders aber der Oberdeutschen*. 4 Bde. Leipzig. Nachdruck hrsg. v. Henne, H., Hildesheim 1970.

Agricola, J. (1534): *Sybenhundert und fünfftzig Teütscher Sprichwörter verneüwert und gebessert*. Hagenaw. Fotomechanischer Nachdruck hrsg. v. Mieder, W., Hildesheim/Zürich 1970.

Bandilla, W. (1999): „WWW-Umfragen – Eine alternative Datenerhebungstechnik für die empirische Sozialforschung?" In: Batinic et al. (Hrsg.) (1999); 9–19.

Batinic, B./Werner, A./Gräf, L./Bandilla, W. (Hrsg.) (1999): *Online Research. Methoden, Anwendungen und Ergebnisse*. Göttingen/Bern/Toronto/Seattle.

Bergenholtz, H./Johnsen, M. (2007): „Log Files Can and Should Be Prepared for a Functionalistic Approach", in: *Lexikos 17*; 1–21.

Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N. R. (Hrsg.) (2007): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*. Berlin/New York.

Burger, H./Häcki Buhofer, A. (Hrsg.) (2006): *Phraseology in Motion. Methoden und Kritik*. Baltmannsweiler.

Burger, H./Häcki Buhofer, A./Gréciano, G. (Hrsg.) (2003): *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie*. Baltmannsweiler.

Campe, J. H. (1807): *Wörterbuch der Deutschen Sprache. 5 Bde.* Braunschweig. Nachdruck hrsg. v. Henne, H., Hildesheim 1969.

Christen, H./Germann S./Montefiori, N./Haas, W./Ruef, H. (Hrsg.) (im Druck): *Dialektologie: Wege in die Zukunft. Beiträge zur 16. Arbeitstagung zur alemannischen Dialektologie in Freiburg im Üchtland. 7. – 10. September 2008.*

Colson, J.-P. (2003): „Corpus Linguistics and Phraseological Statistics: a few Hypotheses and Examples." In: Burger et al. (Hrsg.) (2003); 47–59.

Dräger, M. (eingereicht-a): „Auf der Suche nach historischen Phraseologismen – oder: Wörterbücher als Korpora", in: *Linguistik online*.

Dräger, M. (eingereicht-b): „Phraseologische Nachschlagewerke im Fokus". In: Korhonen, J. (Hrsg.) (voraussichtlich 2010): *Phraseologie. Global – areal – regional.*

Duden Bd. 11. Redewendungen. Wörterbuch der deutschen Idiomatik (2002). Hrsg. von der Dudenredaktion. Mannheim/Leipzig/Wien/Zürich. 2. Auflage.

Ďurčo, P. (2001): „Bekanntheit, Häufigkeit und lexikographische Erfassung von Sprichwörtern. Zu parömiologischen Minima für DaF." In: Häcki Buhofer et al. (Hrsg.) (2001); 99–106.

Elspaß, S./Möller, R. (2006): „Internet-Exploration: Zu den Chancen, die eine Online-Erhebung regional gefärbter Alltagssprache bietet", in: *Osnabrücker Beiträge zur Sprachtheorie 71*; 141–156.

Fellbaum, C./Kramer, U./Neumann, G. (2006) : „Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome". In: Burger/Häcki Buhofer (Hrsg.) (2006); 43–56.

Filatkina, N. (2007): „Formelhafte Sprache und Traditionen des Formulierens (HiFoS). Vorstellung eines Projekts zur historischen formelhaften Sprache", in: *Sprachwissenschaft 32 (2)*; 217–242.

Frisch, J. L. (1741): *Teutsch-Lateinisches Wörter-Buch. Darinnen nicht nur die ursprünglichen, nebst denen davon hergeleiteten und zusammengesetzten allgemein gebräuchlichen Wörter; Sondern auch die bey den meisten Künsten und Handwerken, bey Berg- und Saltzwerken, Fischereyen, Jagd-, Forst- und Hauß-Wesen, u.a.m. gewöhnliche Teutsche Benennungen befindlich, Vor allen, Was noch in keinem Wörter-Buch geschehen, Denen Einheimischen und Ausländern, so die in den mittlern Zeiten geschriebenen Historien, Chroniken, Übersetzungen, Reimen u.d.g. mit ihren veralteten Wörtern und Ausdrückungen verstehen wollen, möglichst zu dienen; Mit überall beygesetzter nöthigen Anführung der Stellen, wo dergleichen in den Büchern zu finden, Samt angehängter Theils versicherten, theils muthmaßlichen Etymologie und critischen Anmerkungen; Mit allem Fleiß viel Jahr über zusammengetragen Und jetzt den Gelehrten zur beliebigen Vermehrung und Verbesserung überlassen. Nebst einem Register der lateinischen Wörter.* 2 Bde. Berlin. Nachdruck der 2 Bände in einem hrsg. v. Powitz, G., Hildesheim 2007.

Grimm, J./Grimm, W. (1854–1960): *Deutsches Wörterbuch*. 16 Bde. Leipzig.

Grzybek, P. (1991): „Sinkendes Kulturgut? Eine empirische Pilotstudie zur Bekanntheit deutscher Sprichwörter", in: *Wirkendes Wort 2/91*; 239–264.

Häcki Buhofer, A./Burger, H./Gautier, L. (Hrsg.) (2001): *Phraseologiae Amor*. Baltmannsweiler.

Hallsteinsdóttir, E. (2006): „Phraseographie", in: *Hermes. Journal of language and communication studies 36*; 91–128.

Hallsteinsdóttir, E./Šajánková, M./Quasthoff, U. (2006): „Phraseologisches Optimum für Deutsch als Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen", in: *Linguistik-online* 27, 2/06; 119–138. *www.linguistik-online.de/27_06/hallsteinsdottir_et_al.pdf*.

Harras, G. (Hrsg.) (1988): *Das Wörterbuch. Artikel und Verweisstrukturen.* Düsseldorf.

Hartmann, D./Wirrer, J. (Hrsg.) (2002): *Wer A sägt, muss auch B sägen. Beiträge zur Phraseologie und Sprichwortforschung aus dem Westfälischen Arbeitskreis.* Baltmannsweiler.

Haß-Zumkehr, U. (2001): *Deutsche Wörterbücher. Brennpunkt von Sprach- und Kulturgeschichte.* Berlin/New York.
Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (Hrsg.) (1989): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie.* Berlin/New York.
Heid, U. (2007): „Computational linguistic aspects of phraseology II". In: Burger et al. (Hrsg.) (2007); 1036–1044.
Juska-Bacher, B. (2009): *Empirisch-kontrastive Phraseologie. Am Beispiel der Bekanntheit der Niederländischen Sprichwörter im Niederländischen, Deutschen und Schwedischen.* Baltmannsweiler.
Juska-Bacher, B. (im Druck): „SDS-Exploratoren und Online-Befragung – Lässt sich im Methodenmix ein Wandel in der Schweizer Dialektlandschaft nachweisen?". In: Christen et al. (Hrsg.) (im Druck).
Kempcke, G. (1986): „Theoretische und praktische Probleme der Phraseologiedarstellung in einem synchronischen einsprachigen Bedeutungswörterbuch". In: Korhonen (Hrsg.) (1987); 155–164.
Korhonen, J. (Hrsg.) (1987): *Beiträge zur allgemeinen und germanistischen Phraseologieforschung. Internationales Symposium in Oulu 13.-15. Juni 1986.* Oulu.
Kühn, P. (1994): „Pragmatische Phraseologie: Konsequenzen für die Phraseographie und Phraseodidaktik". In: Sandig (Hrsg.) (1994); 411–428.
Mokienko, V. M. (2002): „Prinzipien einer historisch-etymologischen Analyse der Phraseologie". In: Hartmann/Wirrer (Hrsg.) (2002); 231–254.
Piirainen, E. (2001): „Phraseologie und Arealität", in: *Deutsch als Fremdsprache* 4; 240–243.
Piirainen, E. (Hrsg.) (2002): *Phraseologie in Raum und Zeit. Akten der 10. Tagung des Westfälischen Arbeitskreises „Phraseologie/Parömiologie" (Münster 2001).* Baltmannsweiler.
Pilz, K. D. (2002): „Vorschläge für ein Phraseolexikon der deutschen Sprache. Oder: Vorschläge für ein Lexikon der deutschen Phraseme/Phraseologismen". In: Hartmann/Wirrer (Hrsg.) (2002); 299–311.
Röhrich, L. (2002): *Das große Lexikon der sprichwörtlichen Redensarten*. 3 Bde. Darmstadt/Basel.
Rothkegel, A. (2007): „Computerlinguistische Aspekte der Phraseme I". In: Burger et al. (Hrsg.) (2007); 1027–1035.
Sandig, B. (Hrsg.) (1994): *Europhras 92. Tendenzen der Phraseologieforschung.* Bochum.
Schemann, H. (1989): „Das phraseologische Wörterbuch". In: Hausmann et al. (Hrsg.) (1989); 1019–1032.
Schmidlin, R. (2001): „Lexikographische Probleme bei phraseologischen Varianten". In: Piirainen (Hrsg.) (2002); 377–391.
Scholze-Stubenrecht, W. (1988): „Phraseologismen im Wörterbuch". In: Harras (Hrsg.) (1988); 284–302.
Scholze-Stubenrecht, W. (2004): „Duden 11 – Lexikographisches Konzept und lexikographische Praxis". In: Steyer (Hrsg.) (2004); 348–359.
Steffens, D. (1989): „Untersuchung zur Phraseologie der deutschen Gegenwartssprache unter lexikographischm Aspekt", in: *Beiträge zur Erforschung der deutschen Sprache* 9; 79–93.
Steyer, K. (2003): „Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches ‚berechnen'?". In: Burger et al. (Hrsg.) (2003); 33–46.
Steyer, K. (Hrsg.) (2004): *Wortverbindungen - mehr oder weniger fest.* Berlin.
Wander, K. F. W. (1867–1880): *Deutsches Sprichwörter-Lexikon. Ein Hausschatz für das deutsche Volk.* 5. Bde, Leipzig. Unveränderter fotomechanischer Nachdruck, Darmstadt 1964.

Online-Datenerhebungen im Dienste der Phraseographie

Marcel Dräger
Universität Basel
Deutsches Seminar
Nadelberg 4
4051 Basel
Schweiz
marcel.draeger@unibas.ch

Britta Juska-Bacher
Fachhochschule Nordwestschweiz, Zentrum Lesen
Kasernenstrasse 20
5000 Aarau
Schweiz

Universität Basel
Deutsches Seminar
Nadelberg 4
4051 Basel
Schweiz
britta.juska-bacher@unibas.ch

Swedish Medical Collocations: A Lexicographic Approach

Maria Toporowska Gronostaj/Emma Sköldberg

Der vorliegende Artikel untersucht schwedische Kollokationen aus dem Bereich der Medizin. Die Kollokationen werden sowohl aus onomasiologischer als auch aus semasiologischer Perspektive beschrieben. Im Zentrum der Analyse stehen lexikalisch-grammatische und lexikalisch-semantische Eigenschaften von Substantiven, Verben und Kollokationen im Sinne in sich geschlossener Einheiten. Behandelt werden u. a. die Präferenzen von Wörtern und Sätzen hinsichtlich des Numerus, der Valenz sowie von Selektionsrestriktionen. Als Ausgangspunkt für die Beschreibung dienen Kollokationen aus der medizinisch-lexikalischen Datensammlung MedLex, auf deren Basis ein elektronisches Lexikon entstehen soll. Es werden zwei Typen von Mehrwortlemmata diskutiert: Der erste Typ besteht aus Kollokationen an sich (z. B. *gå in i en depression* ‚depressiv werden'); der zweite Typ umfasst Kollokationen mit bedeutungsverwandten Konstituenten (z. B. steht das Lemma *ta MEDICIN* ‚Medizin (ein-) nehmen' für *ta tabletter* ‚Tabletten (ein-)nehmen', *ta penicillin* ‚Penicillin (ein-) nehmen' etc.) Die Analyse versteht sich als ein Beitrag zur systematischen Beschreibung von Kollokationen und soll den künftigen Anwendern des Lexikons nützen.

1 Introduction

Work on building a workbench, incorporating a medical text corpus, tools for text analysis and a lexical database for Swedish, is in progress at the Department of Swedish, University of Gothenburg. The workbench, and in particular the medical lexical database, is intended to support lexicographers in compiling a multifunctional electronic lexicon covering medical vocabulary (Kokkinakis/Toporowska Gronostaj 2008). While work on a general module of the database (MedLex) has progressed, work on its collocation module is still in its initial phase and the ongoing study of Swedish medical collocations paves the way for its compilation. The primary aim of the compilation task undertaken is to focus on the important role collocations play in communication between medical professionals and laypersons. This is one of the reasons why medical collocations deserve a more exhaustive and systematic presentation in general-purpose as well as medical dictionaries. Furthermore, there is no doubt that broad access to information on collocations plays a major role in many areas of language technology applications. It can make structural and lexical annotation

of corpora more reliable. It can also provide support for systems dealing with machine learning, translation and text understanding.

The main purpose of this paper is to share some insights into the analysis of medical collocations including a noun and a verb, or in short noun-verb collocations. Collocations are understood here as recurrent combinations of two or more words with an internal semantic and syntactic structure, displaying lexical binding. A typical collocation is to a considerable extent transparent and therefore fairly easy to decode, but more difficult to encode (Malmgren 2003, Heid 2004; cf. Siepmann 2005, 2006). The set of examined collocations comprises verb phrase collocations, e. g. *ta blodprov* 'take a blood test', and sentential collocations, e. g. *febern stiger* 'the body temperature rises'.

Besides the theoretical goal of the study, there is also a practical one. It involves turning the corpus-derived data into a lexicographic resource. The planned collocational module is meant to be descriptive, for both reception and production, having native and non-native laypersons and health care personnel as its potential users. To meet this challenge, a broad spectrum of both semasiological and onomasiological information on medical terms and their collocations needs to be provided, as well as a wide choice of search options (Sköldberg/Toporowska Gronostaj 2008a, 2008b).

The point of departure for our empirical study has been approximately 350 verbs from the medical domain, which are entries in MedLex. After a closer examination, supported by a statistical program based on mutual information as well as on our intuitions, we have reduced the number of verbs to approximately 70, of a more clearly collocational kind. These verbs have been studied in both a general text corpus, Språkbanken (The Swedish Language Bank), containing approximately 100 million Swedish words, and in a medical text corpus, the MedLex Corpus, of 25 million Swedish words. In the next step, the list of collocations extracted has been subjected to a critical evaluation based on the characteristics of collocations outlined in sections 2 and 3. Finally, there are fully 30 different collocations referred to in this article.

The remainder of this paper is structured in the following way. Section 2 provides a short introduction to an onomasiological perspective on medical collocations. Section 3 presents a semasiological perspective on the collocations studied. Some proposals concerning lexicographic modes of representation of noun-verb collocations are dealt with in section 4. Conclusions are given in the last section.

2 Noun-Verb Collocations from an Onomasiological Perspective

We assume that a knowledge-oriented approach to lexicons, in line with the one proposed by Martin (2006), contributes to a more exhaustive and systematic detection and description of collocations. For this reason, we have viewed Swedish medical collocations from an onomasiological perspective, starting from three medical subdomains, namely, diseases, diagnostics and treatments. Viewing the data thematically has helped us to detect semantic parameters and their patterns for the respective subdomains.

Parameters	***depression* 'depression'**	***förkylning* 'cold'**
Falling ill	*få en depression* 'get a depression' *drabbas av (en) depression* 'be stricken with depression' *gå in i en depression* 'go into (a) depression'	*få en förkylning* 'get/catch a cold' *drabbas av en förkylning* 'be stricken with a cold' *dra på sig en förkylning* 'contract a cold' *ådra sig en förkylning* 'be stricken with a cold'
Status	*ha (en) depression* 'have a depression' *lida av (en) depression* 'suffer from a depression'	*ha en förkylning* 'have a cold'
Cure	*bota depression* 'cure depression' *behandla depression* 'treat depression'	*bota en förkylning* 'cure a cold' *kurera en förkylning* 'nurse a cold'
Recovery	*komma ur en depression* 'come out of a depression' *ta sig ur en depression* 'get over a depression'	

Table 1: Parameter-based classification of noun-verb collocations with the nouns *depression* 'depression' and *förkylning* 'cold'

In order to clarify the methodology underlying the onomasiological analysis, table 1 lists parameters characteristic of the subdomain diseases detected among the noun-verb collocations. The parameters Falling ill, Status, Cure and Recovery are exemplified with collocations referring to the diseases *depression* 'depression' and *förkylning* 'cold'.

From the examples in table 1, it follows that there are verbs which are typical of only one of the diseases. For instance, the phrasal verb *gå in* 'go into' occurs almost exclusively with the noun *depression*. Other verbs are more recurrent and somewhat less restrictive as to choice of noun, e. g. *drabbas* 'be stricken with, contract' can be followed by most of the nouns referring to a disease. The empty spaces in the table indicate lack of collocations and reflect a certain asymmetry on the conceptual level in the system. The asymmetry may result either from a factual incompatibility between a parameter and a disease type, or from a preference for using free combinations rather than collocations. It is worth noting that this asymmetric distribution of collocations may give rise to difficulties in encoding situations. (The onomasiological perspective is further elaborated in Sköldberg/Toporowska Gronostaj 2008a, 2008b.)

From what has been said here, it follows that the onomasiological perspective supports a certain systematisation of collocations, at the same time uncovering their idiosyncratic

lexical preferences. But it also provides a useful point of departure for further semasiological description of collocations.

3 Noun-Verb Collocations from a Semasiological Perspective

The description and analysis of noun-verb collocations involves at least three objects, namely the noun, the verb and the collocation as a whole, including the relations between its components (cf. Heid 2004). In this section, these three objects are examined with respect to their lexico-grammatical and lexico-semantic features. The lexico-grammatical features account for the behaviour of words, in line with the notion of colligation, that is "the grammatical associations that a word forms with its environment or the grammatical patterning in which it participates" (Hoey/Brook O'Donnell 2008: 294). The lexico-semantic features involve such phenomena as selection restrictions, semantic roles and semantically related concepts.

3.1 Nouns in Medical Noun-Verb Collocations

We start with an overview of the lexico-grammatical features of nouns and their manifestations concerning number, determination and modification potential. After that, we focus on their semantic features. A list of features with examples is provided in table 2.

For nouns in the noun-verb collocations, the application of general rules of grammar is much more constrained than for corresponding constructions in free combinations. Restrictions concern the use of number, determination and modification and there is a strain of idiosyncrasy in their manifestations. In the collocation *gå in i en depression* 'go into (a) depression', the only possible form of the noun *depression* is that of singular number and indefinite form. In contrast, the noun *komplikation* 'complication' usually occurs in an indefinite, plural form and in subject position in the collocation *komplikationer tillstöter* 'complications set in'.

As far as determination is concerned, the examples in table 2 show the use of nouns with an indefinite article (*gå in i en depression* 'go into (a) depression'), without an article (*ha feber* 'have a fever') and in a definite form (*febern stiger* 'the body temperature goes up'). The collected data on the medical collocations reveal certain correlations between the natural order of events and the form of determination chosen by the nouns. In analogy to general language, when a situation is introduced or comes up, the noun occurs in the indefinite or generic form (*få feber* 'be feverish') and in subsequent situations, the noun takes the definite form (*febern stiger* 'fever goes up').

As regards modification, two main types can be distinguished, namely, a syntactic and a morphosyntactic one. Syntactic modification appears in attributive and appositive constructions. There are relatively many nouns in the collocations studied that can take a preposed adjectival modifier. A characteristic feature of these modifiers is that they often build a collocation on its own with the noun, e. g. *hög feber* 'high temperature' (Heid 2004: 732). Some of the nouns can be complemented by an appositive construction, usually a

Analysis level	Manifestation	Example
Lexico-grammatical features	Number	*gå in i en depression* 'go into (a) depression' *komplikationer tillstöter* 'complications set in'
	Determination	*gå in i en depression* 'go into (a) depression' *ha feber* 'have a fever' *febern stiger* 'the fever rises'
	Modification potential	*ha hög feber* 'have a high fever' *ställa diagnosen cancer* 'make the diagnosis cancer' *ta prov* vs *ta blodprov* 'take specimen' vs 'take blood test'
Lexico-semantic features	Selection restriction	*djup depression* 'deep depression' *långdragen depression* 'protracted depression' *årstidsbunden depression* 'seasonal depression' *lida av sjukdomen artros* 'suffer from the disease arthrosis' *ställa diagnosen ADHD* 'make the diagnosis ADHD'
	Lexico-semantic alternation	*febern/temperaturen sjunker* 'the fever/the body temperature decreases' *göra ultraljud/ultraljuds-undersökning* 'lit. make an ultrasound examination'

Table 2: Relevant features in the analysis of nouns in noun-verb collocations

noun, which specifies the semantic content of the noun in the collocation (cf. *ställa diagnos* 'make a diagnosis', *ställa diagnosen cancer* 'make the diagnosis of cancer'). Since Swedish is a compounding language, morphosyntactic modifications are frequent, e. g. *ta prov* 'take a specimen' and ta *blodprov* 'take a blood test'. In this particular case, the noun *blod* in the compound *blodprov* specifies the type of the test.

Let us turn to the lexico-semantic features of the nouns, in particular those which can undergo attributive and appositive modification. Generalizations about the preferred types of modifiers and appositives can be captured by means of selection restrictions. For example, for nouns referring to diseases, the selection restrictions concern modifiers describing Stage, Course and Type. Thus, for the noun *depression*, these can be illustrated by *djup* 'deep', *långdragen* 'protracted' and *årstidsbunden* 'seasonal'. As to appositives, they are usually restrictive, as illustrated in the following examples: *lida av sjukdomen artros* 'suffer from the disease of arthrosis' and *ställa diagnosen ADHD* 'make the diagnosis of ADHD'. Both appositive nouns belong to the type Disease.

The investigation of lexico-semantic alternation aims at delineating the spectrum of semantically related words which can occur in a specific position in a collocation. By semantically related words, we mean synonyms, antonyms, hyperonyms, hyponyms and cohyponyms. This step of the analysis aims at finding collocations involving words having a similar meaning, as in *febern/temperaturen sjunker* 'the fever/temperature goes down'. There are also cases with a short and a long form which function as synonymous expressions, e. g. *göra ultraljud/ultraljudsundersökning* 'lit. make an ultrasound examination'.

3.2 Verbs in Medical Noun-Verb Collocations

We now turn our attention to verbs as components of medical collocations. An overview of their lexico-grammatical and lexico-semantic features, with examples, is provided in table 3.

Many transitive verbs in the collocations examined can undergo diathesis alternation, for example *lägga om ett sår – såret läggs om* 'change the wound' – 'the wound is changed' and *ta en tablett – tabletten tas* 'take a pill' – 'the pill is taken'. Some collocations, however, resist diathesis alternation, e. g. *ta skada – *skada tas* 'lit. take harm'.

As regards valency, verbs in the collocations can be monovalent (*febern stiger* 'the fever rises'), bivalent (*patienten tar en tablett* 'the patient takes a pill', *patienten svarar på medicinen* 'the patient responds to medicine') or trivalent (*läkaren ordinerar medicin till patienten* 'the doctor prescribes medicine for the patient'). The verbs *ta och ordinera* in these constructions require the presence of a direct object, in contrast to the verb *svara* in *svara på medicinen*, which takes a prepositional object.

Verbs in collocations may differ according to their semantic weight. For this reason, they are often categorised as *full* or *light verbs* with regard to their contribution to meaning in a particular context. Full verbs, such as *stiga* 'rise', *falla* 'fall' and *sjunka* 'drop', are instantiated in collocations like *febern stiger/faller/sjunker*, while light verbs like *ha, ge, ta*, and *göra* ('have', 'give', 'take', 'make') appear in collocations like *ha en depression* 'have a depression', *ge blod* 'give blood', *ta blodprov* 'take a blood test' and *göra abort* 'have an abortion'. However, the distinction between full and light verbs is far from clear-cut, as

Analysis level	Manifestation	Example
Lexico-grammatical features	Diathesis	*lägga om ett sår – såret läggs om* 'change the wound' –'the wound is changed' *ta en tablett – tabletten tas* 'take a pill' – 'the pill is taken' *ta skada – *skada tas* 'lit. take harm'
	Valency	*febern stiger* 'the fever rises' *patienten tar en tablett* 'the patient takes a pill' *patienten svarar på medicinen* 'the patient responds to medicine' *läkaren ordinerar medicin till patienten* 'the doctor prescribes medicine for the patient'
Lexico-semantic features	Semantic weight	*febern stiger/faller/sjunker* 'the fever goes up/goes down/falls' *ha en depression* 'have a depression' *ge blod* 'give blood' *ta blodprov* 'take a blood test' *göra abort* 'have an abortion'
	Selection restriction	*tuberkulos smittar* 'tuberculosis is contagious' *sköterskan lade förband på patienten* 'the nurse applied a bandage to the patient' *patienten kom ur en depression* 'the patient came out of a depression' *patienten hostar blod* 'the patient coughs blood'
	Lexico-semantic alternation	*febern stiger/går upp* 'the fever rises/goes up' *febern sjunker/går ner/släpper* 'the fever falls/goes down/drops'

Table 3: Relevant features in the analysis of verbs in noun-verb collocations

lightness is a matter of degree (cf. *ha – ställa – ordinera* 'have – put – prescribe') (cf. Hanks et al. 2006).

The semantics of collocations can also be described in terms of selection restrictions, which support the classification of verbs according to the types of arguments they select. Here are some patterns from our data:

– DISEASE V

  (V = *debutera* 'make one's debut', *smitta* 'infect'…),
  e. g. *tuberkulos smittar* 'tuberculosis is contagious'

– PERSON V ARTIFACT PERSON

  (V = *ordinera* 'prescribe', *lägga* 'apply'…),
  e. g. *sköterskan lade förband på patienten* 'the nurse applied a bandage to the patient'

– PERSON V DISEASE

  (V = *ådra sig* 'be stricken with', *komma ur* 'come out of'…),
  e. g. *patienten kom ur en depression* 'the patient came out of a depression'

These patterns point to possible types of arguments, but often the types need some further specification. For example, not all diseases are contagious, as implied by the pattern DISEASE V. In addition, some of the patterns require further specification in terms of semantic roles (Hanks/Ježek 2008: 398). This is the case with the notion PERSON which can be found in different semantic roles. In the case of PERSON V ARTIFACT PERSON (*the nurse applied a bandage to the patient*), the PERSON in subject position is agent. In the pattern PERSON V DISEASE, PERSON is patient (*the patient came out of a depression*).

In analogy to the noun alternations already mentioned, it is important to examine whether there are verbs that are semantically related to the verb being analysed, being either synonyms or antonyms (e. g. *febern stiger* 'the fever rises', *febern går upp* 'the fever goes up'; *febern sjunker* 'the fever falls', *febern går ner* 'the fever goes down' and *febern släpper* 'the fever drops').

3.3 Medical Collocations as Multi-Word Units

The analysis of nouns and verbs presented in 3.1 and 3.2 reveals the complexity of inherent syntactic and semantic relations contributing to the meaning and structure of medical collocations. It also underpins our study of collocations seen as multi-word units. In table 4, we record and exemplify the characteristic properties of medical noun-verb collocations taken as a whole.

There are two colligational properties which reveal the lexico-grammatical preferences of the noun-verb collocations, namely syntactic position and valency. As to position, noun-verb collocations form either sentences or verb phrases. In a sentence, the noun takes subject position and the verb functions as a predicate (e. g. *febern stiger* 'the fever rises'). In a verb phrase, the verb is a predicate and the noun is either object (e. g. *ordinera medicin* 'prescribe medicine') or prepositional object (e. g. *svara på medicinen* 'respond to

Analysis level	Manifestation	Example
Lexico-grammatical features	Syntactic position	*febern stiger* 'the fever rises' *ordinera medicin* 'prescribe medicine' *svara på medicinen* 'respond to medicine'
	Syntactic valency	*patienten svarade bra på medicinen* 'the patient responded well to the medicine' *ta prov på patienten/celler* 'take a specimen from the patient, take cell specimens' *ta prov från fingret* 'take a specimen from the finger' *ta prov för klamydia* 'take a specimen for chlamydia' *hosta (upp) slem* 'cough (up) phlegm'
Lexico-semantic features	Lexical boundness	*gå in i en depression* 'go into (a) depression' *drabbas av diabetes/cancer/SJUKDOM* 'be stricken with diabetes/cancer/DISEASE'
	Selection restriction	see Syntactic valency for examples
	Semantic role	*sjuksköterskan tog några blodprov* 'the nurse took some blood tests' *patienten tog några blodprov* 'the patient took some blood tests'

Table 4: Relevant features in the analysis of noun-verb collocations

medicine'). Possible syntactic and semantic alternations of the nouns and the verbs within these positions have been discussed in sections 3.1 and 3.2.

As far as syntactic valency is concerned, the arguments in subject, object and prepositional object position constitute the syntactically preferential arguments in a noun-verb collocation. These arguments often undergo a form of syntactic and semantic binding, which determines the syntactic form of optional arguments. The binding applies to collocations with both full verbs and light verbs. In the verb phrase *svara bra på medicinen* 'respond well to the medicine', the adverb and its semantic realisation is implied by the verb and its prepositional object. Its semantic manifestation is restricted to the reference on the result scale positive – negative. In the case of collocations with light verbs, different manifestations of optional arguments are often possible. For example, the collocation *ta prov* 'take a specimen' can take a number of different types of syntactic and semantic optional arguments. The prepositional object can carry reference to the object being tested (*ta prov på patienten/celler* 'take a specimen from the patient, take cell specimens'), location for taking the specimen (*ta prov från fingret* 'take a specimen from the finger') and different indications relevant for diagnosis (*ta prov för klamydia* 'take a specimen for chlamydia'). As can be seen from the examples, these differences are manifested by three different prepositions. Finally, it is worth noting a change in valency in constructions with cognate objects, e. g. *hosta (upp) slem* 'cough (up) phlegm'. In this example, it is the particle and the following object which turn the intransitive verb hosta 'cough' into a transitive one, changing its basic activity meaning into that of accomplishment.

As implied, the syntactic and semantic bindings are two facets of the same phenomenon. In line with the approach of Martin (2006), two main types of collocation have been distinguished, namely *token bound* and *type bound collocations*. The token bound collocations show strong lexical binding between the verb and the noun, as in *gå in i en depression* 'go into a depression', and impose restrictions on semantic alternation. The type bound collocations display a tendency to weaker lexical binding between the verb and the noun and allow alternation within a type, e. g. *drabbas av diabetes/cancer/SJUKDOM* ('be stricken with diabetes/cancer/DISEASE'. The distinction between these two types of collocation has clear implications for how their entries can be represented in dictionaries.

Medical collocations, in particular with light verbs like *ta* 'take' and *göra* 'make', can be ambiguous in much the same way as simplex words can be. The collocation *ta blodprov* 'take a blood test' is open to two interpretations. An agentive reading is exemplified in *sjuksköterskan tog några blodprov* 'the nurse took some blood tests', while a semi-agentive one can be found in *patienten tog några blodprov* 'the patient took some blood tests'. In the latter case, the agentivity involves the act of giving consent to health care personnel to perform the action. At the same time, the referent of *patient* occupies the role of Undergoer. As follows from the above, the information provided by the context is of relevance for disambiguation tasks.

The account of the lexical, grammatical and semantic preferences of the collocations, undertaken in this section, is a prerequisite for the elaboration of a model for their representation in the lexicon. The lexicographic task is not quite straightforward. The corpus-derived data need to be reduced to generalizations concerning the instantiated lexico-grammatical

and lexico-semantic patterns, which need to be communicated to users in a comprehensible manner.

4 Representation of Noun-Verb Collocations in the Lexicon

The electronic lexicon is planned to comprise two linked modules, a general-purpose lexicon module and a collocational one. The following types of information extracted from the MedLex database will be represented in the general-purpose module: entry word, pronunciation, word class, inflectional forms, meaning indicated by a guide word, and a definition including a definition comment. Furthermore, the entry will provide information on compounds containing the entry word, on semantically related words as well as information about the most frequent collocations. Examples and some information on morphologically related words, focused on derivation, will complete the specification of the lexical part of the entry. In the collocational module, the focus is on a more exhaustive description of particular collocations, with emphasis on those building either noun or verb phrases. While the treatment of noun phrase collocation has been dealt with in Sköldberg/Toporowska Gronostaj (2008a, 2008b), verb phrase collocations are discussed in this section.

Below we reflect on what kind of information should be included in the collocation module and how it should be communicated to the user. But first, we address the issue of presenting multi-word collocational lemmas. There are two types of lemmas distinguished, namely those with exact wording, as in *gå in i en depression* 'go into (a) depression', and those with approximate wording, as in *ta MEDICIN* 'take MEDICINE'. These two types of lemmas match the distinction, already mentioned, between token and type bound collocations. The description of the collocations with exact wording is straightforward as compared to the description of those with approximate wording, since the former type is usually open to only one interpretation of its meaning and structure in contrast to the latter, which covers a number of possible lexical choices. The information needed on collocations with exact wording is intended to be represented in the lexicon as shown below. In what follows, the Swedish examples are presented in the first place and then rendered into English. The examples are meant to give an idea about the content and structure of the entries and nothing is said about the final design of entries in the electronic dictionary.

kollokation: *gå in i en depression*
definition: insjukna i en depression
konstruktion: *PERSON går in i en depression*
exempel: *han gick in i en depression och blev sjukskriven en längre tid.*
besläktade uttryck: *få en depression, drabbas av (en) depression*

collocation: *go into (a) depression*
definition: become depressed
construction: *PERSON goes into (a) depression*

examples: *he went into a depression and was on the sick leave for a long period.*
related expressions: *get a depression, be stricken with depression*

Collocations with approximate wording, being generalizations over a particular set of possible manifestations, require more information on lexical items in the set. To provide such information, the subfield "type specification" has been added to the structure, as shown in the example below.

kollokationstyp: *ta MEDICIN*
definition: tillgodogöra sig medicin, oftast genom att äta, dricka eller injicera den
typspecifikation: *MEDICIN – medicin, en tablett, antibiotika, aspirin, insulin…*
konstruktion: *PERSON tar MEDICIN (mot SJUKDOM/SYMTOM)*
PERSON tar MEDICIN (för SJUKDOM/SYMTOM)
exempel: *hon har glömt att ta sitt penicillin; han tar sprutor mot sin diabetes; han tar tabletter mot högt blodtryck; hon tar medicin för sin akne; hon tar Alvedon för sin ryggvärk*
besläktade uttryck: *svälja, inta, injicera; äta MEDICIN, gå på MEDICIN*
kommentar: i vardagligt språk kombineras *ta MEDICIN* ofta med prepositionen *för* i samma betydelse som *mot*

collocation type: *take MEDICINE*
definition: put medicine in the body, usually by eating, drinking or injecting it
type specification: *MEDICINE – medicine, a pill, antibiotic, aspirin, insulin…*
construction: *PERSON takes MEDICINE (against DISEASE/SYMPTOM)*
PERSON takes MEDICINE (for DISEASE/SYMPTOM)
examples: *she has forgotten to take his penicillin; he takes injections for his diabetes; he takes pills against high blood pressure; she takes medicine for her acne; she takes Alvedon for her backache*
related expressions: *swallow, consume, inject; eat MEDICINE, go on MEDICINE*
note: in everyday language, the preposition *för* 'for' is used in the same sense as *mot* 'against' in the construction *take MEDICINE*

The explication of the approximate wording is given a prominent position in the structure of the above entry. As already noted, the type specification provides a list of some typical manifestations of the semantic type named in the lemma. The list is open-ended, as new lexical items can be added; e. g. for the entry *take MEDICINE*, the type *MEDICINE* can subsume a number of different names of drugs. In consequence, the type-oriented approach seems to be superior to a mere listing of some of the most frequent collocations representing the lemma (e. g. *take medicine, take pills* etc.).

In the construction subfield, the preferred contexts of the collocations are specified and the user is faced with information on valency, optionality of the arguments and types of referents. In the next field, the constructions are exemplified. The note subfield is reserved for additional information on the collocations, which deserves attention beyond that given in the entry.

The example discussed above is but one of many type bound collocations which may usefully be represented in this way. The lemma *ta PROV* shows that even semantically and

syntactically more complex cases can be dealt with using the convention with approximate wording:

> **kollokationstyp:** *ta PROV*
> **definition:** samla in eller lämna en liten mängd organiskt material för analys
> **typspecifikation:** *PROV – prov, cellprov, blodprov, vävnadsprov…*
> **konstruktion:** *PERSON tar PROV på PERSON/DEL AV MÄNNISKOKROPP*
> *PERSON tar PROV från DEL AV MÄNNISKOKROPP*
> **exempel:** *sjuksköterskan tog flera blodprover på patienten; vid traditionell fosterdiagnostik tas prover på antingen moderkaka eller fostervatten; läkaren tog prover från svalget och konstaterade att det var halsfluss; han åkte till sjukhuset för att ta prover*
> **besläktade uttryck:** *undersöka, testa; lämna prov*
> **kommentar:** kollokationen betyder att man antingen utför eller utsätter sig för någon form av provtagning. Det är sammanhanget som avgör tolkningen. Däremot betyder kollokationen *lämna prov* endast 'underkasta sig provtagning'.

> **collocation type:** take SPECIMEN
> **definition:** collect or leave a specimen of organic material for the purpose of analysis
> **type specification:** SPECIMEN – *sample, specimen, blood test, tissue sample…*
> **construction:** *PERSON takes SPECIMEN of PERSON/PART OF HUMAN BODY*
> *PERSON takes SPECIMEN from PART OF HUMAN BODY*
> **examples:** *the nurse took a blood test from the patient; in connection with traditional foetal diagnosis, samples are taken from either placenta or amniotic fluid; the doctor took a sample from the throat and diagnosed tonsillitis; he went to the hospital to take specimens*
> **related expressions:** *examine, test; leave sample*
> **note:** the collocation means either to perform or to undergo tests. Its context determines the interpretation. By contrast, the collocation *leave a sample* means only 'to undergo medical tests'.

As already hinted, the collocation *ta PROV* 'take SPECIMEN' is ambiguous – its interpretation is determined by the semantic role of the subject. This information is, in an implicit way, conveyed to the user through the definition and the examples. It is further explained in the note and an alternative collocation is suggested.

The issues concerning the representation of noun-verb collocations, discussed in this section, show that prior onomasiological and semasiological analysis of data supports the modes of description of collocational entries in the lexicon. They also show that an integrated semantic and syntactic notation can communicate relevant facts about collocations in a user-friendly way.

5 Conclusion

This study deals with medical noun-verb collocations in Swedish, retrieved from general and medical corpora. The corpus-based analysis of the lexico-grammatical and lexico-semantic features of the collocations has shown that there is a broad spectrum of idiosyncratic preferences which need to be accounted for in a lexicon. The classification of the collocations into token- and type-bound ones has been proven relevant for distinguishing two types of multi-word lemmas, namely those with exact wording (e. g. *gå in i en depression* 'go into (a) depression') and those with approximate wording (e. g. *ta MEDICIN* 'take MEDICINE'). In consequence, those two types of lemmas impose partly different requirements on the structure and the content of the entry. The proposed differentiation of lemmas matches well the inherent heterogeneity of the collocations. It also meets lexicon users need to find guidance on both narrowly restricted and open-ended expressions.

The study undertaken has confirmed our conviction that the treatment of collocations in lexicons can go beyond mere listing. The corpus has proved to be an excellent source of information on the lexico-grammatical and lexico-semantic behaviour of collocations. Furthermore, the corpus has made it possible to verify a number of hypotheses concerning types of syntactic and semantic relations between the collocations examined and their context. The study has also shown that an integrated semasiological and onomasiological analysis paves the way for a more exhaustive and stringent description of collocations used in medical language.

Bibliography

Bernal, E./De Cesaris, J. (eds.) (2008): *Proceedings of the XIII Euralex International Congress*. Barcelona.

Corino, E./Marello, C./Onesti, C. (eds.) (2006): *Proceedings of the XII Euralex International Congress*. Torino.

Hanks, P./Urbschat, A./Gehweiler, E. (2006): "German light verb constructions in corpora and dictionaries", in: *International Journal of Lexicography, 19/4*; 439-458.

Hanks, P./Ježek, E. (2008): "Shimmering lexical Sets". In: Bernal/De Cesaris (eds.) (2008); 391-402.

Heid, U. (2004): "On the presentation of collocations in monolingual dictionaries". In: Williams/Vessier (eds.) (2004); 729-738.

Hoey M./Brook O'Donnell, M. (2008): "Lexicography, Grammar and Textual Position", in: *International Journal of Lexicography, 21/3*; 293-311.

Kokkinakis, D./Toporowska Gronostaj, M. (2008): "MedLex+: an Integrated Corpus-Lexicon Medical Workbench for Swedish". In: Bernal/De Cesaris (eds.) (2008); 703-712.

Malmgren, S.-G. (2003): "Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet", in: *Språk & Stil. Tidskrift för svensk språkforskning, 13*; 123-168.

Martin, W. (2006): "Frame-based Lexicons and Making of Dictionaries". In Corino et al. (eds.) (2006); 281-293.

Siepmann, D. (2005): "Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects", in: *International Journal of Lexicography, 18/4*; 409-443.

Siepmann, D. (2006): "Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects", in: *International Journal of Lexicography, 19/1*; 1-39.
Sköldberg, E./Toporowska Gronostaj, M. (2008a): "From Subdomains and Parameters to Collocational Patterns: On the Analysis of Swedish Medical Collocations". In: Bernal/De Cesaris (eds.) (2008); 1421-1432.
Sköldberg, E./Toporowska Gronostaj, M. (2008b): "Modell for beskrivning av kollokationer i ett medicinskt lexikon (MedLex)". In: Svavarsdóttir et al. (eds.) (2008); 433-445.
Svavarsdóttir, Á./Kvaran, G./Ingólfsson, G./Jónsson, J. H. (eds.) (2008): *Nordiske Studier i Leksikografi 9. Rapport fra konference om leksikografi i Norden, Akureyri.* Reykjavík.
Williams, G./Vessier, S. (eds.) (2004): *Proceedings of the XI EURALEX International Congress*. Lorient.

Maria Toporowska Gronostaj
Center for Lexicography and Lexicology
Department of Swedish
University of Gothenburg
Box 200
405 30 Göteborg
Sweden
maria.gronostaj@svenska.gu.se

Emma Sköldberg
Center for Lexicography and Lexicology
Department of Swedish
University of Gothenburg
Box 200
405 30 Göteborg
Sweden
emma.skoldberg@svenska.gu.se

Kollokationen in Wissenschaftssprachen: Zur lernerlexikographischen Relevanz der Textarten- und Diskursspezifik von Kollokationen

Franziska Wallner

The present study analyzes the question if the use of collocations in academic communication should receive special attention with regard to the treatment of collocations in learner lexicography. First, a short introduction into the topics of collocations and academic language is given, paying special attention to the aspects of foreign language acquisition. Next, we present selective results of our study, which analyzes the specifications of application of collocations depending on the type of corpora which is looked at: either scientific papers, or press releases. The specifications of applications of collocations in both corpora types are determined and opposed to each other. Our results show that – at least in some cases – highly significant differences can be found in the use of collocations in scientific papers compared to other types of context in which collocations are used.

1 Kollokationen in Wissenschaftssprachen

Aus der Perspektive des Fremdsprachenerwerbs wurden sowohl Kollokationen als auch Wissenschaftssprachen bereits mehrfach untersucht. Insbesondere die mit den Kollokationen einher gehenden Schwierigkeiten wurden von Seiten der Sprachlehrforschung bereits ausführlich beschrieben (u. a. Hausmann 1984, Bahns 1997, Siepmann 2004, Reder 2006). Vor allem in Hinblick auf die Textproduktion gelten Kollokationen als ein besonders fehlerträchtiger Bereich, was auf unterschiedliche Faktoren zurückgeführt werden kann: Als fixierte, schwach- bzw. nichtidiomatische polylexikalische Einheiten lassen sich Kollokationen nur schwer von freien Wortverbindungen abgrenzen. Für den Nichtmuttersprachler ist ihre Fixiertheit jedoch weder formal noch inhaltlich evident, sodass sie häufig nicht als komplexe Einheiten erkannt und memoriert werden. Zudem haben kontrastive Untersuchungen gezeigt, dass interlinguale Divergenzen in diesem Bereich vollkommen unvorhersehbar auftreten, wobei viele Abweichungen nicht unbedingt gravierend sind, d. h. die Kollokationen bleiben „verstehbar, weil in ihrer Zusammensetzung nachvollziehbar und damit lernunauffällig“ (Börner/Vogel 1994a: 16).

Die mit den Kollokationen einhergehenden Ausdrucksschwierigkeiten betreffen sowohl die Auswahl der Kollokationspartner, als auch Aspekte der Verwendung von Kollokationen. In der Forschungsliteratur wird diesbezüglich auf normbedingte morphosyntaktische Transformationsrestriktionen und stilistische Einschränkungen verwiesen (u. a. Ludewig 2005, Ritz/Heid 2006).

Auch die Produktion und die Rezeption wissenschaftssprachlicher Texte wurden bereits als Problembereich von Nichtmuttersprachlern erkannt und beschrieben (u. a. Ehlich 1993, 1999, Graefen 1997, Moll 2004). Dabei hat sich insbesondere die allgemeinsprachliche Dimension wissenschaftssprachlicher Texte, die so genannte *alltägliche Wissenschaftssprache*[1], als verantwortlich für die Schwierigkeiten von Nichtmuttersprachlern gezeigt. Die sprachlichen Mittel der alltäglichen Wissenschaftssprache wurden zu einem beträchtlichen Teil aus der Alltagssprache entlehnt (Graefen 1999: 225). Dabei hat sich eine wissenschaftsspezifische Bedeutung etabliert und verfestigt, die sich jedoch häufig nicht ohne Weiteres aus der allgemeinsprachlichen Bedeutung ableiten lässt.[2] Der vermeintlich leichte Zugang zur Bedeutung aufgrund des gemeinsprachlichen Ursprungs der lexikalischen Einheiten erweist sich dadurch nicht selten als bloßer Schein von Verständlichkeit (Graefen 2004: 295). Die Problembereiche von Nichtmuttersprachlern bei der Produktion wissenschaftssprachlicher Texte betreffen neben den von einer Forschergemeinschaft etablierten Konventionen hinsichtlich zentraler Form-, Inhalts- und Strukturmerkmale der wissenschaftlichen Text- und Diskursarten auch die in den unterschiedlichen Wissenschaftstraditionen verhaftete Kulturspezifik der Wissenschaftskommunikation generell sowie der einzelnen Text- und Diskursarten.[3] Darüber hinaus erschweren die sprachlichen Defizite vieler Nichtmuttersprachler die Aneignung wissenschaftssprachlicher Strukturen und Handlungsformen sowie die Entwicklung eines ausgeprägten Bewusstseins für wissenschaftsspezifische Bedeutungen und Verwendungsweisen. Auf lexikalischer Ebene äußert sich das vor allem beim Gebrauch komplexer Wortverbindungen. Hier kommt es zu Vermischungen und Analogiebildungen von vorgeformten Syntagmen und zur Verwendung unpassender Präpositionen und Tempora (vgl. u.a. Graefen 1999, 2004, Moll 2004).

Die Auseinandersetzung mit Kollokationen in der Wissenschaftskommunikation steht zwar noch am Anfang, jedoch weist einiges darauf hin, dass auch Kollokationen in ihrem Gebrauch wissenschaftsspezifische Züge tragen können. Anhand einzelner Fachbereiche konnte bereits die Existenz domänenspezifischer Kollokationen nachgewiesen werden (Sternkopf 1998, Caro Cedillo 2004). Die Analyse wissenschaftlicher Texte von nichtmuttersprachlichen Studierenden hat neben einem eindeutig fehlerhaften Gebrauch von

1 Das Konzept der alltäglichen Wissenschaftssprache wurde von Ehlich (1993,1999) geprägt und umfasst „die je spezifische Nutzung von Teilen der Alltagssprache für die Zwecke der Wissenschaft“ (Ehlich 2000) jenseits der Fachterminologien.

2 Oft liegen der wissenschaftsspezifischen Bedeutung einer lexikalischen Einheit metaphorische Prozesse zugrunde, wie etwa bei den Verben *zeigen, herausarbeiten* und *offenlegen*, die primär keine Sprechhandlungsbedeutung aufweisen (Fandrych 2002: 3).

3 Dies äußert sich Moll (2004) zufolge bspw. in fehlenden Quellenangaben, durch deutliche Strukturen des mündlichen Sprachgebrauchs oder auch durch die Veranschaulichung komplexer Inhalte anhand bildhafter Vergleiche, die jedoch von der innerhalb der Wissenschaftskommunikation üblichen Metaphorik abweichen (vgl. Moll 2004: 360).

Kollokationen (vgl. Textbeispiel (1)) auch stellenweise eine unangemessene Verwendungsweise dieser Wortverbindungen offen gelegt (vgl. Textbeispiele (2)-(4)):[4]

(1) *Einerseits wurde **dem Einfluss** der L1 auf den FSE zu viel **Aufmerksamkeit gelenkt**, ...*

(2) *Auch das Suffix -los weist eine **große Produktivität** aus.*

(3) *Es ist auch zu bemerken, dass in diesen Textsorten **gerne** einfache Verben **gebraucht** werden.*

(4) *Hermeneutik ist eigentlich aus dem Verfahren der Bibelinterpretation. Aber in der Neuzeit hatten Friedrich Schlegel und Friedrich Schleiermacher **machten sie zur wichtigen Methode** des Geisteswisschenschaft, mit der man die Wechselbeziehung zwischen der Bedeutung der einzelne Worte und der des Gesamtkontextes erörten und erklären kann.*

Während es sich bei dem Textbeispiel (1) um eine Verschränkung der beiden Wortverbindungen *Aufmerksamkeit + lenken* und *Aufmerksamkeit + schenken*[5] handelt, können die Textbeispiele (2) bis (4) nicht unbedingt als Kollokationsverstöße eingestuft werden. Unter bestimmten Voraussetzungen wären sie durchaus akzeptabel, doch wirken sie bereits ohne größere kontextuelle Einbettung deplatziert und unangemessen. Eine sprachbereichsspezifische Vorgabe, die sehr wahrscheinlich mit der Tatsache einhergeht, dass die Textbeispiele dem Bereich Wissenschaftssprache entstammen, scheint hier nicht erfüllt zu sein.

Es stellt sich die Frage, ob derartige sprachbereichspezifischen Vorgaben den Gebrauch von Kollokationen innerhalb der Wissenschaftskommunikation generell bestimmen und zu lexikographisch relevanten Abweichungen gegenüber den Gebrauchsspezifika von Kollokationen in anderen Verwendungskontexten führen.

Um dieser Frage nachzugehen, wurden in der hier vorgestellten Studie die Gebrauchsspezifika ausgewählter Kollokationen auf Grundlage zweier unterschiedlich zusammengesetzter Korpora ermittelt und miteinander verglichen.

---

4 Die Textbeispiele (1)-(3) stammen aus Hausarbeiten von Nichtmuttersprachlern im BA-Studiengang Deutsch als Fremdsprache am Herder-Institut der Universität Leipzig im WS 2007/2008. Textbeispiel (4) entstammt dem fehlerannotierten Lernerkorpus des Deutschen als Fremdsprache (Falko). Genauere Informationen finden sich unter: *http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko* [Stand 08.03.2009]. Die Textbeispiele wurden unverändert (d.h. inklusive fehlerhafter Textstellen) übernommen. Die Hervorhebungen der den Bereich der Kollokationen betreffenden problematischen Textstellen wurden nachträglich vorgenommen.

5 Auf eine lexikographische Nennform wird im Rahmen des Beitrags verzichtet, da diese m. E. optimalerweise aus einer korpuslinguistischen Auswertung von Kontextbelegen resultieren sollte.

2 Vergleichende Untersuchung von Kollokationen in unterschiedlichen Verwendungskontexten

2.1 Untersuchungskorpora und Analyseverfahren

Zur Ermittlung der Gebrauchsspezifika von Kollokationen in der Wissenschaftskommunikation wurde ein separat konsultierbares Teilkorpus des DWDS-Kernkorpus der Berlin-Brandenburgischen Akademie der Wissenschaften ausgewählt[6]. Mit insgesamt 25 Millionen Token handelt es sich dabei um das derzeit[7] umfangreichste Korpus der deutschen Wissenschaftssprache. Es enthält wissenschaftliche Texte, die dem gesamten 20. Jahrhundert entstammen. Um der Analyse möglichst aktuelle Daten zugrunde zu legen, wurde die Recherche auf Texte ab 1950 beschränkt und zusätzlich auf eine private Korpussammlung ausgewählter germanistischer Fachzeitschriften (im weiteren Verlauf GK-Korpus genannt) aus den Jahren 2000-2007 mit einem Umfang von ca. 1,2 Millionen Textwörtern zurückgegriffen.[8]

Vor dem Hintergrund, dass bereits vorliegende lexikographische Analysen von Mehrwortverbindungen zumeist mit Korpora durchgeführt wurden, die ausschließlich bzw. zum Großteil Pressetexte umfassen (so bspw. Hollós 2004, Hallsteindóttir 2007, Heine 2008 aber auch OWID[9] und das Wortschatz Portal der Universität Leipzig)[10], wurde als Vergleichskorpus das pressesprachliche Teilkorpus des DWDS-Kernkorpus herangezogen. Auch dieses Korpus wurde auf Texte ab 1950 reduziert und um ein Teilkorpus der Berliner Zeitung (BZ-Korpus)[11] mit Texten aus dem Zeitraum 2000-2005 erweitert.

Die Untersuchung konzentrierte sich auf verbonominale Kollokationen des Typs S_a + V.[12] Dabei kamen zunächst solche Verbindungen in Betracht, die aufgrund eines

6 Das DWDS-Kernkorpus wird als ein „ausgewogenes Textkorpus des 20. Jahrhunderts" bezeichnet. Es enthält belletristische (26%), journalistische (27%) und wissenschaftliche Texte (22%) sowie Gebrauchstexte (20%) und Texte gesprochener Sprache (ca. 5%). Vgl. *http://www.dwds.de/textbasis/kerncorpus* [10.03.2009].

7 Stand 2009.

8 Das Korpus wurde von Cordula Meißner zusammengestellt und setzt sich aus 190 elektronisch verfügbaren germanistischen Zeitschriftenartikeln zusammen. Die Artikel stammen aus 13 verschiedenen Zeitschriften von insgesamt 178 AutorInnen mit deutscher Muttersprache.

9 Online-Wortschatz-Informationssystem Deutsch des Instituts für Deutsche Sprache in Mannheim.

10 Demgegenüber stehen lexikographische Projekte, die der Forderung nach einer ausgewogenen Datengrundlage durch eine Berücksichtigung verschiedener Sprachbereiche (weitestgehend) gerecht werden. Verwiesen sei auf das Herder BYU-Korpus als Datengrundlage für das *Frequency Dictionary of German* von Jones/Tschirner (2006) oder auch auf eine bisher lediglich als Alphaversion zugängliche lexikographische Sprachlernplattform, der als Datengrundlage das DWDS-Kernkorpus dient (vgl. *http://odo.dwds.de/wortprofil/project.php?show=80* [Stand 26.02.09].

11 Ein ebenfalls über die DWDS-Homepage zugängliches Korpus – verfügbar unter *http://www.dwds.de/textbasis* [10.03.09].

12 Der Untersuchung wurde ein Kollokationsbegriff zugrunde gelegt, der sich ausschließlich auf nicht- bzw. schwachidiomatische Wortverbindungen konzentriert, zwischen deren Bestandteilen eine syntaktische Relation besteht. Den Kollokationen des Typ S_a+V liegt demzufolge

stark polysemen verbalen Bestandteils ein hohes Fehlerpotential aufweisen.[13] Auch die Vorkommenshäufigkeit des nominalen Bestandteils sowie ein hoher korpusstatistischer Assoziationswert zwischen den Kombinationspartnern spielten eine Rolle. Aufgrund des geringen Umfangs des wissenschaftssprachlichen Korpus (lediglich 13,7 Millionen Token)[14] war es allerdings erforderlich, die Auswahl in erster Linie an der Anzahl der für eine Auswertung verfügbaren Kontextbelege zu orientieren. Hierbei wurde eine Stichprobengröße von mindestens 80 Kontextbelegen angestrebt.[15] Die folgenden sieben verbonominalen Kollokationen wurden schließlich für die Studie als Fallbeispiele ausgewählt:[16]

Anspruch + *erheben* (Stichprobengröße: 80)
Auffassung + *vertreten* (Stichprobengröße: 67)
Beitrag + *leisten* (Stichprobengröße: 85)
Frage + *behandeln* (Stichprobengröße: 62)
Möglichkeit + *bieten* (Stichprobengröße: 79)
Versuch + *unternehmen* (Stichprobengröße: 82)
Ziel + *verfolgen* (Stichprobengröße: 81)

Die als Fallbeispiele ausgewählten Kollokationen wurden auf morphologischer, syntaktischer und semantischer Ebene analysiert. Im Mittelpunkt des Interesses standen dabei Merkmalsklassen, die für eine lernerlexikographische Kodifizierung von Kollokationen besondere Relevanz besitzen:

---

stets eine Verb-Objekt-Beziehung zugrunde – im Gegensatz zu den Funktionsverbgefügen, die in Anlehnung an Heine (2006) als komplexe Prädikate aufgefasst werden.

13 Es handelt sich dabei um mitunter auch als *schwachidiomatisch* bezeichnete Wortverbindungen, deren als Kollokatoren fungierende verbale Bestandteile innerhalb der Wortverbindung eine Lesart annehmen, die von der *üblichen* bzw. zumeist *geläufigeren* Bedeutung des zugrunde liegenden Lexems abweicht. Ihre Semantisierung kann nur unter Rückgriff auf den nominalen Bestandteil erfolgen. Aus Sicht eines Nichtmuttersprachlers ist die Kombinatorik häufig nicht nachvollziehbar und stellt daher eine besondere Schwierigkeit dar.

14 Für eine Analyse von Mehrwortverbindungen bedarf es eigentlich Korpora mit einem Umfang von über 100 Millionen Token (vgl. u.a. Tschirner 2005: 137), da nur so zu einer möglichst hohen Anzahl an Wortverbindungen die Bereitstellung von ausreichend Kontextbelegen gewährleistet werden kann.

15 Die Stichprobengröße resultiert aus Voruntersuchungen sowie aus den Untersuchungsergebnissen von Heine (im Druck). Hier hat die Gegenüberstellung der Analysen von zwei Funktionsverbgefügen mit unterschiedlichen Stichprobengrößen (40, 80 und 120 Belege) gezeigt, dass die Ergebnisse bei 40 und 80 Belegen deutlich auseinander gehen, während beim Vergleich der Auswertungen von 80 und 120 Belegen nur geringfügige Abweichungen feststellbar sind. Genauere Hinweise zu Fragen der Repräsentativität der Analyseergebnisse finden sich bei Heine (im Druck).

16 Aufgrund von Einschränkungen in der Zugänglichkeit der Kontextbelege sowie durch die Unzuverlässigkeit der verwendeten Analysetools, die eine nachträgliche manuelle Aussortierung von Kontextbelegen erforderlich machte, konnte die Stichprobengröße bei den Kollokationen *Auffassung* + *vertreten* und *Frage* + *behandeln* nicht eingehalten werden.

- typische morphologisch-syntaktische Strukturformeln, in denen die Kollokationen bevorzugt verwendet werden,
- der Numerus des nominalen Bestandteils,
- die Artikelwahl zum nominalen Bestandteil,
- die Möglichkeit von adjektivischen Attributen zum nominalen Bestandteil,
- die verbalen Kategorien Tempus, Modus und Genus,
- der Gebrauch von Modalverben,
- die Möglichkeiten einer Negation,
- sowie die semantische Bestimmung der Subjektklassenzugehörigkeit.[17]

Für jedes Fallbeispiel wurde die Vorkommenshäufigkeit der einzelnen Merkmalsklassen sprachbereichsspezifisch – d.h. für das wissenschaftssprachliche Korpus und das pressesprachliche Korpus jeweils gesondert – ermittelt.[18] Anschließend wurden die Frequenzangaben zu den einzelnen Merkmalsklassen miteinander verglichen.

Um zu überprüfen, ob es sich bei eventuellen Abweichungen hinsichtlich der Vorkommenshäufigkeit einzelner Merkmale um eher zufällige Verteilungen handelt oder ob die Abweichung auf die Beschaffenheit des Korpus und damit auf den jeweiligen Sprachbereich zurückgeführt werden können, wurde mit Hilfe eines Chi-Quadrat-Tests[19] die Irrtumswahrscheinlichkeit berechnet. Auf diese Weise war es möglich, signifikante Abweichungen hinsichtlich der Gebrauchsspezifika der untersuchten Kollokationen zu ermitteln. Lag die Irrtumswahrscheinlichkeit unter einem Prozent wurden die Abweichungen als hochsignifikant eingestuft und durch drei Sternchen gekennzeichnet (***). Als signifikant gelten

17 Auf eine semantische Beschreibung der Klassenzugehörigkeit weiterer Aktanten wurde im Rahmen der Untersuchung verzichtet, da eine objektive Zuordnung durch eine Einzelperson nicht realisierbar ist. Die Konsultation weiterer Sprecher hätte jedoch den Rahmen dieser Untersuchung überschritten. Aufwand und Nutzen stünden dabei in keinem sinnvollen Verhältnis.

18 Die Ermittlung der Vorkommenshäufigkeiten erfolgte für jedes Merkmal gesondert, ungeachtet der Gefahr, dass dadurch Hinweise auf eventuelle Korrelationen zwischen einzelnen Merkmalen bzw. Merkmalsklassen verloren gehen könnten. Um Aufwand und Nutzen der Analyse in einem ausgewogenen Verhältnis zu halten, wurde sich jedoch im Rahmen dieser manuellen Analyse auf eine gesonderte Betrachtung der einzelnen Merkmale beschränkt und nur in Einzelfällen das Zusammenwirken verschiedener Merkmale berücksichtigt. Eine Untersuchung, die das Zusammenspiel einzelner Merkmale systematisch analysiert, kann m. E. nur mit Hilfe automatischer Analysetools realisiert werden.

19 Der Chi-Quadrat-Test ist ein Verfahren zur statistischen Analyse von Häufigkeitsverteilungen. Erwartete Häufigkeiten werden dabei mit tatsächlich beobachteten Häufigkeiten verglichen. Berechnet wird die Signifikanz der Abweichung von einer Menge von Zufallsvariablen von einem hypothetisch angenommenen Wert (vgl. Bortz/Döring 2006).

Nominaler Bestandteil		Pressekorpus	Wissenschaftskorpus
Ziel		81 Belege	81 Belege
	im Singular	48 (59,26%)	31 (38,2%)
	im Plural	33 (40,74%)	50 (61,73%)
Möglichkeit		79 Belege	79 Belege
	im Singular	62 (78,48%)	49 (62,03%)
	im Plural	17 (21,52%)	30 (37,97%)
Anspruch		80 Belege	80 Belege
	im Singular	63 (78,75%)	77 (96,25%)
	im Plural	17 (21,25%)	3 (3,75%)

Tabelle 1: Numerus des nominalen Bestandteils

Abweichungen mit einer Irrtumswahrscheinlichkeit unter fünf Prozent (gekennzeichnet durch zwei Sternchen) und als marginal signifikant (gekennzeichnet durch ein Sternchen) sind diejenigen Abweichungen ausgewiesen, bei denen eine Irrtumswahrscheinlichkeit von unter zehn Prozent ermittelt wurde.

2.2 Auswertung der Ergebnisse

Die Datenerhebung hat gezeigt, dass sich die analysierten Kollokationen in ihrem Gebrauch in den einzelnen Verwendungskontexten unterscheiden. Anhand einiger Beispiele soll im Folgenden eine Auswahl besonders signifikanter sprachbereichspezifischer Unterschiede dokumentiert werden. Eine ausführliche Dokumentation der Ergebnisse erfolgt im Rahmen meiner Dissertationsschrift.

A. Obligatorische und fakultative Aktanten

Numerus

Bei der jeweils im Akkusativ auftretenden nominalen Basis wurden hinsichtlich des Numerus bei drei der sieben ausgewerteten Kollokationen signifikante Abweichungen festgestellt. Handelt es sich bei den Verbindungen *Ziel* + *verfolgen* und *Möglichkeit* + *bieten* um marginal signifikante Unterschiede, kann bei der Verbindung *Anspruch* + *erheben* der im wissenschaftlichen Korpus scheinbar unübliche Gebrauch des nominalen Bestandteils im Plural als hoch signifikante Abweichung gegenüber der Pluralverwendung von *Anspruch* im Pressekorpus eingestuft werden.

Bei den Kollokationen *Ziel* + *verfolgen* und *Möglichkeit* + *bieten* zeichnet sich im wissenschaftssprachlichen Korpus eine leichte Affinität zum Gebrauch des nominalen Bestandteils im Plural (bei *Ziel* + *verfolgen*) bzw. im Singular (bei *Möglichkeit* + *bieten*) ab. Von einer Gebrauchspräferenz kann bei diesen Verbindungen aufgrund des jeweiligen Anteils von Singular und Plural jedoch nicht ausgegangen werden. Bei der Verbindung *Anspruch* + *erheben* ist eine recht ausgeprägte Affinität zum Gebrauch der nominalen

Strukturformel	Pressekorpus	Wissenschaftskorpus
$S_n + V_{fin} + S_a +$ *auf* $+ S_a$	42 (52,50%)	18 (22,50%) ***
$S_n + V_{fin} +$ Adj $+ S_a +$ *auf* $+ S_a$	2 (0,25%)	0
$S_n + V_{fin} + S_a +$ Infinitivkonstruktion	10 (12,50%)	37 (46,25%) ***
$S_n + V_{fin} +$ Adj $+ S_a +$ Infinitivkonstruktion	3 (3,75%)	0

Tabelle 2: Anteil ausgewählter Strukturformeln zur Kollokation *Anspruch + erheben*

Basis im Singular in beiden Korpora zu beobachten. Da der Gebrauch von *Anspruch* im Plural im Pressekorpus immerhin 17 Mal belegt werden konnte, wäre die Annahme einer Gebrauchsrestriktion allerdings unbegründet. Im Wissenschaftskorpus fanden sich hingegen lediglich drei Kontextbelege, in denen *Anspruch* im Plural verwendet wurde. Daher kann man in diesem Fall durchaus von einer bevorzugten Verwendung des Singulars und damit von einer wissenschaftsspezifischen Gebrauchspräferenz ausgehen.

Strukturformeln

Mit Ausnahme der Kollokation *Beitrag + leisten* waren bei allen analysierten Fallbeispielen korpusspezifische Abweichungen hinsichtlich der morphosyntaktischen Struktur der Kollokation und/oder der unmittelbaren Kollokationsumgebung zu beobachten. Neben der Gebrauchsfrequenz eines adjektivischen Attributs zur nominalen Basis betreffen die Unterschiede die Vorkommenshäufigkeit einzelner Anschlussstrukturen, die in Verbindung mit den Kollokationen auftreten sowie die Notwendigkeit der Besetzung weiterer Leerstellen. So überwiegt bei der Verbindung Anspruch + erheben im wissenschaftssprachlichen Korpus der Anschluss durch eine Infinitivkonstruktion ($S_n + V_{fin} + S_a +$ Infinitivkonstruktion). Dies geht einher mit der Verwendung der nominalen Basis im Singular und eines bestimmten Artikels (vgl. Tabelle 1 auf der vorherigen Seite und Beleg (4)). Dem gegenüber steht die im Pressekorpus häufigere Struktur $S_n + V_{fin} +$ (Adj. +) $S_a +$ auf $+ S_a$ (vgl. Beleg (6)). Auch hier dominiert die nominale Basis im Singular, tritt aber deutlich öfter im Plural auf als das im Wissenschaftskorpus der Fall ist.

(5) „Jede Theorie, die den Anspruch erhebt, die Welt zu beschreiben, und in diesem Sinne universelle Geltung anstrebt, muß diese Notwendigkeit der Invisibilisierung mit in Rechnung stellen." (4, DWDS, Wiss, Luhmann 1997: 187)[20]

20 Die Kontextbelege wurden für die Datenpräsentation gekürzt und wie folgt gekennzeichnet: Die erste Ziffer verweist auf die interne Belegnummer, das zweite Kürzel steht für das jeweilige Teilkorpus (DWDS, GK oder BZ), mit „Wiss" sind Belege markiert, die dem wissenschaftssprachlichen Teilkorpus des DWDS-Kernkorpus entstammen, „Zei" kennzeichnet hingegen das pressesprachliche Teilkorpus des DWDS-Kernkorpus. Bei den weiteren Angaben handelt es sich um konkrete Quellenangaben, die von den jeweiligen Korpora ausgewiesen wurden.

	Pressekorpus	Wissenschaftskorpus
Spezifizierung von Personen bzw. Personengruppen, der die Möglichkeit geboten wird	35 (44,30%)	13 (16,46%) ***

Tabelle 3: Besetzung weiterer Leerstellen bei der Kollokation *Möglichkeit + bieten*.

Kollokation	Konjunktivbelege im Pressekorpus	Konjunktivbelege im Wissenschaftskorpus
Anspruch + erheben	23 (31,51%)	6 (7,50%) ***
Auffassung + vertreten	6 (9,23%)	2 (3,03%)
Beitrag + leisten	31 (43,06%)	5 (6,41%) ***
Frage + behandeln	13 (23,27%)	1 (1,82%) ***
Möglichkeit + bieten	25 (32,05%)	3 (3,80%) ***
Versuch + unternehmen	21 (29,58%)	1 (1,23%) ***
Ziel + verfolgen	25 (30,86%)	1 (1,23%) ***

Tabelle 4: Anteil des Konjunktivsgebrauchs innerhalb der Kontextbelege, in denen die jeweilige Kollokation in finiter Form verwendet wurde.

(6) „Seine Partei erhebe Anspruch auf etwa ein Viertel der Sitze in der Regierung und in den Majlis, mindestens auf fünf Minister und drei Vizeministerposten." (7, DWDS, Zei, Archiv der Gegenwart 62 (1992: 36841))

Hinsichtlich der Notwendigkeit der Besetzung weiterer Leerstellen konnten bei der Kollokation *Möglichkeit + bieten* hochsignifikante korpusspezifische Unterschiede festgestellt werden. So wurde im Pressekorpus in fast der Hälfte der Kontextbelege spezifiziert, *wem* eine Möglichkeit geboten wird (vgl. Tabelle 3 und Beleg (7)). Im wissenschaftssprachlichen Korpus war die Notwendigkeit einer Besetzung dieser Leerstelle deutlich seltener gegeben.

(7) „Die neue Landwirtschaftspolitik biete der Regierung die Möglichkeit, große landwirtschaftliche Projekte in Zusammenarbeit mit einzelnen Bürgern oder ausländischen Firmen in Angriff zu nehmen." (23, DWDS, Zei, Archiv der Gegenwart 54 (1984: 27628))

B. Verbale Kategorien

Auch im Bereich der verbalen Kategorien hat die Auswertung der Kontextbelege korpusspezifische Unterschiede offen gelegt. Sie betreffen in erster Linie den Gebrauch der Modi. So sind – mit Ausnahme der Kollokation *Auffassung + vertreten* – bei allen Fallbeispielen hochsignifikante korpusspezifische Unterschiede hinsichtlich des Konjunktivgebrauchs zu beobachten.

Anwendungsbereich des Konjunktivs	Pressekorpus	Wissenschaftskorpus
Indirekte Rede	137 (95,14%)	9 (47,37%)
Konditionalsatz	3 (2,08%)	2 (10,53%)
Konsekutivsatz	1 (0,69%)	0
Einfacher Satz als Wunschsatz	1 (0,69%)	1 (5,26%)
Höfliche oder vorsichtig abwägende Äußerung	1 (0,69%)	7 (36,84%)
Einfacher Satz mit Modalverbkonstruktion zum Ausdruck einer Forderung	1 (0,69%)	0

Tabelle 5: Vorkommenshäufigkeit der einzelnen Anwendungsbereiche des Konjunktivs nach Helbig/Buscha (2001) unter Berücksichtigung aller Fallbeispiele.

Der vergleichsweise häufigere Konjunktivgebrauch im Pressekorpus lässt sich mit der für diesen Sprachbereich typischen indirekten Redewiedergabe erklären (vgl. auch Wermke et al. 2005: 529). Der Gebrauch des Konjunktivs zur indirekten Redewiedergabe dominiert daher auch innerhalb des Pressekorpus deutlich gegenüber anderen Anwendungsbereichen (vgl. Tabelle 5). Zwar überwiegt auch im wissenschaftssprachlichen Korpus der Anteil der analysierten Kollokationen im Konjunktiv zur indirekten Redewiedergabe, doch ist der Anwendungsbereich des Konjunktivs zu einer „höflichen oder vorsichtig abwägenden Äußerung" (Helbig/Buscha 2001: 182)[21] beinah gleichermaßen vertreten (vgl. Belege (8) und (9)). Auffällig ist, dass im Rahmen dieses Anwendungsbereichs des Konjunktivs die Kollokationen stets in Verbindung mit einem Modalverb (i. d. R. mit *können*) anzutreffen sind. Die übrigen Anwendungsbereiche sind mit maximal drei Kontextbelegen pro Korpus in beiden Sprachbereichen deutlich unterrepräsentiert.

(8) „Besonders in der linguistischen Disziplin der Semantik zog kaum jemand in Betracht, dass empirische Verfahren einen wesentlichen Beitrag bei der Erstellung von Bedeutungstheorien oder zur Beschreibung konkreter Bedeutungen leisten könnten." (57, GK, *Linguistik Online*, Bärenfänger 2003)

(9) „Man könnte die Auffassung vertreten, daß die meisten Defektallele ja heterozygot unwirksam sind und daß das menschliche Erbgut durch Diploidie infolgedessen gegen solche Schäden wohl abgesichert ist." (45, DWDS, Wiss, Bresch 1964: 309)

Während die Auswertung der pressesprachlichen Kontextbelege im Konjunktiv die in der Literatur vertretene Auffassung, dass der Konjunktiv in Pressetexten bevorzugt zur indirekten Redewiedergabe verwendet wird (vgl. u. a. Wermke et al. 2005: 529), bestätigt, gestattet die Anzahl der Konjunktivbelege im Wissenschaftskorpus keine verallgemeinernden Aussagen. Der jeweilige Anteil der Kontextbelege im Konjunktiv zum Ausdruck einer höflichen

21 Helbig/Buscha (2001) zufolge handelt es sich hierbei um einen formelhaften Gebrauch des Konjunktivs, der in diesen Sätzen nicht mehr in einem deutlichen Gegensatz zu entsprechenden Sätzen im Indikativ steht. Die Modi seien hier vielmehr austauschbar, „ohne dass ein nennenswerter Bedeutungsunterschied erkennbar ist (...)" (Helbig/Buscha 2001: 182).

oder vorsichtig abwägenden Äußerung legt jedoch nahe, dass dieser Anwendungsbereich in den Wissenschaftssprachen eine größere Rolle spielt, als dass es bei Pressetexten der Fall ist. Mithilfe der Auswertung einer größeren Anzahl von wissenschaftssprachlichen Kontextbelegen im Konjunktiv könnte darüber Aufschluss gewonnen werden.

Auch bezüglich der anderen verbalen Kategorien wie Tempus und Genus sowie in Hinblick auf den Modalverbgebrauch konnten korpusspezifische Unterschiede festgestellt werden. Diese sind jedoch nicht bei allen Fallbeispielen gleichermaßen ausgeprägt. Aufgrund der immer noch geringen Vorkommenshäufigkeit der einzelnen Merkmalsklassen können anhand der Auszählungsergebnisse kaum Rückschlüsse in Hinblick auf das Vorliegen eines wissenschaftsspezifischen Gebrauchs gezogen werden. Daher sei an dieser Stelle lediglich auf die Tabelle 9 auf Seite 212 am Ende des Beitrags verwiesen, in der bei den einzelnen Fallbeispielen jeweils die Kategorien markiert wurden, die korpusspezifische Unterschiede aufweisen.

C. Semantische Beschreibung der Subjektklassenzugehörigkeit

Bei allen untersuchten Kollokationen hat die semantische Analyse der Subjektklassenzugehörigkeit signifikante korpusspezifische Abweichungen offen gelegt.

Mit Ausnahme der Kollokationen *Auffassung + vertreten* und *Versuch + unternehmen* betreffen die signifikanten korpusspezifischen Unterschiede bereits die prinzipielle Zuordnung des Subjekts zu den Kategorien „jemand“ und „etwas“.

Bei den Kollokationen *Beitrag + leisten* und *Frage + behandeln* überwiegt im wissenschaftssprachlichen Korpus die Kategorie „etwas“, während im Pressekorpus das Subjekt hauptsächlich der Kategorie „jemand“ zugeordnet werden kann. Auch die Kollokation *Möglichkeit + bieten* tritt bevorzugt mit einem Subjekt der Kategorie „etwas“ auf, was allerdings auch auf das Pressekorpus zutrifft. Bei der Kollokation *Anspruch + erheben* dominiert zwar die Kategorie „jemand“ im wissenschaftssprachlichen Korpus, doch ist diese Kategorie dort beinah gleichermaßen vertreten wie die Kategorie „etwas“, die im Pressekorpus vergleichsweise selten anzutreffen ist.

Die Zahlen erlauben bereits die vorsichtige Schlussfolgerung, dass die untersuchten Kollokationen (mit Ausnahme von *Auffassung + vertreten* und *Ziel + verfolgen*) in wissenschaftssprachlichen Texten häufiger ein Subjekt der Kategorie „etwas“ zulassen, als es bei ihrer Verwendung in der Pressesprache der Fall ist. Die geringe Vorkommenshäufigkeit der Kategorie „etwas“ bei den Kollokationen *Auffassung + vertreten* und *Versuch + unternehmen*, die sowohl das Presse- als auch das Wissenschaftskorpus betrifft, kann als Hinweis gelten, dass bei diesen Verbindungen eine generelle, d. h. für Press- und Wissenschaftssprachen gleichermaßen gültige Präferenz zum Gebrauch eines Subjekts der Kategorie „jemand“ vorliegt.

Bei den Kollokationen *Anspruch + erheben*, *Beitrag + leisten* und *Frage + behandeln* kann aufgrund der Zahlenangaben das Vorhandensein einer pressesprachlichen Präferenz zur Verwendung eines Subjekts der Kategorie „jemand“ angenommen werden. Das Verhältnis der beiden Kategorien innerhalb des Wissenschaftskorpus ist bei diesen Verbindungen hingegen relativ ausgewogen.

Kollokation	Subjekt *jmd.* im Pressekorpus	Subjekt *jmd.* im Wissenschaftskorpus	Subjekt *etw.* im Pressekorpus	Subjekt *etw.* im Wissenschaftskorpus
Anspruch + erheben	63 (90,00%)	39 (52,0%) **	7 (10,00%)	36 (48,00%) ***
Auffassung + vertreten	61 (98,39%)	55 (98,21%)	1 (1,61%)	1 (1,79%)
Beitrag + leisten	56 (81,16%)	30 (40,54%) ***	13 (18,84%)	44 (59,46%) ***
Frage + behandeln	27 (79,41%)	10 (41,67%) ***	7 (20,59%)	14 (48,33%)
Möglichkeit + bieten	20 (30,30%)	5 (6,41%) ***	46 (69,70%)	73 (93,59%) **
Versuch + unternehmen	48 (96,00%)	41 (91,11%)	2 (4,00%)	4 (8,89%)
Ziel + verfolgen	48 (60,76%)	58 (80,56%)	31 (39,24%)	14 (19,44%) **

Tabelle 6: Frequenz der Kategorien „jemand“ und „etwas“ in Subjektposition.

Bei der Kollokation *Möglichkeit* + *bieten* scheint wiederum eine wissenschaftssprachliche Präferenz zur Verwendung eines Subjekts der Kategorie „etwas" vorzuliegen, während die Vorkommenshäufigkeiten der beiden Kategorien im Pressekorpus die Annahme von Gebrauchspräferenzen bzw. Restriktionen nicht rechtfertigen.

Auch der Versuch einer Zuordnung der Subjekte zu einzelnen Verknüpfungspartnerklassen offenbarte zum Teil hochsignifikante korpusspezifische Unterschiede. Innerhalb der Kategorie „jemand" konnten bei sechs der analysierten Kollokationen Differenzen hinsichtlich der Klassenzugehörigkeit des Subjekts festgestellt werden. Eine Ausnahme bildet die Verbindung *Möglichkeit* + *bieten*, bei der zwar bei der Kategorie „jemand" insgesamt gesehen ein hochsignifikanter Unterschied zwischen Presse- und Wissenschaftskorpus besteht, jedoch aufgrund der geringen Vorkommenshäufigkeit dieser Kategorie im Wissenschaftskorpus keine weiteren Aussagen zu einzelnen Klassen abgeleitet werden können. Am häufigsten betreffen die korpusspezifischen Abweichungen hinsichtlich der Klassenzugehörigkeit des Subjekts die Klassen „Einzelpersonen", „Staaten/Länder" und „Organisationen/Institutionen".[22] Bei allen untersuchten Kollokationen überwiegt im Wissenschaftskorpus der Anteil der als Subjekt fungierenden Einzelpersonen – sowohl innerhalb des Korpus als auch beim kollokationsspezifischen Vergleich des Anteils der Einzelpersonen im Pressekorpus und im Wissenschaftskorpus. Währenddessen dominiert im Pressekorpus bei immerhin vier Kollokationen (*Anspruch* + *erheben*, *Auffassung* + *vertreten*, *Beitrag* + *leisten* und *Ziel* + *verfolgen*) die metonymische Verwendung einer Staaten-/Länderbezeichnung an der Subjektstelle, die im Wissenschaftskorpus insgesamt nur fünf Mal belegt werden konnte (vgl. Beleg (10)):

(10) „Er persönlich betonte, daß die Bundesrepublik keinen Anspruch auf die Gebiete, die heute Polen seien, erhebe." (9, DWDS, Zei, Archiv der Gegenwart 60 (1990: 34232))

Auch die Klasse „Organisationen/Institutionen" ist im Pressekorpus sehr viel häufiger als im Wissenschaftskorpus vertreten. Diesbezüglich erreichen die korpusspezifischen Unterschiede bei fünf der analysierten Kollokationen (*Versuch* + *unternehmen*, *Frage* + *behandeln*, *Anspruch* + *erheben*, *Beitrag* + *leisten* und *Auffassung* + *vertreten*) ein hohes Signifikanzniveau.

Eine gesonderte Betrachtung der *Wir*-Form in Subjektposition hat quantitativ gesehen keine nennenswerten lexikographisch relevanten Abweichungen offen gelegt.

Dieser Form wurde jedoch besondere Aufmerksamkeit gewidmet, da der Gebrauch der *Wir*-Form als eine in Wissenschaftssprachen häufig anzutreffende Möglichkeit zur *Ich*-Vermeidung gilt (vgl. Graefen/Thielmann 2007: 95). Eine eingehendere Analyse der Kontextbelege hat schließlich gezeigt, dass sich die Funktion der *Wir*-Form in beiden

22 Der Kategorie „jemand" wurden generell auch diejenigen Belege zugeordnet, in denen das Subjekt zunächst zwar keine Personenbezeichnung im engeren Sinne beinhaltet, jedoch ein metonymischer Gebrauch anzunehmen ist. In der Regel handelt es sich dabei um institutionalisierte Zusammenschlüsse von Personen wie Parteien, Organisationen und Verbände oder auch Staaten bzw. Staatengemeinschaften.

Kollokation	*Wir*-Form im Pressekorpus	*Wir*-Form im Wissenschaftskorpus
Anspruch + erheben	6 (9,52%)	3 (7,69%)
Auffassung + vertreten	2 (3,28%)	2 (3,64%)
Beitrag + leisten	5 (8,93%)	1 (3,33%)
Frage + behandeln	0	3 (30,00%)
Möglichkeit + bieten	0	0
Versuch + unternehmen	2 (4,17%)	1 (2,44%)
Ziel + verfolgen	8 (16,67%)	1 (1,72%)

Tabelle 7: Frequenz der *Wir*-Form in Subjektposition.

Korpora deutlich unterscheidet. So ist sie im Pressekorpus stets innerhalb der direkten Redewiedergabe anzutreffen:

(11) Breshnew zur Lage in Asien; UdSSR lehnt chinesisches Verhandlungsangebot ab; Stellungnahme zum Treffen in Wladiwostok [26.11.74]:
„Wir erheben keinerlei Ansprüche auf irgendwelche ausländischen Gebiete, daher gibt es in diesem Sinne für uns auch keine, umstrittenen Gebiete“ (34, DWDS, Zei, Archiv der Gegenwart 44 (1974: 19080))

Im Wissenschaftskorpus überwiegt hingegen der Autorenplural:

(12) „Bei unserer Arbeit haben wir uns systematisch darum bemüht, solche Stellen zu lokalisieren und nach Möglichkeit minimal invasiv zu kurieren. Im Einzelnen haben wir dabei Ziele in folgenden Bereichen verfolgt:“ (19, GK, *German Life and Letters*, Sitta 2005)

Graefen/Thielmann (2007) weisen darauf hin, dass die Verwendung der *Wir*-Form eine stärkere Einbeziehung des Rezipienten ermöglicht, was sich besonders im Rahmen von Textkommentierungen anbietet (2007: 95):

(13) „Bisher haben wir uns mit der Umstellung der Kommunikation von Gesten auf Sprache befaßt und die Frage nach den Bedingungen einer bedeutungsidentischen Verwendung von Symbolen behandelt [...]“ (21, DWDS, Wiss, Habermas, 1981: 44)

Innerhalb des wissenschaftssprachlichen Korpus fand sich lediglich ein Kontextbeleg, in dem die *Wir*-Form innerhalb der direkten Redewiedergabe verwendet wurde.

Innerhalb der Kategorie „etwas“ wurden bei fünf der analysierten Kollokationen signifikante korpusspezifische Differenzen hinsichtlich der Klassenzugehörigkeit des Subjekts festgestellt. Aufgrund der großen Anzahl und Vielfalt der Verknüpfungspartnerklassen ist jedoch eine objektive Zuordnung kaum möglich, sodass die Auszählungsergebnisse lediglich als Indikatoren für mögliche Gebrauchstendenzen gewertet werden können.

Funktion der *Wir*-Form	Pressekorpus	Wissenschaftskorpus
Gesamtanteil[a]	23 (8,12%)	11 (4,62%)
direkte Redewiedergabe	23 (100%)	1 (9,09%)
Autorenplural	0	10 (90,91%)

Tabelle 8: Funktion der *Wir*-Form in Subjektposition.

a Die Prozentangaben beziehen sich auf den Gesamtanteil der *Wir*-Form innerhalb der Kategorie „jemand" unter Berücksichtigung aller analysierten pressesprachlichen Kontextbelege.

Signifikante korpusspezifische Abweichungen innerhalb der Kategorie „etwas" betreffen bei den Kollokationen *Anspruch + erheben* und *Beitrag + leisten* die Vorkommenshäufigkeit der Klasse „Wissenschaftszweig, Begriffe und Theorien", die ausschließlich im wissenschaftssprachlichen Korpus vorgefunden wurde. Zusätzlich betreffen die korpusspezifischen Unterschiede bei der Kollokation *Anspruch + erheben* auch die Klasse „Schriftstücke", die im Wissenschaftskorpus deutlich überwiegt. Bei der Kollokation *Möglichkeit + bieten* dominiert diese Klasse wiederum im Pressekorpus – allerdings ohne ein Signifikanzniveau zu erreichen, das den Rückschluss auf korpusspezifische Gebrauchstendenzen gestatten würde. Im Wissenschaftskorpus überwiegen hingegen bei dieser Verbindung die Klassen „Technik/Verfahren", „Prozess" sowie „Zusammensetzung/Eigenschaft" gegenüber ihrer Vorkommenshäufigkeit im Pressekorpus. Die Kollokation *Ziel + verfolgen* zeigt wiederum signifikante korpusspezifische Unterschiede hinsichtlich der Frequenz der Klasse „Handlungen" in Subjektposition. Diese dominiert im Pressekorpus deutlich, was jedoch auch in Zusammenhang mit der generellen Vorkommenshäufigkeit der Kategorie „etwas" bei dieser Verbindung steht.

Die zahlenmäßige Verteilung der einzelnen Verknüpfungspartnerklassen gestaltet sich sehr vielfältig und unterscheidet sich von Kollokation zu Kollokation teilweise erheblich. Kollokationsübergreifende Gebrauchstendenzen lassen sich daher kaum ablesen.

3 Zusammenfassung und Ausblick

Anhand der Ergebnisauswahl konnte gezeigt werden, dass Kollokationen in ihrem Gebrauch in unterschiedlichen Verwendungskontexten teilweise erheblich voneinander abweichen. Die Unterschiede sind auf verschiedenen linguistischen Beschreibungsebenen angesiedelt, betreffen unterschiedliche Merkmalsklassen und differieren auch in ihrer Ausprägung (vgl. Tabelle 9 auf der nächsten Seite).

In Wörterbüchern finden derartige Hinweise zum Gebrauch bislang keine Erwähnung – was nicht zuletzt auf Druckraumbeschränkungen und den verhältnismäßig langen Entstehungszeitraum von Wörterbüchern zurückzuführen ist. Zudem basieren korpusbasierte lexikographische Informationsquellen (wie bspw. OWID oder das Wortschatz Portal der Universität Leipzig) zumeist auf ausgewogenen oder überwiegend aus Presse- und Ge-

Kollokation	Strukturformel	Numerus des Nomens	Artikelwahl zum Nomen	Adjektiv. Attribut zum Nomen	Tempus	Modus	Genus	Modalverbgebrauch	Subjektklasse
Anspruch + erheben	***	***	***	-	-	***	-	-	***
Auffassung + vertreten	***	-	-	***	-	-	*	-	***
Beitrag + leisten	-	-	-	-	*	***	-	**	***
Frage + behandeln	**	-	-	-	***	***	-	-	***
Möglichkeit + bieten	**	*	**	-	-	***	***	-	***
Versuch + unternehmen	***	-	-	**	-	***	**	-	***
Ziel + verfolgen	**	*	*	**	-	***	**	**	***

Tabelle 9: Merkmalsklassen und Signifikanzniveau korpusspezifischer Unterschiede.

brauchstexten bestehenden Korpora, was einer sprachbereichsspezifischen Beschreibung entgegensteht.

In Anbetracht der teilweise hoch signifikanten Unterschiede hinsichtlich der Gebrauchsspezifika von Kollokationen besitzt eine sprachbereichsspezifische Beschreibung jedoch durchaus Relevanz für die Lexikographie und insbesondere für die Lernerlexikographie. So gewinnt gerade vor dem Hintergrund einer zunehmenden Mobilität von Studierenden und Wissenschaftlern die Frage nach den sprachlichen Fähigkeiten, die bei einem Studien- oder Forschungsaufenthalt im Ausland vorausgesetzt werden, zunehmend an Bedeutung. Eine hohe produktive Kompetenz zählt dabei zu den grundlegenden Erfordernissen im Wissenschaftsbetrieb und ist für ein erfolgreiches wissenschaftliches Handeln von eminenter Wichtigkeit. Demnach wäre dem Nichtmuttersprachler mit einer wissenschaftsspezifischen lexikographischen Kodifizierung von Kollokationen (und selbstverständlich auch anderer lexikalischer Einheiten) eine große Hilfestellung insbesondere für die Produktion wissenschaftlicher Texte in die Hand gegeben. Die Realisierung eines solchen lexikographischen Nachschlagewerks bzw. Informationssystems erfordert jedoch zunächst die Bereitstellung eines größeren wissenschaftssprachlichen Korpus, welches die Analyse einer größeren Bandbreite von Wortverbindungen gestattet. Überdies bedarf es der Entwicklung spezieller Recherchewerkzeuge, die eine automatisierte Datenerhebung ermöglichen und somit die hier zugrunde gelegten manuellen Analysen weitgehend ersetzen.

Literaturverzeichnis

Auer, P./Baßler, H. (Hrsg.) (2007): *Reden und Schreiben in der Wissenschaft.* Frankfurt a. M. [u.a.].

Bahns, J. (1997): *Kollokationen und Wortschatzarbeit im Englischunterricht.* Tübingen.

Börner, W./Vogel, K. (1994): „Mentales Lexikon und Lernersprache". In: Börner; Vogel (Hrsg.) (1994); 1–17.

Börner, W./Vogel, K. (Hrsg.) (1994): *Kognitive Linguistik und Fremdsprachenerwerb. Das mentale Lexikon.* Tübingen.

Bortz, J./Döring, N. ([4]2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Heidelberg.

Caro Cedillo, A. (2004): *Fachsprachliche Kollokationen. Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch.* Tübingen.

Ehlich, K. (1993): „Deutsch als Fremde Wissenschaftssprache". In: *Jahrbuch DaF*; 13-42.

Ehlich, K. (1999): „Alltägliche Wissenschaftssprache". In: *Info DaF 26/1*; 3-24.

Ehlich, K. (2000): „Deutsch als Wissenschaftssprache für das 21. Jahrhundert". In: *German as a Foreign Language 1/2000. http://www.gfl-journal.de/1-2000/ehlich.html* [05.06.08].

Ehlich, K. (Hrsg.) (2002): *Mehrsprachige Wissenschaft – europäische Perspektiven. Eine Konferenz im Europäischen Jahr der Sprachen.* München. *http://www.euro-sprachenjahr.de/onlinepub.htm* [Stand 20.09.09].

Fandrych, C. (2002): *„Herausarbeiten* vs. *illustrate*: Kontraste bei der Versprachlichung von Sprechhandlungen in der englischen und deutschen Wissenschaftssprache". In: Ehlich (Hrsg.) (2002); *http://www.euro-sprachenjahr.de/Fandrych.pdf* [Stand 20.09.09].

Graefen, G. (1997): „Wissenschaftssprache – ein Thema für den Deutsch-als-Fremdsprache-Unterricht?". In: *Materialien Deutsch als Fremdsprache 43*; 31-44.

Graefen, G. (1999): „Wie formuliert man wissenschaftlich?". In: *Materialien Deutsch als Fremdsprache 52*; 222-239.

Graefen, G. (2004): „Aufbau idiomatischer Kenntnisse in der Wissenschaftssprache". In: *Materialien Deutsch als Fremdsprache 73*; 293-309.

Graefen, G./Thielmann, W. (2007): „Der wissenschaftliche Artikel". In: Auer/Baßler (Hrsg.) (2007); 67-97.

Hallsteinsdóttir, E. (2007): „Wörtliche, freie und phraseologische Bedeutung. Eine korpusbasierte Untersuchung des Vorkommens von freien und phraseologischen Lesarten bei deutschen Idiomen". In: Kržišnik/Eismann (Hrsg.) (2007); 107-121.

Hausmann, F. J. (1984): „Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen". In: *Praxis des neusprachlichen Unterrichts 31*; 395-406.

Heine, A. (2006): *Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-)Lexikografie.* Frankfurt a. M. [u. a.].

Heine, A. (2008): *Funktionsverbgefüge richtig verstehen und verwenden. Ein korpusbasierter Leitfaden unter Angabe der finnischen Äquivalente.* Frankfurt a. M. [u. a.].

Heine, A. (2009): „Möglichkeiten und Grenzen der Korpusanalyse für die Lexikografie am Beispiel eines Wörterbuches deutscher Funktionsverbgefüge mit finnischen Äquivalenten." In: Mellado Blanco (Hrsg.) (2009).

Heine, A./Hennig, M./Tschirner, E. (Hrsg.) (2005): *Deutsch als Fremdsprache. Konturen und Perspektiven eines Faches. Festschrift für Barbara Wotjak zum 65. Geburtstag.* München.

Helbig, G./Buscha, J. ([19]2001): *Deutsche Grammatik – Ein Handbuch für den Ausländerunterricht.* Leipzig.

Hollós, Z. (2004): *Lernerlexikographie: Syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch.* Tübingen.

Jones, R./Tschirner, E. (2006): *A Frequency dictionary of German: Core vocabulary for learners.* London.
Kržišnik, E./Eismann, W. (Hrsg.) (2007): *Phraseologie in der Sprachwissenschaft und anderen Disziplinen. Akten der Europhras-Tagung in Strunjan/Slowenien, 19.-22. September 2005*. Ljubljana.
Ludewig, P. (2005): *Korpusbasiertes Kollokationslernen. Computer-Assisted Language Teaching als prototypisches Anwendungsszenario der Computerlinguistik.* Frankfurt a. M.
Mellado Blanco, C. (Hrsg.) (2009): *Theorie und Praxis der idiomatischen Wörterbücher.* Tübingen.
Moll, M. (2004): „Deutsch als fremde Wissenschaftssprache "für Fortgeschrittene" – am Beispiel des Linguistischen Internationalen Promotionsprogramms LIPP". In: Wolff et al. (Hrsg.) (2004); 349-470.
Reder, A. (2006): *Kollokationen in der Wortschatzarbeit*. Wien.
Ritz, J./Heid, U. (2006): „Extraction tools for collocations and their morphosyntactic specificities". In: *Proceedings of the Linguistic Resources and Evaluation Conference, LREC-2006 LREC Genova, Italia* [CD-ROM].
Siepmann, D. (2004): „Kollokationen und Fremdsprachenlernen. Imitation und Kreation, Figur und Hintergrund". In: *Praxis Fremdsprachenunterricht 1; 2004/2*; 107–113.
Sternkopf, J. (1998): „Kollokationen in wissenschaftlichen Rezensionen". In: *Beiträge zur Fremdsprachenvermittlung 33*; 32-44.
Tschirner, E. (2005): „Korpora, Häufigkeitslisten, Wortschatzerwerb". In: Heine et al. (Hrsg.) (2005); 133–152.
Wermke, M./Kunkel-Razum, K./Scholze-Stubenrecht, W. (Hrsg.) (2005): *Die Grammatik. Duden Bd. 4.* Mannheim/Leipzig/Wien.
Wolff, A./Ostermann, T./Chlosta, C. (Hrsg.) (2004): *Integration durch Sprache*. Regensburg.

Franziska Wallner
Institut für Auslandsgermanistik/DaF/DaZ
Friedrich-Schiller-Universität Jena
Ernst-Abbe-Platz 8
07743 Jena
Deutschland
franziska.wallner@uni-jena.de

"…you are quoting Shakespeare": Quotations in Practice

Sixta Quassdorf/Annelies Häcki Buhofer

> If you cannot understand my argument, and declare „It's Greek to me", you are quoting Shakespeare.
>
> *(Bernard Levin)*

Die Aussage, dass Shakespeare einen enormen Einfluss auf die Kultur und Sprache in Europa ausübe, grenzt an eine Binsenweisheit: Man scheint sich einig darüber, dass nicht nur Motive, sondern auch Worte und Wendungen aus seinen Werken wiederholt in andere Kontexte eingearbeitet und auch durch diese mittelbar tradiert werden. Erstaunlicherweise jedoch gibt es kaum systematische, empirisch ausgerichtete Studien zu diesem Phänomen. Die Fragen des *wo, wann, warum* und *von wem* Shakespeare *wie* zitiert wird, sind – abgesehen von vereinzelten Studien zu literarischen Werken – bisher weitgehend unbeantwortet. Der vorliegende Aufsatz versucht sich dieser Aufgabe ansatzweise zu stellen: Es werden drei Zeilen aus Shakespeares *Hamlet* in ihrer Verwendungsweise näher untersucht. Dabei zeigt sich, dass sowohl das historische Vorkommen, als auch die thematischen und textfunktionalen Verwendungen jeweils divergieren. Die Untersuchung des sprachlichen Einflusses Shakespeares darf also nicht generell gefasst, sondern muss vielmehr differenziert für einzelne sprachliche Bereiche und Perioden untersucht werden. Darüber hinaus weisen die von *Hamlet* abgeleiteten Phraseologismen zwar tendenziell die typischen Merkmale historischer Wandelprozesse auf, wie z. B. Verkürzungen und die Auflösung von Festigkeit, andererseits sind diese Prozesse aber auch immer reversibel durch die Präsenz der literarischen Quelle. Diese Möglichkeit der Reversion scheint einen wesentlichen Unterschied zwischen „Zitaten in der Praxis" und anonymen Mehrwortverbindungen auszumachen.

1 Empirical Linguistics and the "Shakespeare Phenomenon"

The little sketch by *Times* columnist Bernard Levin, from which we choose our motto, nicely echoes the mantra-like quality with which Shakespeare's influence on the English language is often pronounced: it embeds more than 40 Shakespearean quotations into the

reiterated structure of "if [you say so-and-so] you are quoting Shakespeare" (Levin 1986: 98-99). Empirically verified studies backing these claims are, however, hard to find. The questions of *what, when, where, why* and *how* a phrase from a poetic work leaves its original context and is taken up in a more or less comparable new setting, call for linguistic investigation. As a result, the *Shakespeare phenomenon* makes the sister disciplines of literature and linguistics draw nearer again: *historical linguistic, corpus linguistic, discourse analytical, pragmatic, semantic* and *syntactic* approaches can complement the *intertextual, structuralist* and *reception theoretical* traditions of literary and cultural studies.

Not surprisingly then, the study of quotations comprises a number of challenges. In view of theory formation, the challenge consists of integrating several, mostly independently working research fields with one another. Language philosophical and cognitive models of understanding and meaning transfer have to be linked to the question of uptake and re-application, and tested for their suitability. Sociologically motivated theories of language change need to be complemented by cognitive models of acquisition. The ontogeny of phraseologisms has to be understood in combination with learning processes; and the question of whether or not the genesis and history of a multi-word unit parallels that of single words is as yet unanswered. In terms of *genesis*, we also need to take into account the possibility of chance parallel developments.

In view of methodology, Shakespearean phrases and expressions have to be traced in both their historical and present-day use and application. References to *Hamlet* occur in a great variety: they may be verbal, visual or acoustic, allude to phrases, characters, motifs or the complete plot, and accordingly range from single words to book-length adaptations. Although the linguistic study of the *Shakespeare phenomenon* focuses on so-called *line references*,[1] i. e. syntagmatic chains which are taken from *Hamlet* and reapplied in new contexts as "quotation in practice", the heterogeneity of the material is principally unbounded. Consequently, a thorough and differentiated empirical description is required along a multiplicity of dimensions, such as form, function, domain of application and historical period, to name only a few.[2]

2 On Form and Frequency

The present study focuses on a description of three selected line references from *Hamlet*: "A little more than kin, and less than kind" (I, ii, line 65), "It is a custom more honour'd in the breach than the observance" (I, iv, lines 15-16) and "For 'tis sport to have the engineer hoist with his own petard" (III, iv, lines 206-207). Syntagms can typically be re-applied by either verbatim rendition, shortening as in (1), lexical substitution as in (2), word order permutation and/or additions as in (3):

1 The term *line reference* is a concession to the tradition of indicating dramatic passages by line numbers, i. e. for ease of reference retrieval.

2 See Hohl Trillini/Quassdorf (2008b) for a detailed exposition of descriptive parameters, which are also suitable for the annotation of a corpus of quotations from and allusions to a literary work.

(1) More Kin Than Kind. (Fitzstephen 1903)

(2) More than a hunch, and less than a certainty. (Barash 2006)

(3) I believe the rule itself to be one of universal application, always honoured in the observance, if not always equally dishonoured in the breach. (Conington 1882)

Verbatim quotations can, of course, be easily found through a computerised search. However, they are not very frequent unless they consist of rather short units such as "primrose path" (I, iii, line 50) or "sea of troubles" (III, i, line 61). Variants of longer phrases, such as the chosen ones, are comparatively more difficult to trace. To capture maximum variety, we formed several search strings per line according to combinatorial criteria and tested them with the database *Literature Online* (LION).[3] We systematically combined the lexical word stems regardless of syntactic structure, chose specific line fragments, and also tested the syntactic schema consisting partly or wholly of functional words. Occasionally, very common synonyms or variants were included in the searches. The phrase "a custom more honour'd in the breach than the observance" yields the exemplary search strings in table 1.

Logical combinations	**Fragments**	**Schemas**
custom NEAR hono*[4]	it is a FBY.3 more hono*	it is a FBY more FBY in the
custom NEAR breach*	more hono* in the	more FBY in the FBY than the
custom NEAR obser*	more hono* FBY than in	more FBY in the FBY than in
hono* NEAR breach*	in the breach than	it is a FBY in the FBY than the
hono* NEAR obser*	than FBY.3 obser*	it is a FBY in the FBY than in
breach* NEAR obser*	than the obser*	

Table 1: Exemplary selection of search strings for *it is a custom more honour'd in the breach than the observance*.

The test run with LION revealed that the rarer a keyword or the more unusual the structure of a phrase, the more formal variety can be covered with only very few search strings. Consequently, further searches for *honour'd breach* and *hoist petard* could be narrowed down to *hono?r* NEAR.4 breach**, or *in the breach than* and *own NEAR.3 petard?*. The search for "a little more than kin, and less than kind", however, had to remain close to the original because both the structure and the keywords are very common in

3 LION contains about 350,000 literary works in English from the 14th to the early 20th c.

4 Initially, we created even more search strings defining different word spans and word orders in case some searches would be too open, i. e. yield too large quantities of data (which turned out to be rarely the case). The search strings are indicated according to the rules of LION: NEAR standing for *Range* regardless of word order, FBY for *Range* observing word order, numbers indicate the specific word span, * stands for any number of additional letters, ? for one letter. In order to capture spelling variants, stems were often shortened. Most of the other databases offer comparable search options, but the actual code may differ.

English. The variants *more than kin, less than kin, little more than kind* and *less than kind* proved most effective for the searches through the other corpora used in this study, i. e. the *British National Corpus* (BNC), the *Eighteen's Century Collections Online* (ECCO), the *House of Commons Parliamentary Papers 1801-2004* (HCPP),[5] and the *Times Digital Archive 1785-1985* (TDA). Moreover, the *World Wide Web* was searched via *WebCorp*, which is a combination of a concordancer and a search machine accessing 100 random sites per query. The specialised corpus of Shakespearean quotations, *HyperHamlet* (HYHA),[6] has also been used to answer the question of which phrase occurs when and where.

Computerised searches necessarily yield incomplete results because there is no limit to word play and ad hoc modification (cf. also Burger et al. 1982: 68f., and Sternberg 1982: 109ff.). The case that the cotext provides unexpected variation such as (2), which is taken from a review about A *little more than kin and less than kind – The Evolution of Family Conflict* by Mock (2004), is not very frequent. What corpora and computer searches can show, however, is "which forms and variants are more commonly found" (Moon 2007: 1054). These "common" variants are indicative of the integration of quotations into the general linguistic reservoir of ready-made syntagms: following general patterns of language change, fixedness and length should decrease over time with the dissociation from the quotational source and the increase of phraseological usage. The line "'tis sport to have the engineer hoist with his own petard" is a case in point: the "engineer" is mentioned only in very early quotations; generally the clause is shortened to a verb phrase and the preposition *by* may be used instead of *with*.

It has repeatedly been noted that even in very large general corpora the actual frequency of specific proverbs, idioms and metaphoric collocations is often very low, if documented at all (cf. Colson 2003, Wray 2002, Moon 2007, Gries 2008). For example, the familiar expression *the steep and thorny way to heaven* (*Hamlet* I iii) or the *kin/kind* contrast discussed in this paper do not occur even once in the BNC. Against this background, an absolute frequency of 12 incidents of *honour'd breach* and 17 of *hoist petard* in the BNC, or 34 occurrences of the *kin/kind* line in LION is noteworthy. By taking into account text types and textual function, the results can be fine-tuned and gain in information content.

---

5 The BNC is a 100 Million words corpus of written and spoken British English from the later decades of the 20th c. ECCO comprises 150,000 fictional and non-fictional works published in the UK between 1701 and 1800. HCPP is a collection of more than 180,000 printed records of the British government from 1801 to 2004.

6 HYHA – a transdisciplinary project supported by the Swiss National Science Foundation – is a specialised corpus of quotations from and allusions to Shakespeare's *Hamlet* which has been built at the University of Basel since 2002. To date the corpus contains about 6,000 *Hamlet* references from fiction and non-fiction, English and non-English contexts and several centuries. The reader is invited to contribute to the database via the website *www.HyperHamlet.unibas.ch*. All contributions are edited by the project staff to guarantee scholarly standard.

3 On Proverbs and Paradoxes

Shakespeare made ample use of rhetorical figures and tropes such as metaphor, alliteration, antithesis, paronomasia or paradox to name only a few. Strengthened expressivity, conciseness and euphony make rhetorically worked up phrases conspicuous and raise their memorability as well as the potential for re-application. The abundance of rhetorical figures found in our data may therefore be attributed to both the original purpose of public performance and the conditions of uptake.

3.1 A little more than kin and less than kind

Rhetoric does not only characterize poetic language but also a large number of proverbs, sayings, idioms and collocations. In a monograph on rhetoric in German phraseology, Dietz (1999) showed that rhetoric is much used in many types of conventionalised lexical groups. The rhetoric of figures and tropes is possibly a natural linguistic phenomenon which is elaborated on rather than created by poets and orators. A case in point is the proverb "the nearer in kin the less in kindness" (cf. Tilley 1950: K38) with its *k* alliteration and parallel structure. The proverb conceptually contains a paradox in the literal sense of the word.[7] It defies the accepted rule of "blood is thicker than water" by suggesting that *kin* and *kindness* form an opposition rather than a combination. The paradox is echoed in the rhetorical figure of antithesis by the *the nearer/the less* contrast.

Shakespeare, however, increases the rhetorical finesse: Hamlet's first words "A little more than kin and less than kind" create paranomasia by the reduction of *kindness* to *kind*, add the alliteration *little/less*, strengthen the antithesis by the more explicit opposites of *more/less* as well as the parallel structure with the repetition of the heavier *than*, and improve the rhythm by creating a regular iambic pentameter. It is these rhetorical gimmicks which distinguish the Shakespearean from the original proverbial trace:[8]

(4) She was certainly a little less than kin to Miss Pemberton, and she was as certainly a little more than kind to people who paid her attention on board ship. (Anon. 1870).

(5) A little less than kin, and more than kind. (Lawrence 1924).

(6) A little more than SimpleCartItem, and less than 'full-on' e-commerce. (Stal 2006).

As already mentioned above, there is no reference to this line in the BNC and very few in ECCO and HCPP. The two incidents in HCPP do not occur within a text passage but are part of examination tasks for graduating students. HYHA on the other hand, documents 37 instances, the earliest reference dates from 1787 and the newest one from 2008. 23

7 *para* = besides, against; *doxa* = received opinion, the usually expected. Thus we use the term *paradox* in the following sense: a paradox is a well-founded statement which contradicts received opinion or commonsensical notions yet does not cause logical problems of any kind. The philosophical sense of *paradox* is more specific.

8 The question of data validity is not discussed here: the chosen examples are rather straightforwardly attributtable to Shakespeare due to their length and rhetorical make-up. See Quassdorf (2009) for a discussion of the complexity of data selection.

database	#	period	marking	domain	text function
BNC	0	-	-	-	-
ECCO	1	1751	marked	literature	footnote
HYHA	37	1787–2008	8 marked	arts & culture	23 titles
LION	34	1756–1936	23 marked	literature	body of text
HCPP	2	1861 + 1895	marked	examination tasks	body of text
TDA	45	1827–1985	8 marked	arts & culture, society life	8 titles
WebCorp	51	late 20th c. onwards	4 marked	arts & culture	24 titles

Table 2: Search results *more than kin / less than kin / less than kind / little more than kind*.

entries refer to book titles. 8 references are marked for quotation.[9] A search through LION yields 34 occurrences covering a period from 1756 to 1936. 23 of these occurrences are explicitly marked as quotations. Explicit marking comprises typographic signals, e. g. inverted commas as in (7), and metalinguistic embedding such as "to use the words of a poet" in (8).[10]

(7) "The dark Death," said Hamlet, "'a little more than kin and less than kind!'" (Aldrich 1862).

(8) I think, my good sister, we have been all our lives a little more than kin and less than kind, to use the words of a poet whom your dear father loved dearly. (Thackeray 1859).

The search via WebCorp on 15 November 2008 prompted 51 (out of 100) different references to that line outside a Shakespearean context. The fact that some 40 per cent of the occurrences stem from *Hamlet*-related pages shows that the link between the line and its original context is still relatively strong. Outside that original context, culture and the arts are the preferred areas of application with only very few exceptions such as (6). A high frequency of paratextual usage, such as titles of books, songs, poems, films and TV series, as well as the onomastic function for music groups or websites is noticeable. The 45 data sets from TDA show a comparable distribution.

9 Frequencies in HYHA have to be evaluated with caution: systematicity is guaranteed in some subdomains such as quotations from and allusions to *Hamlet* in the works of Dickens, Byron, Scott, Raabe, Th. Mann and other famous writers. Other domains have been assembled randomly, i. e. the advantages and disadvantes of opportunist corpora have to be taken into account.

10 For a detailed exposition of the various marking strategies see Hohl Trillini/Quassdorf (2008a) and Quassdorf (2009).

The line "A little more than kin, and less than kind" is primarily applied in cultural settings and in the domain of private or personal affairs. Possibly, the semantics of the keywords influence the predominant occurrence in the more emotional areas. As soon as the keywords are substituted as in (6), the thematic domain can broaden. Data from earlier periods tend to be marked by quotation marks, whereas marking decreases over time. Furthermore, the elaborated rhetoric may have a share in the documented application for names and titles. Titles, however, tend to be shortened versions of the line. The combination *than kin* is apparently the most conspicuous constituent: titles and names such as *More than kin* and *Less than kin* abound. The psycholinguistic question of whether or not the phrase is felt as a whole and usually mentally completed has to remain unanswered for the time being.

3.2 'Tis sport to have the engineer hoist with his own petard

The human hope that vice may turn against itself is the *conceptual base* (Honeck 1997: 131ff.) of Shakespeare's "'tis sport to have the engineer hoist with his own petard." Although the underlying idea pre-dates Shakespeare and several earlier variants are known such as

(9) Behold, he travaileth with iniquity, and hath conceived mischief, and brought forth falsehood. He made a pit, and digged it, and is fallen into the ditch which he made. His mischief shall return upon his own head, and his violent dealing shall come down upon his own pate. (King James Bible, Psalm 7, 14-16)

(10) The fowler is caught in his own net. (Tilley 1950: F626)

Shakespeare is assumed to have introduced the warfare metaphor himself (cf. OED, Shakespeare 1994: 361 and Dent 1981: P243.1). Apart from the vivid image, no further rhetorical gimmicks are applied. Nevertheless, the derived verb phrase has gained so much currency that dictionaries like the OED and LEO list it as idiomatic expression.

Interestingly, the proverb (10) as well as the Shakespearean version leave the directive perlocution as expressed in (9) implicit. The state of affairs is pictured from the perspective of the person breaking the norm: as a rule, the malefactor occupies the subject position in a passive clause. Thus, the expression turns into a paradox: the same person is both agent and patient of the same event. Only rarely, active or quasi-active constructions are found:

(11) "Greenberg also warns the Unix contingent not to try to hoist NT on same [sic] petard Microsoft is using against Unix." (Anon. 1993)

The BNC documents 17 instances of the idiom, five of which are marked for quotation. Two occurrences are found in ECCO dating from 1752 and 1797. References collected in HYHA cover the period 1824–2007; the majority stems from non-fictional texts. Nine references are explicitly marked as quotations. LION yields nine incidents, which all but one date from the 19th century. Four references are marked. 57 occurrences are found in HCPP dating from 1845 to 2004 with only occasional marking. The WebCorp search from 7 November 2008 resulted in 92 valid hits outside the context of the drama. They are predominantly headlines of either newspaper articles or blog contributions on politics,

database	#	period	marking	domain	text function
BNC	17	late 20th c.	5 marked	politics, law, economics	body of text
ECCO	2	1752 + 1797	Marked	politics	body of text
HYHA	31	1824–2007	9 marked	10 fiction / 21 non-fiction	body of text
LION	9	19th c.	4 marked	literature	body of text
HCPP	57	1845–1919/1969–2004	not marked	politics, law, economics	body of text
TDA	317	1834–1985	Occasional	politics, law, economics, sports, arts	6 titles
WebCorp	92	late 20th c. onwards	34 marked	politics, law, economics	54 titles

Table 3: Search results *own petard*.

law and economics. 25 websites discuss the meaning of the idiom. TDA prompts 317 hits from 1834 onwards. Although the expression mostly occurs in articles about politics, law and economics, the *Times* also documents occurrences in sports reports and art criticism. Nevertheless, a general predominance of replication in a specific domain is again obvious. In contrast to the cultural domain of the *kin/kind* line, *hoist petard* is preferably used in the "domain world affairs."[11] We do assume then that the application of quotations in a limited variety of genres is not mere coincidence but that it may partly derive from some inherent property of the quotation. On the other hand, this noticeable tendency of restricted usage reveals the fact that the study of the *Shakespeare phenomenon* cannot be conducted on an abstract or generalised linguistic level, but has to be differentiated according to domain and specific communicative setting.

The high ratio of Shakespeare-unrelated hits in the WWW, the frequency in TDA, BNC and HCPP let us suggest that *to be hoist with one's own petard* has been widely used as an independent idiomatic verb phrase since the mid-19th century. The preferential occurrence in the field of world affairs may be due to the semantics and pragmatics of the phrase: public affairs have to deal with norms in society; as such the norm-enforcing moral of the quotation is thematically linked. Secondly, the expression is an instance of Lakoff's conceptual metaphor *argument is war* (Lakoff/Johnson 1992: 4 ff.) and fits well with both the domain and its debating style. Thirdly, due to the passive construction, actual agency can stay hidden, which leads to ambiguity in agency: the potential for purposive

11 The term is taken from the BNC.

indirectness corresponds to the communicative style in politics and news coverage.[12] In contrastive confirmation, LION provides only very few results from the literary domain.

The discussions in the web about the meaning of the phrase provide further evidence of its phraseological status: the expression is familiar, yet the metaphoric base, Elizabethan warfare, needs explanation. The idiom has become semi-transparent, if not non-transparent for a number of language users. These discussions exemplify that a literary quotation can always be returned to its source as long as the literary source is read. In other words, in contrast to most proverbs, sayings and idioms, the transition from a quotation to an anonymous institutionalised part of the *langue* can possibly not be fixed as it is subject of the dynamics of general cultural developments. The ratio of about 30% of marked occurrences supports our assumption that frequently used literary quotations tend towards a "mongrel" status between quotation and idiom. On the other hand, the lack of marking in HCPP and the *Times* may be due to a general expectation that parliamentarians and educated readers know Shakespeare's plays. In this case, marking for quotation may simply be felt superfluous. Consequently, the lack of marking is a sufficient but not a necessary indication of unintentional quoting, i.e. the usage of a former quotation as a mere phraseological item learned from other contexts.

The familiarity, on the one hand, and the semi-transparent metaphorical apparel, on the other hand, may be one of the reasons why *hoist petard* frequently functions as a newspaper heading: headline writers may take the phrase as a perfect hook which balances catchiness with opacity to make people read on (cf. Lennon 2004: 78 ff.). However, we should mention that the media do not yield a unified picture: the quality paper *Times* only counts 6 instances of headlines since 1916. A future, more fine-grained analysis has to differentiate between popular and quality papers, print and digital media, as well as between the more specific text genres and discourse functions of the syntagm.

Lastly, another aspect should be addressed, which the HCCP data reveal: between 1919 and 1969, i. e. over a period of 50 years, no single instance of *own petard* is documented in the governmental papers. Possibly this interruption is another distinguishing characteristics of (famous) literary quotations: as the source text can always be re-read, the usefulness of a quotation can always be re-discovered. Whatever the reasons for this gap may be – it indicates that "quotation in practice" may not only be domain specific, but also time specific and dynamic – their usage ultimately depends on extralinguistic social factors (and fashions).

3.3 It is a custom more honour'd in the breach than the observance

Also the third line chosen for this study exemplifies Shakespeare's strategy of using proverbial wisdom for his works: the proverb "A bad custom is like a good cake, better broken than kept" is the assumed source (Tilley 1950: C931, Shakespeare 1994: 181). Again, a paradoxical stance is implied: a tradition is not to be kept but to be broken, which contradicts the usual function of customs and norms. If *to be hoist with one's own petard* is

12 Note that it is actually *Hamlet* who "hoists the petard" in the original text; true agency is put out of focus and thus "manipulated away" by the passive construction.

a norm-consolidating maxim and moralises against the violation of rules, *a custom more honour'd in the breach* suspends this moral. Principally, the line can either serve to lament the violation of norms, or to pave the way for a change: Shakespeare explicitly evoked the latter attitude by the positive undertones of *to honour*.

As for the rhetorical reconfiguration of the proverbial model, Shakespeare's strategy is again different this time: he actually removes catchiness. While preserving the antithesis, he cuts the simile *like a good cake* as well as the explicit value judgement *bad/good*. Thus the phrase is bereft of its humorous down-to-earth tinge. Instead, Shakespeare uses a complex conceptual metaphor: the abstract noun *custom* is conceptualised as both an object which can be broken and a person who can be honoured. The direct combination of the positively connoted *honour* with the negatively connoted *breach* also increases cognitive complexity. Thus the expression gains in sophistication. The data demonstrate that the phrase is predominantly used in expert settings, i. e. it is used by the "learned audience".

The antithesis is still at the heart of the expression, which makes for a remarkably long and conspicuous quotation in the shape of a participial phrase or a post-modified noun phrase:

(12) Unfortunately, even in structuralist work (e. g., recent theories of narrative) this teleological logic – so inimical to the typologist's instinct for orderly divisions and fixed pairings and neat symmetries – is more honored by homage than observance. (Sternberg 1982)

(13) However, all serious historians have noted that the 1925 decisions had been merely a formality, more honoured in the breach than in their application. (Tarbuck 1989)

However, implicit antitheses, i. e. shortened versions, are also found, though as yet not very frequently:

(14) But, as in North America, the treaty was sometimes honored more in the breach. (Powell, 2003)

The BNC contains 12 unmarked occurrences. A search through ECCO prompts 75 references with only occasional marking, which decreases over time with the rise in frequency of the expression. HYHA contains 32 references from 1796 onwards. Two references are titles and seven references stem from works of fiction. 23 passages are marked. LION prompts 23 hits, eight of which date from the 18th century already. Five references are marked. HCPP prompts 151 hits starting in 1830 with regular occurrences until 2004 and TDA offers 286 occurrences starting in 1785. In both cases marking decreases over time. The search through WebCorp gives 74 instances outside the Shakespearean context: 13 websites discuss the meaning of the phrase, five occurrences are titles.

Honoured breach is not only the earliest and best documented line of our selection, but also the broadest in thematic application. A certain predominance of non-fictional political and economic texts is noticeable, but the domains ultimately vary from law to linguistics, from theology to natural sciences, from social sciences to the arts since the 18th century. Presumably the abstract shape of the expression provides the possibility of condition for

database	#	period	marking	domain	text function
BNC	12	late 20th c.	none	non-fiction (expert / academic)	body of text
ECCO	75	1750–1800	occasional	non-fiction (expert / academic)	body of text
HYHA	32	1796–2007	23 marked	7 literature	2 titles
LION	23	1758–1906	5 marked	literature	body of text
HCPP	151	1830–2004	occasional	politics, law, business	body of text
TDA	286	1785–1985	occasional	politics, law, business, culture, sports	body of text
WebCorp	74	late 20th c. onwards	occasional	non-fiction (expert / academic)	5 titles

Table 4: Search results *in the breach than / hono* NEAR.4 breach*

this wide-ranging usage. The fact that the expression is primarily used in expert journalism or scholarly contexts may be founded in its sophisticated make-up.

4 Conclusion

This paper is a linguistic contribution to a larger project-in-progress which aims to trace the *Shakespeare phenomenon* in non-Shakespearean contexts with empirical methods. A selection of three lines from *Hamlet* has been discussed in view of their rhetorical quality, their occurrence in different databases, their usage over time, their genre distribution, as well as their textual functions. Assumptions about correlations between these fields of analysis have been offered.

While *honoured breach* and *hoist petard* most frequently occur in non-fiction, the *kin/kind* contrast is more often found in literature and the cultural domain. The semantic field of the keywords is possibly one of the decisive factors (cf. also Moon 2007). *Hoist petard* and the *kin/kind* lines are popular for titles, while *honoured breach* occurs nearly exclusively within the text flow. The preference for paratextual functions may be a corollary of the rhetoric make-up: the more rhetorically or pictorially conspicuous, the more the expression seems to be appropriate for a heading. However, a differentiation between quality and popular genres is advisable for future studies. Products of popular culture seem to title with the *kin/kind* and *hoist petard* lines more often than high-brow literature or quality papers do. Lastly, the expressions *a custom more honoured in the breach than the observance* and *to be hoist with one's own petard* have already been regularly used

outside a Shakespearean context since the 18th and 19th centuries respectively, noticeably in the *Times* and among British parliamentarians. The *kin/kind* line is less documented, which may also be due to its relation to the private and emotional domains, which are generally less frequently represented in corpora. On the other hand, it has kept the link to its original areas of literature and culture, which may be both cause and corollary of its lower frequency.

We are lacking data from the first centuries of *Hamlet's* "life" – especially the time between 1600 and 1750, which may be primarily due to the "bad data" problem. The role Shakespeare's bicentennial played in terms of pushing the frequency of quotations needs some further exploration. The example of the *hoist petard* line, however, has shown that quotations need not be used continuously: despite considerable gaps in time, they may be re-discovered again. This is presumably a distinctive trait of quotations in contrast to other phraseologisms: references to literary or, at least, written sources can be better identified and re-established than references from oral sources, which integrate more rapidly into the anonymous phrase stock of a language and, consequently, become also more easily subject to linguistic change.

Our analysis of three lines from *Hamlet* seems to indicate that marking for quotation decreases over time and that, subsequently, the notion of quotational *use* or "quotation in practice" increases. According to general historical principles, this should lead to a decrease in fixedness. Although a number of data behave according to this principle, others do not: quotations may always recuperate their fixedness by the source text. On the other hand, intertextual references in literature are mostly praised for their subtleness; as a result, quotations in literary practice are deliberately dissolved and subtly integrated into a new context, which complicates their linguistic status from another angle. Further fine-grained explorations of the pragmatic, semantic and syntactic characteristics of "quotations in practice" are waiting for us to be accomplished.

Bibliography

Primary Literature

Aldrich, Th. B. (1862): *Out of His Head. A Romance.* New York.

Anon (1870): "Riddles of Love", in: *London Society: An Illustrated Magazine of Light and Amusing Literature for the Hours of Relaxation, 18. November*; 390-398.

Anon (1993): "Common Open Software Environment", in: *Computergram International.* (Quoted in: The British National Corpus. Document No. CPA. Oxford University Computing Services. 2007.)

Barash, D. P. (2006): "More than a Hunch, and less than a Certainty: A review of Douglas W. Mock, *More Than Kin and Less Than Kind: The Evolution of Family Conflict*", in: *Evolutionary Psychology 4*; 459-461.

Boulton, J. T. (ed.) (1988): *The Works of D. H. Lawrence.* 33 Vols. Cambridge.

Conington, J. (1882): "Preface", *The Odes and Carmen Saeculare of Horace.* London.

Fitzstephen, G. (1903): *More Kin Than Kind.* London.

Lawrence, D. H. (1924): "On Being a Man". In: Boulton (ed.) (1988); 211-222.

Mock, D. W. (2004): *A little more than kin and less than kind – The Evolution of Family Conflict.* Cambridge, Massachusetts.
Powell, E. A. (2003): "Searching for the First New Zealanders", in: *Archeology. 56.2*; 40-48.
Stal, J. (2006): "Politics, the environment, technology, activism", in: *John Stal's Journal.* 27. Dezember, section "And stuff." *http://blogs.onenw.org/jon/archives/2006/12/27/initial-thoughts-on-a-plone-for-nonprofits-bundle/ss\%20than* (15. November 2008).
Tarbuck, K. J. (1989): *Bukharin's Theory of Equilibrium.* London.
Thackeray, W. M. (1859): "The Virginians: A Tale of the Last Century". In: *The Works of William Makepeace Thackeray.* 13 Vols. London, 1899.

Secondary Literature

Burger, H./Buhofer, A./Sialm, A. (eds.) (1982): *Handbuch der Phraseologie.* Berlin/New York.
Burger H./Dobrovol'skij D./Kühn, P./Norrick, N. R. (eds.) (2007): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin/New York.
Burger, H./Häcki Buhofer, A./Gréciano, G. (eds.) (2003): *Flut von Texten – Vielfalt der Kulturen.* Baltmannsweiler.
Colson, J.-P. (2003): "Corpus Linguistics and Phraseological Statistics: a few Hypotheses and Examples". In: Burger et al. (eds.) (2003); 45–59.
Dent, R. W. (1981): *Shakespeare's Proverbial Language: An Index.* Berkeley.
Dietz, H.-U. (1999): *Rhetorik in der Phraseologie. Zur Bedeutung rhetorischer Stilelemente im idiomatischen Wortschatz des Deutschen.* Tübingen.
Granger, S./Meunier, F. (eds.) (2008): *Phraseology: An Interdisciplinary Perspective.* Amsterdam.
Gries, S. Th. (2008): "Phraseology and linguistic theory: a brief survey". In: Granger/Meunier (eds.) (2008); 3–25.
Hamm, A./Higgs, L. (eds.) (2008): *Variability and Change in Language and Discourse.* Strasbourg.
Hohl Trillini, R./Quassdorf, S. (2008a): "Quotations and their Co(n)Texts". In: Hamm/Higgs (eds.) (2008); 77-89.
Hohl Trillini, R./Quassdorf, S. (2008b): "A 'Key to all Quotations'? A Corpus-Based Parameter Model of Intertextuality". Submitted manuscript.
Honeck, R. P. (1997): *A Proverb in Mind: The Cognitive Science of Proverbial Wit and Wisdom.* Mahwah.
Lakoff, G./Johnson, M. (1992): *Metaphors We Live By.* Chicago/London.
Lennon, P. (2004): *Allusions in the Press: An applied linguistic study.* Berlin.
Levin, B. (1986): "On Quoting Shakespeare". In: McCrum et al. (eds.) (1986); 98-99.
McCrum, R./Cran, W./MacNeil, R. (eds.) (1986): *The Story of English.* New York.
Moon, R. (2007): "Corpus Linguistic Approaches with English Corpora". In: Burger et al (eds.) (2007); 1045–1059.
Quassdorf, S. (2009): "*HyperHamlet* – Intricacies of Data Selection", in: *Linguistik Online 38.2*; 45-55. *www.linguistik-online.org*.
Shakespeare, W. (1994): *Hamlet.* Ed. by Hibbard, G. R. Oxford/New York.
Sternberg, M. (1982): "Proteus in Quotation-Land: Mimesis and the Forms of Reported Discourse", in: *Poetics Today 3.2*; 107-158.
The King James Bible. http://www.studylight.org (5. November 2008).
Tilley, M. P. (1950): *A Dictionary of the Proverbs in England in the 16th and 17th Centuries: A Collection of the Proverbs Found in English Literature and the Dictionaries of the Period.* Ann Arbor.
Wray, A. (2002): *Formulaic Language and the Lexicon.* Cambridge.

Electronic Databases

BNC – The British National Corpus. http://www.natcorp.ox.ac.uk/ (14. February 2009).
ECCO – Eighteen's Century Collections Online. http://www.gale.cengage.com/DigitalCollections/products/ecco/index.htm (14. February 2009).
HCPP – 19th/20th Century House of Commons Parliamentary Papers 1801-2004. http://parlipapers.chadwyck.co.uk (14. February 2009).
HYHA – HyperHamlet. http://www.hyperhamlet.unibas.ch (14. February 2009).
LEO – Online Dictionary. http://www.leo.org (15. February 2009).
LION – Literature Online. http://lion.chadwyck.co.uk (14. February 2009).
OED – The Oxford English Dictionary. Online Edition. *http://dictionary.oed.com* (15. February 2009).
TDA – The Times Digital Archive 1785–1900. http://www.galeuk.com/times (14. February 2009).
WebCorp – http://www.webcorp.org.uk (14. February 2009).

Sixta Quassdorf
Englisches Seminar
Universität Basel
Nadelberg 6
4051 Basel
Schweiz
sixta.quassdorf@unibas.ch

Annelies Häcki Buhofer
Deutsches Seminar
Universität Basel
Nadelberg 4
4051 Basel
Schweiz
annelies.haecki-buhofer@unibas.ch

Verbale und visuelle Formelhaftigkeit: Zwischen Tradition und Innovation

Natalia Filatkina/Ane Kleine/Birgit Ulrike Münch

The article gives insights into four interdisciplinary projects in the field of verbal and visual formulaic patterns. The projects are currently being conducted at the University of Trier (Germany) and are part of the Cultural-Historical Research Centre of Excellency (HKFZ Trier). The common interest of all the projects lies in the field of the history of traditions of verbal and visual communication and its dynamics. Another common point consists in using modern database technology and corpus linguistic methods in order to tackle this question. Nevertheless, the projects are based on different methodology, due to the differences in the current state of scholarly research in the field of formulaic patterns in the Art History and Historical Linguistics as well as to the different disciplinary approaches. After a short overview about the scholarly state of the art, we show how modern database technology is implemented in all four projects. The article is to be understood as a preliminary report about the work in progress aimed at interlinking (electronic systematization) of different types of data within the framework of eHumanities.

1 Einleitung

Im Mittelalter und in der Frühen Neuzeit war Formelhaftigkeit ein wesentliches Element der verbalen und visuellen Kommunikation. Im Bereich der Sprache kam sie vor allem mit Hilfe von syntaktisch, semantisch und pragmatisch mehr oder weniger fest werdenden bzw. gewordenen Wendungen oder Texten zustande. Im visuellen Bereich manifestierte sie sich in ein- oder mehrszenigen Sprichwortbildern unterschiedlichster Gattungen, z. B. auf Gemälden, Tapisserien, in der Druckgrafik, in der Buchmalerei oder im Kunsthandwerk.

Im Gegensatz zu dem in der Phraseologieforschung bevorzugt verwendeten Begriff „Phraseologismus“ sprechen wir im Folgenden mehrheitlich von formelhaften Wendungen und weiten diesen Begriff von rein linguistischen auf kunsthistorische Phänomene aus. Ziel dieses Aufsatzes ist es, den Einsatz der korpus- und computergestützten Verfahren für die interdisziplinäre Erforschung der historischen verbalen und visuellen Formelhaftigkeit zu thematisieren.

Nach der Erläuterung der Forschungslage im Bereich der verbalen und visuellen Formelhaftigkeit (Abschnitt 2) stellen wir vier Projekte vor, in denen der Einsatz der korpus- und computergestützten Verfahren bereits aktiv erprobt wird (Abschnitt 3). Alle vier Projekte

bilden (zusammen mit weiteren Projekten) die Arbeitsgruppe „Wissensraum Kommunikation: Kulturelle Praktiken, Tradition und Wandel" im Historisch-Kulturwissenschaftlichen Forschungszentrum (HKFZ) an der Universität Trier.[1] Sie unterscheiden sich in ihren Ausgangslagen im methodischen, medialen (Bilder und Texte) und technischen Bereich. Es soll dargelegt werden, welche Herausforderungen die unterschiedlichen Vorgehensweisen für den Einsatz der korpus- und computerlinguistischen Methoden bedeuten und wie sie für die Beantwortung gemeinsamer Fragestellungen fruchtbar gemacht werden können. Unsere ersten und vorläufigen Überlegungen fassen wir in Abschnitt (4) zusammen. Sie bilden den Ausgangspunkt einer Vernetzung des gesammelten Materials und haben die Entwicklung einer gemeinsamen interdisziplinären, multimedialen und internetbasierten Forschungsressource zur historischen Formelhaftigkeit zum Ziel.

2 Zum Stand der Erforschung der historischen Formelhaftigkeit in Linguistik und Kunstgeschichte

2.1 Formelhaftigkeit aus der Perspektive der Sprachgeschichte sowie der Korpus- und Computerlinguistik

In der Phraseologieforschung war bis jetzt die synchrone und gegenwartsbezogene Blickrichtung dominierend. Für das ältere Deutsch existieren nur verhältnismäßig wenige diachrone Untersuchungen.[2] Im Bereich der Korpus- und Computerlinguistik werden formelhafte Wendungen seit den 60er Jahren berücksichtigt, in dem sie im weiten Sinn als *multiword expressions* (MWEs) oder *collocations* bezeichnet werden. Zurückblickend auf die nun mehr fast 50-jährige Tradition (Sinclair 1987; Fellbaum 2006; Fellbaum 2007; Heid 2007; Heid 2008), werden formelhafte Wendungen seitens der Korpus- und Computerlinguisten bei der Korpusannotation und erst recht bei der Korpusbenutzung immer noch als „harte Nuss" oder „pain in the neck"[3] bezeichnet (Filatkina, im Druck).

Für die historische verbale Formelhaftigkeit kommt eine weitere Schwierigkeit hinzu. Das größte Problem besteht darin, dass heutzutage für die ältesten Sprachstufen des Deutschen seit Beginn seiner Überlieferung im 8. Jahrhundert bis in die Frühe Neuzeit (ca. 1650) kein umfangreiches philologisch verlässliches elektronisches Referenzkorpus vorhanden ist, das die Untersuchung der Dynamik der formelhaften Wendungen epochenübergreifend und empirisch abgesichert erlaubt. Während für das moderne Deutsch die Größe des Korpus, das der Beantwortung formelhafter Fragestellungen dienen soll, bei mehreren Millionen Wörtern angesetzt wird (Geyken 2004), ist es für das historische Material alleine aufgrund der spärlichen Überlieferung (etwa für das Alt- und Frühmittelhochdeutsche) problematisch, diesem Anspruch gerecht zu werden.

Ferner liegen der Korpuslinguistik, aber auch der Sprachwissenschaft nur sporadische Informationen über die text(sorten)spezifische Distribution formelhafter Wendungen vor.

1 Nähere Informationen dazu finden sich unter *http://www.hkfz.uni-trier.de*.
2 Ein ausführlicher Überblick wird in Filatkina (im Druck) gegeben.
3 Vgl. den Überblick in Heid (2008).

Der Korpuserstellung müssten entsprechende sprachhistorische Untersuchungen vorangehen. Eine weitere Schwierigkeit, die bereits bei der Identifikation von formelhaften Wendungen in Texten in Erscheinung tritt, ist theoretischer bzw. terminologischer Art. Sobald man versucht, in älteren Texten nach Phraseologismen zu suchen und die für Phraseologismen geltenden Merkmale (Burger 2007: 11-37) anzuwenden, bemerkt man, dass man relativ bald an Grenzen stößt. Denn diese Merkmale sind nicht eins zu eins aus der bis jetzt vorwiegend gegenwartssprachlich orientierten Phraseologieforschung auf das ältere Material übertragbar. So ist z. B. das Problem der Wortgrenzen für historische Sprachstufen so alt wie die Erforschung derselbigen. Das Kriterium der relativen syntaktischen Festigkeit steht dem Variantenreichtum nicht kodifizierter älterer Sprachstufen entgegen. Das Festigkeitskriterium bezieht sich immer auch auf die Gebräuchlichkeit der Einheit, und auch da wird man aufgrund der Zufälligkeit der schriftlichen Überlieferung historischer deutscher Texte mit großen Problemen konfrontiert. Das hermeneutische Zugangsproblem zur Bedeutungsermittlung in historischen Texten relativiert ferner die Gültigkeit des Merkmals Idiomatizität.

2.2 Formelhaftigkeit aus der Perspektive der Kunstgeschichte

Die Untersuchung der historischen Formelhaftigkeit stellt im Fach Kunstgeschichte ein ganz besonders großes Desiderat dar, das selbst noch einer grundsätzlichen definitorischen Durchdringung harrt. Der Teilbereich des bislang pauschalisiert als „Sprichwortbilder" bezeichneten Phänomens wurde bisher nur anhand weniger Fallbeispiele erarbeitet. Die durchaus verdienstvolle Beschäftigung von Seiten der Sprachwissenschaft, der europäischen Ethnologie oder der Geschichtswissenschaft (Mieder/Sobieski 1999) mit dem Thema „Visualisierung von Sprichwörtern" führte in der Vergangenheit nicht immer zu idealen Ergebnissen, da das Bild oft aus seinem Funktionszusammenhang gelöst und die Frage nach der Ikonografie (dem Bildsinn) bisweilen vernachlässigt wurde. Auf diese Weise wurden teilweise Bilder begrifflich unscharf als „Sprichwortbilder" tituliert, die sich jedoch bei genauerer Kenntnis eines umfangreichen Bildkorpus, wie es die Kunstwissenschaft leisten kann, als Bild ähnlicher Ikonografie ohne „Sprichwortgehalt" entpuppten. Auf der anderen Seite wurden zahlreiche Kunstwerke der „zweiten Reihe", also Werke weniger bekannter Künstler oder Notnamenmeister, bislang nicht analysiert. Erst in jüngster Zeit werden die Funktionsebenen von Sprichwortbildern erheblich erweitert und – in Widerlegung der älteren Forschung – Rezeptionswege und Rezipienten weitaus differenzierter betrachtet sowie in ihrer kulturellen Entwicklung primär im Laufe des 16. Jahrhunderts bewertet (Zemon Davis 1975: 231; 239). Einen ersten Schritt in Richtung einer eingehenden, auf den Methoden des rezeptionsästhetischen Ansatzes basierenden Forschung zu Sprichwortbildern in der Kunst leistete zuerst Sullivan (1991).

Während der Forschungsbereich der visualisierten Formelhaftigkeit völlig neu erarbeitet werden muss, kann bei der Erstellung einer Bild-Text-Datenbank auf Vorgängerprojekte zurückgegriffen werden. Das Fach Kunstgeschichte begann sehr früh, seine in vielen Fällen umfangreichen, für die kunstwissenschaftliche Erforschung unerlässlichen Fotografie-, Microfiche- oder Diapositiv-Korpora digital aufzuarbeiten. Ein Beispiel hierfür ist der

Bildindex Foto Marburg,[4] dessen Grundlage 1,4 Millionen Fotografien aus verschiedenen Institutionen bilden. Ferner begann das genuin interdisziplinär forschende Fach Kunstgeschichte bereits in früheren Datenbanken neben dem Bildkorpus ein mit anderen Fächern vernetzbares Textkorpus aufzubereiten.[5]

Die Genese dieser Computerdatenbank ist beispielhaft für andere, kleinere kunsthistorische Projekte, während das System derselben für eine kunsthistorische Datenbank zur visualisierten Formelhaftigkeit zwar vorbildhaft ist, inhaltlich jedoch von anderer Struktur sein muss, da jedes illustrierte Sprichwort bzw. jede entsprechende Formelhaftigkeit aufgeführt werden sollte, auch wenn das jeweils analysierte Bild aus mehreren hundert Einzelsprichwörtern besteht. Ferner muss die Datenbank die der Bildikonografie zugrunde liegenden Texte in den verschiedenen europäischen Sprachen mit den jeweils belegten Varianten aufführen, da die Illustrationen im gesamten europäischen Bereich kursierten und sich gegenseitig mehrdimensional vielfach beeinflussten. Auch kann eine solche Datenbank nicht auf eine bestehende Fotografie-Sammlung aufbauen, sondern muss sich ihr Korpus von Grund auf neu erarbeiten.

Aufgrund dieser Ausgangssituation mussten für die Erforschung der historischen verbalen und visuellen Formelhaftigkeit in vier Projekten des Historisch-Kulturwissenschaftlichen Forschungszentrums (HKFZ) Trier andere Methodiken und Vorgehensweisen gewählt werden.

3 Verbale und visuelle Formelhaftigkeit in vier Projekten des HKFZ Trier

3.1 Nachwuchsforschergruppe „Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)“

Der in Abschnitt 2 skizzierte Forschungsstand macht deutlich, dass für das ältere Deutsch in der gegenwärtigen Untersuchungsetappe eine erste Bestandsaufnahme anhand der möglichst breiten Textsortenpalette aus unterschiedlichen Epochen, die die textuelle Distribution der formelhaften Wendungen quellenbasiert dokumentiert sowie Einblicke in ihre Verfestigungsprozesse und das funktionale Spektrum gewährt, ein großes Forschungsdesiderat darstellt. Diese Forschungslücke versucht die Nachwuchsforschergruppe „Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)“[6] erstmals zu füllen, indem es eine epochenübergreifende Dokumentation und Kommentierung der historischen Variation und Gebrauchsdynamik der verbalen Formelhaftigkeit in älteren deutschen Texten unterschiedlichster Textsorten aus der Zeitspanne zwischen ca. 750 bis ca. 1700 anstrebt. Dabei handelt es sich hier in mehrfacher Hinsicht um Grundlagenforschung, die die Beantwortung weiterer Fragen ermöglichen soll und in ihren Mittelpunkt zunächst die Dokumentation und die linguistische Annotation des Vari-

4 *http://www.bildindex.de*

5 So etwa das an der Humboldt-Universität Berlin verortete Census-Projekt, das sich der Antikenrezeption widmet: *http://www.census.de*.

6 Eine ausführliche Projektvorstellung findet sich unter: *http://www.hifos.uni-trier.de*.

antenspektrums in älteren Wendungen stellt. Eine Beispielgruppe soll dies veranschaulichen.

(1) a. *Wol ûf, swer tanzen welle nâch der gîgen* (Walther v.d.Vogelweide, 19, 37, vor 1230)
b. *Und must am lesten tantzen als die von zwürch pfiffen* (Justinger, Berner Chron. 339, S. 207, 20; 1420-30)
c. *Vor dich, vor dich, vordencke mich nicht, Noch deyner pheyffe tantze ich nicht* (Prov. Frid., 131; 2. H. 15. Jh.)
d. *Went gy moten na myner pypen springen* (Totentanz, 20;1463)
e. *Went se hebben ... noch myt leve noch myth pranghe Nummende konen darto brynghen, De de wil na erer pipen springhen* (Redentiner Ostersp., 1464; 1214)
f. *He moste al na syner pypen dantzen* (Hagen, Helmst. Chron. 162, S. 119; 1491)
g. *He heft leeff den … de so dantzet, alze he vore synget* (Reinke Vos, 3893; 1498)
h. *Der gemein man můß im nach tantzen, wie er pfeiffet* (Geiler, Brösamlin, I, 59d; 1517)
i. *De na der horen pypen danset, de is der schmede vry. Depudet, obsequitur scorto quicunque bilingui* (Tunnicius, 645; 1513)
j. *Tantzen nach irer alten geigen* (Sachs VI, 381, 33; 1523)
k. *Du dantzt nach deyner alten geygen* (Sachs IV, 46, 27, 1534)
l. *Alls ir mir vor tanczt, also sol ich nach springen* (Füetrer, Lanzelot, 152; um 1467)

Das gegenwartssprachliche Idiom *nach jmds. Pfeife tanzen* ist syntaktisch und semantisch als relativ stabil zu bezeichnen.[7] Diesem Befund steht der in Beispiel (1) angeführte Variantenreichtum in älteren deutschen Texten gegenüber.[8] Vom ersten Beleg bei Walther von der Vogelweide bis in die 2. Hälfte des 16. Jahrhunderts hinein ist das heute eher umgangssprachliche Idiom vielfältig in der Chronistik, in einem Totentanz, in einem geistlichen Spiel, in der didaktischen Literatur, in Predigten und satirischen Werken überliefert. Es findet Eingang in vier parömiologische Sammlungen mit Autoritätsstatus, was möglicherweise seine Verbreitung begünstigt hat und auf die Zugehörigkeit der Wendung zu einem gehobenen Stilregister schließen lässt. Bei der Konstanz der aktuellen Bedeutung ‚alles tun, was jmd. von einem verlangt, jmdm. gehorchen' und dem seit Beginn der Überlieferung hohen Idiomatizitätsgrad bleiben die morphosyntaktische Struktur und lexikalische Besetzung bis hin zu Hans Sachs und über ihn hinaus beweglich. Die Beispiele veranschaulichen die lexikalische Variation im Bereich der substantivischen und verbalen Konstituenten (z. B. die Substitution *pfeife* durch *geige* bzw. durch den Nebensatz *wie er ihm vorsingt, wie er pfeift*; *tanzen* durch *springen* oder die Erweiterung der Struktur durch Adjektiveinschub) sowie die morphosyntaktische Variation (etwa die Negation oder die heute zumindest für den binnendeutschen Sprachraum untypische Null-Besetzung der Leerstelle *jemandes*).

Solche Varianten einer formelhaften Wendung werden im HiFoS-Projekt aus den originalen Handschriften und Drucken manuell exzerpiert und in eine Datenbank eingetragen. Für die Eingabe der Daten wurde eine webbasierte Anwendung implementiert. Das tech-

7 Seine Distribution und die geringe Variationspalette im Gegenwartsdeutschen wurden ausführlich in Filatkina (im Druck) analysiert.

8 Bei dieser Liste handelt es sich um eine exemplarische Auswahl; weitere Belege, u. a. auch aus anderen mittelalterlichen Sprachen, sind im TPMA (1996; Bd. 11: 268f.) verzeichnet.

nische Konzept sieht die Nutzung einer relationalen MySQL-Datenbank vor, die über standardisierte Schnittstellen erreichbar ist. Diese Architektur ermöglicht eine dezentrale Datenpflege und eine simultane Nutzung der Arbeitsplattform. Die historischen Belege werden in der Datenbank (und nicht in den Texten) nach dem im Projekt entwickelten Kriterienkatalog kommentiert. Die Benutzerschnittstelle besteht aus einem Formular, in dem bibliografische Angaben zu den ausgewerteten Quellen erfasst werden, einem Belegkorpus, einer Maske zur Belegsuche und einem Tool zur Belegkorpusverwaltung. Den Kern der Benutzerschnittstelle machen fünf Erfassungsmasken aus, die die Analyse eines jeden Belegs nach drei semiotischen Dimensionen – Lexik/Semantik, Syntax und Pragmatik – sowie nach seinem kulturhistorischen Hintergrund ermöglichen.[9]

Die Datenbank enthält im Moment ca. 10 000 Einträge (Stand: April 2009) überwiegend aus der althochdeutschen Zeit. Die Eingabe erfolgt handschriftengetreu und unicodekonform, inklusive Sonderzeichen und Zeilenumbrüchen. Im deutlichen Unterschied zur gegenwärtigen lexikografischen Praxis wird nicht die abstrakte Nennform einer formelhaften Wendung, sondern ihre in Texten real vorkommende nicht normierte Form sowie ihre Varianz in den Mittelpunkt gestellt. Auf diese Weise wird eine exakt dokumentierte Belegbasis erstellt, die die Vielfalt des Gebrauchs einer Wendung in der Synchronie dokumentiert. Die kontextnahen Kommentierungen stellen eine dem Forschungsstand Rechnung tragende Kombination zwischen freien Kommentierungen (z. B. im Feld „Semantische Paraphrase") und standardisierten Angaben dar. Standardisierung bzw. Normalisierung wird z. B. im Feld „Konstituenten" insofern durchgeführt, dass die sinntragenden Komponenten eines Belegs in diesem Feld ins Neuhochdeutsche übersetzt notiert werden. Die Normierung wird somit auf der Metaebene der einzelnen Konstituenten vorgenommen, sprachhistorische Daten bleiben davon unberührt. Diese standardisierten Angaben können bei der Vernetzung mit anderen an der Universität Trier bereits bestehenden Datenbanken zur verbalen und visuellen Formelhaftigkeit benutzt werden (vgl. die Projektbeschreibungen unten).

Die Untersuchung der Gebrauchsdynamik der formelhaften Wendungen in der Diachronie ermöglicht das Belegverwaltungstool. Mit seiner Hilfe können die im Belegkorpus isoliert in der Reihenfolge der Eingabe stehenden Belege nach unterschiedlichen Kriterien für sprachhistorische Analysen manuell gebündelt werden. Diese Kriterien können von den Mitarbeitern abhängig von der Fragestellung beliebig definiert werden. Abfragbar sind alle Ebenen (Belege sowie Quellen) und alle Felder der Datenbank, so dass die Belege zu ganz unterschiedlichen Gruppen und zu mehreren Gruppen gleichzeitig zugeordnet werden können. Interessiert man sich z. B. für die grammatischen und lexikalischen Varianten des Idioms *nach jemandes Pfeife tanzen*, so können in der Suchmaske die Konstituenten *Pfeife* und *tanzen* in Neuhochdeutsch angeben werden. Im Output erscheint ein Subkorpus wie in Beispiel (1) gezeigt. Zusätzlich zu diesem Tool ist im Projekt ein semiautomatisches

9 Für die theoretische und praktische Unterstützung bei der Konzipierung der Datenbank sei herzlich Prof. Dr. Christiane Fellbaum und ihren Mitarbeiterinnen und Mitarbeitern im Projekt „Kollokationen im Wörterbuch" gedankt. Die HiFoS-Erfassungsmasken wurden an einer anderen Stelle ausführlich beschrieben (Filatkina, im Druck); hier soll deshalb kurz ihre Spezifik zusammengefasst werden.

Programm entwickelt worden, das zu dem aktuell bearbeiteten Referenzbeleg selbständig in der Datenbank nach ähnlichen Belegen sucht und dem Bearbeiter eine Liste mit den besten Treffern für die weitere Analyse liefert (Dostert 2009). Hierzu werden Verfahren verwendet, die Ähnlichkeiten zwischen Zeichenketten (n-Grammen) messen und somit ähnliche Wörter erkennen können. Dieses Verfahren ergänzt die Möglichkeit der manuellen Bündelung und ermöglicht die Zusammenführung der Belege, deren Ähnlichkeit für den menschlichen Bearbeiter nicht sofort ersichtlich ist.

3.2 HKFZ-Projekt „Jiddische Phraseologie im Kontext europäischer Sprachen (*JPhras*)"

Das besprochene Idiom (1) soll uns auch weiter interessieren, findet es sich doch unerwartet in einer weiteren Datenbank, deren Aufbau im Folgenden beschrieben wird. Zu den im HKFZ mit der Erforschung historischer Phraseologie und formelhafter Sprache befassten Projekten gehört auch „*JPhras* – Jiddische Phraseologie im Kontext europäischer Sprachen"[10] bei dem zahlreiche Fragen zu Entstehung, Ausformung und Varianz historischer formelhafter Wendungen des Westjiddischen[11] in unterschiedlichen Texttypen und an unterschiedlichen Schreiborten, sowie kontrastiv zum Mittel- bzw. Frühneuhochdeutschen in den Fokus der Betrachtung treten.

In inhaltlicher Anlehnung an HiFoS und unter Übernahme einer modifizierten Version der HiFoS-Datenbank hat auch *JPhras* sich die Dokumentierung und Kommentierung historischer formelhafter Wendungen zum Ziel gesetzt, wobei angesichts der Überlieferungssituation westjiddischer Texte ein synchroner Schnitt mit einem Schwerpunkt um 1600 ausgewertet und mit rezenten jiddischen Phraseologismen abgeglichen wird. Die Auswahl dieses Zeitfensters ist nicht zuletzt bedingt durch die Tatsache, dass *JPhras* von vornherein auch als korpusbasiertes Projekt konzipiert ist und für aschkenasisches Jiddisch das größte computerlesbare Textkorpus, welches in Trier angelegt wurde, seinen zeitlichen Schwerpunkt ebenfalls um 1600 legt. Das Korpus „Westjiddische Texte Trier (WTT)"[12] umfasst mittlerweile die wichtigsten Werke der Zeit um 1600 (teilweise mit Varianten aus späteren Nachdrucken bzw. aus Vorgängerausgaben), die nach den Trierer Konventionen für westjiddische Texte (Timm 1987) in TUSTEP transkribiert wurden und findet auch

10 *JPhras* wurde nach der Etablierung des Sofja Kovalevskaja-Preis-Projekts „Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)" und dem daraus resultierenden fachlichen Austausch angeregt.

11 Mit Westjiddisch ist hier die ältere jiddische Sprache bezeichnet ab ihren Anfängen im 10. Jh. auf deutschsprachigem Gebiet bis zur Zurück- und Verdrängung im Sog der Assimilationsbestrebungen als Folge der jüdischen Aufklärung unter Moses Mendelssohn. Ostjiddisch bezeichnet dann die moderne jiddische Sprache, die in Osteuropa aus den Wurzeln des Westjiddischen hervorging, im Lauf der Jahrhunderte und vor allem nach dem Niedergang des Westjiddischen, eigene Wege in ihrer Entwicklung nahm und heute über den gesamten Erdball verstreut ist mit geschätzten 3 Millionen (mehr oder weniger) aktiven Sprechern.

12 Das Korpus „Westjiddische Texte Trier (WTT)" wurde v. a. von Neuberg erstellt und aus den öffentlichen Mitteln finanziert (*http://www.jiddistik.uni-trier.de*). Über Kooperationsabkommen können diese Daten zur Verknüpfung mit *JPhras* verwendet werden.

im noch laufenden DFG-Projekt „Erstellung einer Datenbank jiddischer lexikografischer Hilfsmittel“ eine weitere Pflege und Erweiterung.

Für den Abgleich mit rezenten ostjiddischen Belegen steht ferner der noch unveröffentlichte, in Trier digitalisierte Thesaurus der jiddischen Sprache von Stuchkoff (1950) mit seinen phraseologischen Einträgen über ein Kooperationsabkommen zur Verfügung. In der laufenden Projektphase (Anschubphase) werden die Belegkandidaten[13] aus elektronisch verfügbaren westjiddischen Texten exzerpiert, manuell in die Datenbank eingetragen und kommentiert. Projektiert ist jedoch die Entwicklung eines Annotationssystems zur Kennzeichnung von formelhaften Wendungen in den Textdateien zur (semi-)automatischen Codierung über alle Texte. Ferner sollen die Belegkandidaten mit dem jeweiligen Text seiten- und zeilengetreu verknüpft werden, so dass auf die manuelle Eingabe eines Kontextes in die Datenbank verzichtet werden kann. Zur schnellen Orientierung wird der jeweilige Kontext über syntaktische Regeln automatisch in die Datenbank eingetragen; die Phraseologiedatenbank ist dann ferner mit dem jeweiligen Quelltext aus dem WTT-Korpus verbunden, so dass für die tiefere Recherche der gesamte Text als Kontext zur betrachteten formelhaften Wendung gilt.

Zwar steht die Identifizierung, Sammlung und Dokumentation der Belegkandidaten in einer relationalen internetfähigen MySQL-Datenbank im Zentrum des Projektes, doch konnten bereits wiederholt exemplarisch auch kontrastive Auswertungen erfolgen, die unterschiedliche Überlieferungswege im Jiddischen und im Deutschen aufzeigen und dabei kultur- und sprachspezifische Übernahmetendenzen beleuchten. Neben den besonders interessanten Aspekten wie der Nähe des Jiddischen zu historischen deutschen Sprachstufen und zur mündlichen Überlieferung des Deutschen respektive deutscher Mundarten gilt auch die Bereicherung durch zusätzliche Entlehnungsquellen bei Kontakt mit weiteren Varietäten (Dialektausgleich) und Sprachen, denen die Jiddischsprecher durch Migration, sowie durch weitläufige (Handels- oder andere) Beziehungen ausgesetzt waren, als untersuchenswert.[14] Auch die für das Jiddische einzigartige ‚prinzipielle Zweisprachigkeit‘[15] sowie die daraus resultierende Fülle an semitischen Lehnübernahmen umschließt ein eigenes

13 Es ist evident, dass bereits die Identifizierung und Benennung eines Textbausteins als ‚formelhafte Wendung‘ oder ‚Phraseologismus‘ bei der Arbeit mit historischen Texten eine Herausforderung darstellt. Festigkeit, usuelle Verwendung oder gar Idiomatizität zu konstatieren stellt in Ermangelung gerade ähnlich strukturierter Untersuchungen wie der in diesem Beitrag beschriebenen einen Akt der Willkür dar und nimmt in gewisser Weise bereits das statistische Endergebnis einer abschließenden Datenbankauswertung vorweg. Um aber dennoch Textstellen als potentiell formelhafte Fundstücke in die Datenbank aufnehmen zu können, sprechen wir von Belegkandidaten. Auch wo im Folgenden der Begriff ‚Beleg‘ verwendet wird, ist im engeren Sinne ‚Belegkandidat‘ gemeint.

14 Grundsätzliche Überlegungen zur (historischen) jiddischen Phraseologie und dessen Beziehungen zum Deutschen und anderen europäischen Sprachen werden mit Beispielen dargelegt in Kleine (im Druck).

15 Da der Erwerb hebräischer Sprachkenntnisse im Judentum von je her einen großen Stellenwert hatte, ist es üblich, auf die ‚prinzipielle Zweisprachigkeit‘ zu verweisen, wenn auf den selbstverständlichen Gebrauch hebräischer und hebräisch-aramäischer Einschübe in jiddischen Texten Bezug genommen wird.

Forschungsfeld. Besondere Betrachtung verdient schließlich die kulturelle Perspektive, denn bestimmte gesellschaftliche Bilder und somit auch sprachliche Konzepte können nur in die jüdische, nicht aber in die christliche Sprachwelt eindringen; so finden etwa Begriffe der jüdischen Gerichtsbarkeit, der Kaschrut[16] usw. kaum Eingang in die deutsche Sprache der christlichen Umgebung.[17] Umgekehrt vermuten wir in der jiddischen Phraseologie keine Anklänge etwa an neutestamentarische Zitate. Doch zu unserer Überraschung finden sich auch in jiddischen Texten Belege für das oben gewählte Beispiel (1) *nach jemandes Pfeife tanzen*, vgl. einen Beleg aus *Paris un' Wiene* aus einem westjiddischen Stanzenroman von (oder aus dem Umkreis von) Elija Levita, gedruckt im Jahre 1594:

(2) Der münch (Mönch) widėr zu Pȧris špȧch: „wi'-wol ich sich di' grȯś šacone (Gefahr), denȯcht, was du' wilśt, dȧś wil ich ach (auch), wen nȯrt (nur) zu bȯk (Gott) is mein cawone (Hingabe)." „nain", špȧch Pȧris, „ich wil di' sach wol vūrėn recht un' wil ünś schȯnẹ." un' sagėt im, wi' er es wolt an-grȯüfėn (angehen, anpacken); der münch, der tanzėt, as er was pȯüfėn (pfeifen) (1594, 599).

Eine Sammlung *bilderishe oysdrukn in yidish* von Guri (2002) verzeichnet für die moderne Jiddische Sprache unter Nr. 169:

(3) tantsn nokh yenems fayfl, dudek.[18]

Gewöhnlich wird das Idiom *nach jemandes Pfeife tanzen* in Verbindung mit einer neutestamentarischen Quelle gesehen und zwar Matthäus 11,15-17 (vgl. auch Lukas 7,32):

(4) [15]Wer Ohren hat, der höre! [16]Mit wem aber soll ich dies Geschlecht vergleichen? Es ist den Kindern gleich, die auf den Märkten sitzen und den anderen zurufen [17]und sagen: **Wir haben euch gepfiffen, und ihr habt nicht getanzt;** wir haben Klagelieder gesungen, und ihr habt nicht wehgeklagt.[19]

Sicher haben jene Evangelien zur Verbreitung des Ausdrucks in vielen europäischen Sprachen beigetragen. Aber wie kommen sie ins Jiddische? Bei Beleg (2) aus *Paris un' Wiene*

16 Kaschrut bezeichnet die Gesetze zur Einhaltung der rituellen Reinheit, v. a. im Zusammenhang mit dem Genuss von Lebensmitteln aber auch in zahlreichen weiteren Bereichen des täglichen Lebens.

17 Dass freilich auch Christen etwas „nicht koscher" finden können, gehört wohl zu den selteneren Fällen einer Übernahme jüdischer Begrifflichkeiten ins Deutsche. Bemerkenswert ist daran auch, dass bislang keine westjiddischen Belege für die Verwendung dieses Ausdrucks in der für das Deutsche üblichen übertragenen Bedeutung *‚nicht geheuer'* (DWB, Grimm, 1854-1960. Online-Version v. Kompetenzzentrum Trier; s. v. ‚kauscher') gefunden wurden. Es sind erst die modernen ostjiddischen Texte, die diese Bedeutungserweiterung ebenfalls wagen.

18 Identisch auch im Thesaurus der jiddischen Sprache von Stuchkoff (1950) (Nr. 466 *folgevdikayt, untertenikayt*).

19 So zitiert nach der Elberfelder Bibel. Anders hingegen in der revidierten Lutherübersetzung von 1984, die an der entscheidenden Stelle wiedergibt „[17]*Wir haben euch aufgespielt, und ihr wolltet nicht tanzen*".

kann natürlich darauf verwiesen werden, dass dem Schöpfer des jiddischen Werkes dieses Bild eventuell aus seiner italienischen Vorlage in den Mund gelegt wurde; dennoch musste er es im Jiddischen kennen, um es einem jiddischsprachigen Publikum vorzusetzen. Der Pfad kann weiter zu Äsop und der Fabel vom Fischer geführt werden, der mit Hilfe seiner Flöte/Pfeife versuchte, die Fische zu sich zu locken. Dies misslang, er griff zum Netz und machte damit einen großen Fang.

Wichtiger als die Herkunft jenes Idioms ist uns im Projektverbund jedoch die Frage nach Varianz und Pragmatik. Bezüglich letzterem fällt auf, dass es in dem westjiddischen Beispiel aus *Paris un' Wiene* ausgerechnet ein christlicher Mönch ist, der nach der Pfeife tanzt. Zur Varianz, wie auch zur weiteren Annäherung an pragmatische Verwendungszusammenhänge fehlen derzeit aber noch Vergleichsbelege. Der weitere Ausbau der Datenbank dürfte hier neue Erklärungsansätze liefern.

3.3 HKFZ-Projekt „Darstellung der luxemburgischen Phraseologie in der Lexikografie und darüber hinaus. Wissensräume zwischen Regionalität und Mehrsprachigkeit (*LuxPhras*)"

Während also das Westjiddische bisher keine Variation gegenüber der im Deutschen heute üblichen lexikalischen Besetzung für das Idiom *nach jemandes Pfeife tanzen* bietet, finden sich interessante Verweise in der Datenbank eines Projektes zur Luxemburgischen Phraseologie, *LuxPhras*. Zu den dort verzeichneten Phraseologismen gehört auch das oben bereits in 3.1 und 3.2 besprochene Idiom. Es ist im Luxemburger Wörterbuch (LWB) unter den Lemmata *danzen* ‚tanzen' (5-a) und *Päif* ‚Pfeife' (5-b) sowie im älteren Wörterbuch der Luxemburger Mundart (WLM) unter *danzen* (5-c) verzeichnet:

(5) a. *mir mussen all no sénger Päif danzen*
b. *ech gin nët op d'Päif* (ich bin nicht sofort zu Diensten)
c. *no èngem senger Peif danzen*

In diesem Zusammenhang wird der Blick in die Datenbank von *LuxPhras* aber erst durch den Abgleich mit den Wörterbüchern der Großregion besonders interessant. Sie verzeichnen das Idiom mit Varianten, wie die Beispiele unter (6) zeigen; vgl. das Wörterbuch der deutsch-lothringischen Mundart s. v. *danzen* (6-a) bzw. das Pfälzische Wörterbuch (6-b) und auch das Rheinische Wörterbuch (6-c) s. v. *tanzen*:

(6) a. *danzen wie äner pifft*
b. *Er muß danze, wie sei Fraa geit* (geigt) [KU-Bedb, LU-Limbghf Friesh], *wie sei Fraa peift* [Pirmas, verbr.], *wies'm sei Fraa vormacht* [WD-Niedkch] bzw. *'s muß alles nooch ihrer Geig (nooch ihrer Peif) danze*
c. *De danzt wie der annere ofspillt* (ist ihm folgsam; Bernk-Maring); *wie der ene peift (flöt), esu danzt der annere; du sollst noch danze, wie eich peife (flöte)* (Mosfrk, Allg.)

Aufgrund der Nähe des Luxemburgischen zu deutschen Mundarten innerhalb des moselfränkischen Dialektkontinuums sind zur Erschließung des phraseologischen Materials

regionale Zusammenhänge, wie auch die spezifische Multilingualität der Sprechergruppen innerhalb des konzeptionellen Wissensraums „Kommunikation" aufschlussreich. Vergleiche mit anderen Mundarten der Großregion[20] sollen dabei Gemeinsamkeiten wie auch Alleinstellung des Luxemburgischen innerhalb der Sprachregion aufzeigen. *LuxPhras* ist ein Tochter- und Folgeprojekt des an der Université du Luxembourg angesiedelten Projekts *LexicoLux*[21] zur lexikografischen und metalexikografischen Erschließung des luxemburgischen Wortschatzes;[22] auch eine erste Erfassung des phraseologischen Bestandes innerhalb der luxemburgischen Wörterbücher soll darin geleistet werden. Doch stellt eine tiefere Auswertung und Aufbereitung phraseologischer Einheiten die Lexikografie vor besondere Herausforderungen durch die bekannte semantische, pragmatische, distributive und syntaktische Irregularität ihres Untersuchungsgegenstandes.

LuxPhras nimmt nun methodisch eine Zwischenposition zwischen HiFoS und *JPhras* ein. Einerseits zieht es die Belegkandidaten über EDV-philologisch definierte Routinen (weitgehend) automatisiert aus den mehrfach annotierten Wörterbuchdateien und überführt sie in eine Datenbank, wo sie manuell bearbeitet werden. In diesem Schritt werden fehlerhaft erfasste Textpassagen getilgt und die gefundenen Belege unterschiedlichen Phraseologismustypen zugeordnet. Die Verknüpfung mit dem Ursprungslemma im Wörterbuch bleibt über eindeutige Identifikationsnummern erhalten, so dass zwischen Phraseologiedatenbank und Wörterbuch hin und her gesprungen werden kann. Andererseits soll auch der phraseologische Bestand des Luxemburgischen über die kodifikatorisch-lexikografische Literatur hinaus Eingang in die Phraseologiedatenbank des Luxemburgischen finden. Hierzu werden in der derzeitigen Anschubphase des Projektes primär Kinderbücher manuell ausgewertet und die rezenten Belegkandidaten in die Datenbank eingetragen und kommentiert. Über manuelle Auswertungsverfahren werden in der Zukunft auch bereits vorliegende Sammlungen zur luxemburgischen Phraseologie[23] aufgenommen.

Projektiert ist für eine spätere Phase auch die automatische Phraseologismenidentifikation in Texten der Luxemburger Klassiker, sofern diese in elektronischer Form vorliegen. Die Zuordnung z. B. von flektierten Vorkommensformen in authentischen Texten zu einer eingetragenen ‚Nennform' im Wörterbuch setzt eine Lemmatisierung voraus. Für die Verben des Luxemburgischen kann dies als annähernd geleistet betrachtet werden, da mit *Luxo-*

20 LoWB: Wörterbuch der deutsch-lothringischen Mundart (1909), RhWB: Rheinisches Wörterbuch (1923-1971), PfWB: Pfälzisches Wörterbuch (1965-1997), EWB: Wörterbuch der elsässischen Mundarten (1899-1907).

21 Eine detaillierte Projektbeschreibung zu „LexicoLux: Erschließung und Vernetzung lexikografischen Wissens über das Luxemburgische" findet sich unter *http://lexicolux.uni.lu*.

22 Als Basis für das lexikografische Wissen über das Luxemburgische gelten die drei vorhandenen Wörterbücher Lexicon der Luxemburger Umgangssprache (LLU), Wörterbuch der luxemburgischen Mundart (WLM) und Luxemburger Wörterbuch (LWB). Sie dokumentieren zugleich den Beginn der kodifikatorisch-lexikografischen Erfassung des Luxemburgischen und werden somit auch im Sinne historischer Quellen erschlossen.

23 Neben der grundlegenden Monografie von Filatkina (2005), sollen auch regionale Sammlungen wie Comes (1935-1959) oder Lehrwerke wie Christophory (1973a, 1973b) u. ä. eingearbeitet werden.

gramm[24] ein Kooperationsprojekt vollständige Verbparadigmen (fast) aller Tätigkeitswörter elektronisch bereitsteht. Routinen für eine frequenzgesteuerte Kookurrenzanalyse an Texten des Luxemburgischen Autors Michel Rodange werden zur Zeit getestet und sollen die korpusgestützte (semi-)automatische Annotation und Exzerption von Phraseologismen in einer späteren Projektphase erweitern. Auch *LuxPhras* beteiligt sich an der Entwicklung eines Annotationssystems zur Kennzeichnung von Phraseologismen in Textdateien. Darin sollen u. a. die Herausforderungen der Verknüpfung von Datenbanken unterschiedlicher Sprachen, unterschiedlicher Schriftsysteme und Medien (Bild und Text) innerhalb eines kulturthematischen Bereiches mit modernen Methoden der eHumanities bearbeitet werden.

3.4 HKFZ-Projekt „Gnomisches Wissen im Raum der Bilder" – Gnomik Visuell. Die Visualisierung von Sprichwörtern und Phraseologismen in Mittelalter und Früher Neuzeit (GnoVis)

Neben den skizzierten sprachwissenschaftlichen Projekten wurde 2007 das interdisziplinäre Teilprojekt „Gnomik Visuell (GnoVis) – Die Visualisierung von Sprichwörtern und Phraseologismen in Mittelalter und Früher Neuzeit" eingerichtet. GnoVis[25] verstand sich von Beginn an als enges Kooperationsprojekt zu HiFoS, indem unter anderem von Anfang an vorgesehen war, die einzusetzende Bild-Text-Datenbank mit der linguistischen Datenbank zu verknüpfen.

Mit dem Projekt werden Sprichwortbilder und visualisierte formelhafte Wendungen verschiedenster künstlerischer Gattungen primär des 15. bis 18. Jahrhunderts untersucht. Desiderat der bisherigen Interpretationen sind bislang die Funktion, topografische Kontinuität und Divergenz der einzelnen Bildthemen, ihre Öffentlichkeit(en) und ihre Rezipientenkreise. Ein weiterer wichtiger Aspekt ist das Verhältnis der formelhaften Wendungen zum Phänomen des Emblems. Übergeordnet schließt sich hieran eine weitere zentrale Frage an, jene nach dem Grad der Exklusivität der vormalig pauschal als Sprichwortbilder bezeichneten visualisierten formelhaften Wendungen gerade im Barockzeitalter und im Gegensatz hierzu das entsprechende spätmittelalterliche Kunstwerk, das bislang allein mit den Arbeitsmethoden der historischen Volkskulturforschung analysiert wurde (Dundes/Stibbe 1981f, Nr. 230-1; Zemon Davis 1984, 78ff.).

Eine Datenbank, die die Visualisierung von Formelhaftigkeit aufbereitet, muss bestimmte Voraussetzungen erfüllen, die für eine Beschäftigung mit dem Material charakteristisch sind, denn zahlreiche Artefakte enthalten Erklärungen, so beispielsweise auch Hieronymus Boschs berühmte Federzeichnung, die unter dem Titel *De boschen hebben ooren, en de velden oogen* firmiert (vgl. Abbildung 1 auf der nächsten Seite).[26] Die Datenbank muss nicht nur das Bild mit allen relevanten Details in Einzelaufnahmen aufbereiten, sondern es müssen auch das zugrunde liegende Sprichwort und seine Varianten in verschiedenen

24 Zu Luxogramm: Grammatisches Informationssystem zum Luxemburgischen, vgl. *http://luxogramm.uni.lu/*

25 Das griechische *gnomon* wird hier im Sinne von stehender Redewendung oder Sentenz übersetzt.

26 Sammlung Museen Preußischer Kulturbesitz, Kupferstichkabinett Berlin, Signatur: 549.

Abbildung 1: H. Bosch: Das Feld hat Augen, die Wälder Ohren, ohne Datierung, Kupferstichkabinett Berlin

Sprachen ebenso eingearbeitet werden wie auf dem Blatt befindliche Inschriften.[27] Das Blatt wurde in der Forschung vielfach gedeutet, zunächst als stilisiertes Selbstporträt (*Bosch* bedeutet im Niederländischen ‚Wald') oder als Illustration der sieben Planeten und der sieben Augen des Herrn (Zach. 3,9).[28]

Der früheste textliche Beleg dieses Sprichworts findet sich möglicherweise in der *Fecunda ratis* des Egbert von Lüttich aus dem 11. Jahrhundert: *Silva suas aures et habet sua lumina campus* – „Der Wald hat seine Ohren und das Feld hat seine Augen", während es hiernach bei Reinmar von Zweter auftaucht, aber auch im theologischen Raum etwa bei Bernhard von Clairvaux.

Diese verschiedenen Interpretationsansätze, die teilweise auch Einfluss auf die Titulierung des Blattes haben, müssen bei der Erarbeitung des Eintrags in die Datenbank berücksichtigt werden. Am besten eignen sich dazu die Verknüpfungen mit Datensätzen zu den entsprechenden Textquellen, die vollständige Transkriptionen und Provenienznachweise enthalten. Ferner soll in der Erfassungsmaske die aufgeführte Bibliografie verzeichnet

27 *Miserimi quippe est ingenii semper uti bestiis et numquam jucundis* – „Das Zeichen oder Merkmal eines unglücklichen Charakters / Temperaments ist es, immer in der Gesellschaft mit den Schlechten, und nie in der Gesellschaft der Freundlichen zu sein." (Übers. d. Verf.)

28 Zur Erforschung der visuellen Umsetzung des Sprichworts primär Bambeck (1987).

werden, um Parallelverbindungen zu anderen Kunstwerken aufzeigen zu können. Dies wird auch anhand des bereits angeführten Beispiels *nach jemandes Pfeife tanzen* deutlich. In Röhrichs Sammlung sprichwörtlicher Redensarten (Röhrich 1991: 166) findet sich als Illustrierung des Idioms ein Holzschnitt vom tanzenden und Flöte spielenden Tod. Problematisch ist hieran, dass ein etymologischer Zusammenhang zwischen dem Motiv des *danse macabre* und dem Idiom suggeriert wird, der nicht nachzuweisen ist und höchstens durch kulturgeschichtliche Analogiebildung nahe gelegt wird. Solche bloßen Illustrierungen sind strikt zu unterscheiden von den eigentlichen Sprichwortbildern oder Visualisierungen von Formelhaftigkeit. So bildet auch der Rattenfänger von Hameln zwar eine Analogie stilistisch-semantischer Qualität, aber keineswegs eine Visualisierung des Sprichwortes. Fündig wird man jedoch im flämischen Raum, wo eine Variation des Sprichwortes *Er lässt alle nach seiner Pfeife tanzen* zu *Er lässt die Welt auf seinem Daumen tanzen* existiert, was in zahlreichen Gemälden, etwa von Frans Hogenberg oder Pieter Brueghel d. J. mit einem auf dem Finger drehenden Globus wiedergegeben wurde.

Die Datenbank für das Projekt GnoVis befindet sich noch im Aufbau und enthält derzeit von den bereits archivierten 1500 Bilddokumenten 650 Einträge.[29] Sie steht in Kooperation mit dem am Fach Kunstgeschichte angesiedelten Projekt Tridoc und nutzt derzeit dessen Plattformen.[30] In der Forschungsdatenbank GnoVis soll nach unterschiedlichen Kriterien gesucht werden:

1. Personen (Künstler, Auftraggeber, Mäzene, Besitzer von Kunstwerken, Dichter, Schriftsteller, Verfasser von Randbemerkungen o. ä.)
2. Institutionen (Museen, Universitäten, Kupferstichkabinette)
3. Standorte der Objekte
4. Bibliografie
5. Technik (Holzschnitt, Kupferstich, Tapisserie, Scheibenrisse, Glasmalerei)
6. Bilddetails (z. B.: Eule, Wald, Augen)
7. Formelhafte Wendungen (Sprichwörter, Phraseologismen).

---

29 Für die Einrichtung der vorläufigen Datenbank, die wissenschaftliche Betreuung sowie die Kooperationsmöglichkeit sei an dieser Stelle Dr. Georg Schelbert M.A. (Fach Kunstgeschichte, Universität Trier/Biblioteca Hertziana, Rom) und seinen Mitarbeitern gedankt.

30 Tridoc (Leitung: Georg Schelbert) versteht sich als Plattform zur digitalen Dokumentation von Daten zur Kultur und Geschichte in Trier und seiner Region. Dabei werden nicht nur Objekte, sondern auch Daten zu Personen und Körperschaften und deren Beziehung untereinander erfasst. Grundlage bildet ein webbasiertes Datenbanksystem mit einem stark differenzierten objektrationalen Datenmodell, das den ereignishaften, von Subjekt, Objekt, Raum und Zeit bestimmten Bezug der in der Datenbank enthaltenen Entitäten dokumentiert; Informationen unter: *http://www.uni-trier.de/index.php?id=22299*. Tridoc ist eng mit dem in Rom (Biblioteca Hertziana) und Trier (Fach Kunstgeschichte) entwickelten Projekt *Zuccaro* (*http://zuccaro.biblhertz.it*) verbunden.

Letzteres bedeutet, dass in der Datenbank bei der Eingabe einer bestimmten formelhaften Wendung (z. B.: *Der König trinkt!* oder *zwischen zwei Stühlen sitzen*) alle Bildwerke samt Detailabbildungen sofort abrufbar sind. Ikonografische und stilistische Verbindungslinien können so erstmalig gezogen werden.

Das System der verwendeten Forschungsdatenbank (Filemaker-Pro) ist auf den geisteswissenschaftlichen Benutzer zugeschnitten und besitzt viele Vorteile. Zu nennen wäre die Fähigkeit, Detailaufnahmen aus einem Kunstwerk mit einzubeziehen und hierzu ebenfalls Datensätze zu erstellen. Die Details können online erstellt und in die Datenbank eingespeist werden. Darauf aufbauend können weitere Informationen zu einem entsprechenden Sprichwortdatensatz eingefügt werden.[31]

Das Projekt sieht seit der Konzeptionsphase eine Vernetzung mit den anderen Datenbanken der Kooperationsgruppe als unerlässlich an. Es initiierte jedoch darüber hinaus eine Vernetzung mit anderen kunsthistorischen Projekten. Das wichtigste Datenbankprojekt auf kunsthistorischer Seite, das sich unter anderem auch mit Artefakten aus dem Bereich „Sprichwortbild" befasst, das Projekt STALLA,[32] konnte als ein Kooperationspartner im HKFZ angeworben werden. Eine weitere Kooperation bildet eine interdisziplinäre Publikation zu den Sprichwortbildern bei Jacob Jordaens.[33]

4 Zusammenfassung und Ausblick

Die Forschungsziele der hier vorgestellten HKFZ-Projekte gehen einerseits weit über rein korpus- und computerlinguistische Fragestellungen hinaus, andererseits liefern sie genaue Angaben zur Verwendung verbaler und visueller formelhafter Wendungen und ihrer Variationen, auf deren Grundlage in weiteren Untersuchungen u. a. formalisierte Annotationsstandards für Formelhaftigkeit entwickelt werden könnten. Die Restriktionen, die das Arbeiten mit historischem Material, mit wenig kodifizierten Sprachen sowie der Kombination von Sprache(n) mit visuellen Medien mit sich bringt, verstehen wir nicht als Hindernisse für korpus- und computergestützte Methoden, sondern als Herausforderungen. Mit dem Bewusstsein für diese Herausforderungen einerseits und mit Blick auf die aktuelle Forschungslage in den beteiligten geisteswissenschaftlichen Disziplinen andererseits entwickeln die vorgestellten Projekte neue Vorgehensweisen, die im Sinne der Grundlagenforschung empirisch abgesicherte Angaben zur historischen verbalen und visuellen Formelhaftigkeit liefern. Sie können und sollen bei zukünftiger Entwicklung

31 Hierzu zählt beispielsweise die Übersetzung des Sprichwortes in verschiedene Sprachen (bisher möglich: Deutsch, Niederländisch, Spanisch, Englisch, Französisch, Italienisch und Lateinisch).

32 Vgl. *http://www.let.ru.nl/ckd/koorbank/*. Untersucht werden hier Bilder und Daten zum europäischen Chorgestühl und die darauf befindliche figürliche Skulptur und Plastik. Leiter des Projekts und Kooperationsmitarbeiter von GnoVis und der AG Kommunikation ist Prof. Jos de Koldeweij (Universität Nijmegen).

33 Münch/Pataki (im Druck) ist die erste Publikation zu diesem Thema – eine grundsätzliche Analyse aus kunsthistorischer Perspektive.

anwendungsbezogener Richtlinien und Annotationsstandards sowie bei Konzipierung der multidirektionalen Verknüpfung von philologischen und kunsthistorischen Datenbanken berücksichtigt werden. Darüber hinaus sind Kooperationen mit Datenbanken einer Reihe weiterer geisteswissenschaftlicher Disziplinen unter dem Dach der eHumanities realisierbar, weil die ausschöpfende Dechiffrierung von Bild- und Textquellen nur im Zusammenspiel und Dialog zu solchen Fächern wie Archäologie, Medienwissenschaften, Geschichte, Literaturwissenschaften und Europäische Ethnologie möglich ist. Damit öffnen sich Türen für zukünftige Forschung, die über Probleme der sprachlichen Repräsentation oder technischen Organisation hinausgehen und verbale und visuelle Kommunikationstraditionen innovativ erschließen.

Literaturverzeichnis

Bambeck, M. (1987): *Das Sprichwort im Bild. „Der Wald hat Ohren, das Feld hat Augen". Zu einer Zeichnung von Hieronymus Bosch.* Wiesbaden.

Die Bibel (1985): Nach der Übersetzung Martin Luthers. Bibeltext in der revidierten Fassung von 1984. Stuttgart.

Burger, H. (2007): *Phraseologie. Eine Einführung am Beispiel des Deutschen.* 3. Auflage. Berlin.

Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N. (Hrsg.) (2007): *Phraseologie/Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin/New York.

Christophory, J. (1973a): *Who's afraid of Luxembourgish? – Qui a peur du luxembourgeois?* Luxembourg.

Christophory, J. (1973b): *Sot et op Lëtzebuergesch – Dites-le en luxembourgeois – Say it in Luxemburgisch.* Luxembourg.

Comes, I. (1935-1959): „Idiomatik der Echternacher Sprache", in: *Vierteljahresblätter für Luxemburgische Sprachforschung, Volks- u. Ortnamenkunde 1*; 12-15. Erschienen in 21 weiteren Folgen unter dem Titel: „Idiomatik der Echternacher Mundart." In: *Vierteljahresblätter für luxemburgische Sprachforschung, Volks- und Ortsnamenkunde. Nr. 1-8*.

Dostert, H. (in Vorb.): *Klassifizierung sprachhistorischer Belege mit Hilfe maschineller Lernverfahren* (Arbeitstitel). Diplomarbeit, Universität Trier.

Dundes, A./Stibbe, C.A. (1981): *The Art of Mixing Metaphors. A Folkloristic Interpretation of the Netherlandish Proverbs by Pieter Bruegel the Elder*. Helsinki.

DWB: *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm* (1854-1960). Leipzig. *http://www.dwb.uni-trier.de*.

Elberfelder Bibel. Das neue Testament (2006). Wuppertal.

EWB: *Wörterbuch der elsässischen Mundarten* (1974). Bearbeitet von Ernst Martin und Hans Lienhart. 2 Bd. 1899-1907. Straßburg. Nachdruck Berlin/New York.

Fellbaum, Chr. (Hrsg.) (2006): *Corpus-based studies of German idioms and light verbs.* Special issue 19-4 of the *International Journal of Lexicography*.

Fellbaum, Chr. (Hrsg.) (2007): *Idioms and collocations. Corpus-based linguistic and lexicographic studies.* London/New York.

Filatkina, N. (im Druck): „Historische formelhafte Sprache als ‚harte Nuss' der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt", in: *Linguistik online*.

Filatkina, N. (2005): *Phraseologie des Lëtzeburgeschen. Empirische Untersuchungen zu strukturellen, semantisch-pragmatischen und bildlichen Aspekten.* Heidelberg.

Geyken, A. (2004): *What is the optimal corpus size for the study of idioms?* Lecture presented at the annual meeting of the DGfS, Mainz.
Granger, S./Meunier, F. (Hrsg.) (2008): *Phraseology. An interdisciplinary perspective.* Amsterdam.
Guri, Y. (2002): *Vos darft ir mer? 2000 bilderishe oysdrukn in yidish.* [=2000 Idiomatic Expressions in Yiddish]. Jerusalem
Heid, U. (2007): „Computational linguistic aspects of phraseology II". In: Burger et al. (Hrsg.); 1036-1044.
Heid, U. (2008): „Computational phraseology. An overview". In: Granger/Meunier (Hrsg.); 337-360.
Kleine, A. (im Druck): „Jiddische Phraseologie. Phraseologie einer Fusionssprache." In: *Tagungsband zur Internationalen Tagung der Europhras in Helsinki 2008 „Phraseologie: global – areal – regional".*
LoWB: *Wörterbuch der deutsch-lothringischen Mundarten* (1909). Bearbeitet von Ferdinand Follmann (1909). Leipzig. Nachdruck Hildesheim/New York 1971.
LLU: *Lexicon der Luxemburger Umgangssprache (wie sie in und um Luxemburg gesprochen wird) mit hochdeutscher und französischer Uebersetzung und Erklärung* (1857). Hg. v. J. F. Gangler. Luxemburg.
LWB: *Luxemburger Wörterbuch* (1950-1977). Hg. v. der Luxemburgischen Wörterbuchkommission. 5 Bd. Luxemburg.
Mieder, W./Sobieski, J. (1999): *Proverb Iconography. An International Bibliography.* New York/Washington/Baltimore.
Münch, B.U./Pataki, Z.A. (im Druck): *Jacob Jordaens. Ein Maler großen Formats.* Stuttgart.
PfWB: *Pfälzisches Wörterbuch* (1965-1997). Begründet von Ernst Christmann, fortgeführt von Julius Krämer, bearbeitet von Rudolf Post unter Mitarbeit von Josef Schwing und Sigrid Bingenheimer. 6 Bd. Wiesbaden/Stuttgart.
RhWB: *Rheinisches Wörterbuch* (1928-1971). Im Auftrag der Preußischen Akademie der Wissenschaften, der Gesellschaft für Rheinische Geschichtskunde und des Provinzialverbandes der Rheinprovinz auf Grund der von Johannes Franck begonnenen, von allen Kreisen des Rheinischen Volkes unterstützten Sammlung bearbeitet und herausgegeben von Josef Müller, Heinrich Dittmaier, Rudolf Schützeichel und Mattias Zender. 9 Bd. Bonn/Berlin.
Röhrich, L. (1991): *Das große Lexikon der sprichwörtlichen Redensarten*. Stuttgart.
Sinclair, J.McH. (1987): „Collocation: A progress report". In: Steele/Threadgold (Hrsg.) (1987); 319-331.
Steele, R./Threadgold, T. (Hrsg.) (1987): *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam.
Stuchkoff, N. (Hg.) (1950): *der oytser fun der yidisher shprakh*. Thesaurus der jiddischen Sprache. New York.
Sullivan, M. (1991): „Bruegel's Proverbs: Art and Audience in the Northern Renaissance", in: *Art Bulletin, 73/3*; 431-466.
Timm, E. (1987): *Graphische und phonische Struktur des Westjiddischen unter besonderer Berücksichtigung der Zeit um 1600.* Tübingen.
Timm, E. (1996): *Paris un Wiene. Ein jiddischer Stanzenroman des 16. Jahrhunderts von (oder aus dem Umkreis von) Elia Levita.* Eingeleitet, in Transkription herausgegeben und kommentiert von Erika Timm unter Mitarbeit von Gustav Adolf Beckmann. Tübingen.
TPMA. *Thesaurus proverbiorum medii aevi. Lexikon der Sprichwörter des germanisch-romanischen Mittelalters.* Begründet von Samuel Singer. Hg. vom Kuratorium Singer der Schweizerischen Akademie der Geistes- und Sozialwissenschaften (1996ff). Berlin/New York.
van Dülmen, R.; Schindler, N. (Hrsg.) (1984): *Volkskultur. Zur Wiederentdeckung des vergessenen Alltags (16.-20. Jahrhundert).* Frankfurt a. M.

WLM: *Wörterbuch der luxemburgischen Mundart* (1906). Druck v. M. Huss. Luxemburg.
Zemon Davis, N. (1984): „Spruchweisheiten und populäre Irrlehren." In: van Dülmen/Schindler (Hrsg.) (1984); 78-116.
Zemon Davis, N. (1975): *Society and Culture in Early Modern France.* Stanford.

Natalia Filatkina
Germanistik, Ältere deutsche Philologie
Nachwuchsforschergruppe HiFoS
Universität Trier
Universitätsring 15
54286 Trier
Deutschland
filatkin@uni-trier.de

Ane Kleine
Fuerschungslaboratoire vun der Lëtzebuerger Sprooch a Literatur
IPSE – LexicoLux – X.20
Universitéit Lëtzebuerg
Campus Walfer
route de Diekirch / B.P. 2
7201 Walferdange
Lëtzebuerg
ane.kleine@uni.lu

Birgit Ulrike Münch
Univesität Trier
Kunstgeschichte
Universitätsring 15
54286 Trier
Deutschland
muench@uni-trier.de

The Collection of Distributionally Idiosyncratic Items: An Interface between Data and Theory

Frank Richter/Manfred Sailer/Beata Trawiński

Dieser Beitrag gibt einen Überblick über CoDII, die *Collection of Distributionally Idiosyncratic Items*. CoDII ist eine elektronische Sammlung verschiedener Untergruppen lexikalischer Elemente, die sich durch idiosynkratische Distribution auszeichnen. Das bedeutet, dass sich die Verteilung dieser Lexeme im Text nicht alleine aufgrund ihrer syntaktischen Kategorie vorhersagen lässt. Die Methoden, die in der Entwicklung von CoDII angewandt werden, greifen über traditionelle Fachgrenzen hinaus und umfassen Korpuslinguistik, Computerlinguistik, Phraseologie und theoretische Sprachwissenschaft. Ein wichtiger Schwerpunkt unserer Diskussion liegt auf der Darstellung, inwiefern die in CoDII gesammelten, annotierten und unter anderem mit Suchwerkzeugen abfragbaren Daten dazu beitragen können, die linguistische Theoriebildung durch die Bereitstellung sorgfältig aufbereiteter Datensammlungen bei der Überprüfung ihrer Datengrundlage zu unterstützen.

1 Introduction

The Collection of Distributionally Idiosyncratic Items (CoDII) is an electronic resource for linguistic research. In its very design it crosses traditional boundaries of several linguistic subdisciplines. The methods and techniques that were used in its creation come from corpus linguistics, computational linguistics, phraseology and theoretical linguistics. Its goal is to provide a resource that is useful for researchers working in areas as diverse as lexicography, syntax, semantics and psycholinguistics. In this paper, we will present the main features of CoDII. An important part of this discussion will be to show that beyond being a valuable data repository that may be used for building specialized (electronic) resources or applications in specific areas of interest, CoDII can support theoretical linguistics by giving researchers structured access to a wealth of data to test and improve their theories.

Distributionally idiosyncratic items (DIIs) are special from two perspectives: First, they don't follow the distribution pattern that would be expected based on their syntactic categorial properties. Because of their irregular distributional properties, they are accessible to statistical corpus linguistic methods. Second, since they are expressions with strict context requirements, their failure to occur in their respective licensing context triggers clearcut ungrammaticality judgments by native speakers. Their location in an area which is

simultaneously accessible to measurements of statistical distribution and to the investigation of the human grammatical system by grammaticality judgments makes DIIs ideally suited for gaining new insight into human language.

Section 2 gives an overview of the structure and content of the five subcollections that CoDII currently consists of, and of the sources that were used to collect the data. Section 3 outlines some of the linguistic questions that can be addressed with the data in CoDII. We will show how these collections can be useful in approaching long-standing problems in linguistics from a new angle and on the basis of new types of empirical evidence. In Section 4 we conclude with a summary of the most important features of CoDII.

2 Data

2.1 Five Collections

CoDII[1] comprises five subcollections: (1) 446 bound words in German (CoDII-BW.de); (2) 77 bound words in English (CoDII-BW.en); (3) 58 negative polarity items in Romanian (CoDII-NPI.ro); (4) 165 negative polarity items in German (CoDII-NPI.de); and (5) 88 positive polarity items in German (CoDII-PPI.de).

Bound words (BWs) are words which may only be used in combination with a fixed set of other words. Typical examples are the English word *headway*, which only occurs in the idiom *to make headway*, and the German word *Bärendienst*, which may only be used in the idiom *jemandem einen Bärendienst erweisen* ('to do so. a disservice'). In a first informal characterization, NPIs are words (or multi-word expressions) which require the presence of some form of negation in their context. A good example is the verb *scheren*, which is only acceptable in negative contexts as provided by the negation adverb *nicht* ('not'), the adverb *niemals* ('never'), or a nominal phrase such as *wenige Studenten* ('few students'). We will take a closer look at the licensing environments of NPIs below. PPIs are in a sense the positive counterpart to NPIs in that they shun negation in their immediate semantic environment.

The main source of the bound words in CoDII-BW.en and CoDII-BW.de are the studies Dobrovol'skij (1988, 1989) and Dobrovol'skij/Piirainen (1994). The items in CoDII-PPI.de were taken from van Os (1989), van der Wouden (1997) and Ernst (2005), with additional items from our own research. The sources for acquiring the NPIs for CoDII-NPI.de include the collections of NPIs in Welte (1978) and Kürschner (1983). To extend the coverage beyond previous literature, NPI candidates were extracted automatically from the *Tübingen Partially Parsed Corpus of Written German (TüPP)*[2]. The extraction algorithm is described in Lichte (2005) and Lichte and Soehn (2007). The items in our smallest collection, CoDII-NPI.ro, are mostly counterparts to the English, German and Dutch NPIs in the linguistic literature, since no specialized collection of Romanian NPIs was available as data source.

1 *www.sfb441.uni-tuebingen.de/a5/codii*
2 *www.sfs.uni-tuebingen.de/en_tuepp.shtml*

2.2 Data Format

Each CoDII item has four basic information blocks: 'General Information', 'Information about Licensing Contexts', 'Syntactic Information', and 'Classificatory Information'. The optional block 'Sample Queries' recommends search patterns that are optimized for important publicly available corpora.

The block 'General Information' identifies an item by providing its word form, an English gloss and a translation (for the non-English items), expressions in which the item occurs, and, if appropriate, paraphrases. This is also where we report occurrences of an NPI outside its theoretically expected licensing environments.

Within the block 'Syntactic Information', each item is assigned a syntactic category. The syntactic structure of the expression in which the item occurs is added where appropriate. Possible syntactic variations are listed, including passivization, pronominalization, modification, topicalization, occurrence in raising or control constructions, and appearance within relative or interrogative clauses. For each syntactic variation, examples from corpora, Internet or the linguistic literature are included. Three tagsets provide the theory and notation for the syntactic description of CoDII items. The *Stuttgart-Tübingen Tagset (STTS)*[3] is used for the syntactic description of German items and of expressions in which they occur. The English BWs are annotated with the syntactic annotation scheme from the *Syntactically Annotated Idiom Database* (SAID, cf. Kuiper et al. 2003). For the syntactic description of Romanian NPIs we take the (modified) tagset from the *Multilingual Text Tools and Corpora for Central and Eastern European Languages (MULTEXT-East)*[4].

The block 'Licensing Contexts' contains information on the licensing environment of each item. In the case of polarity items, the licensing contexts are chosen from general, descriptive categories rather than from classifications in a particular theoretical framework. We distinguish the following licensing environments: clausemate (sentential) negation, non-clausemate negation, n-words (such as *nobody*, *never*), the scope of negation expressed by the determiner *kein-*, the scope of *without*, interpretation in the restrictor of universal quantifiers, other contexts of interpretation which are logically downward-entailing (and are not subsumed by one of the more specific categories), the scope of *only*, the complement clause of negative verbs (such as *doubt, fear* and *regret*), questions, antecedents of conditionals, comparative constructions, superlative constructions, and imperatives. To allow the documentation of all available data, exceptional cases that do not fit any of these predetermined categories are listed as 'Exceptions'. Some of the licensing environments will be discussed in more detail below.

The examples for the usage in their licensing contexts of the items listed in CoDII were collected from electronic and printed sources. The Romanian examples were gathered from Rada Mihalcea's Romanian electronic corpus, and from Internet search with Google. Some examples were constructed by Gianina Iordăchioaia, a native speaker of Romanian, who worked on CoDII-NPI.ro. The sources of the German BWs, NPIs and PPIs were

3 *www.sfs.uni-tuebingen.de/Elwis/stts/stts.html*
4 *nl.ijs.si/ME*

corpora of the Institute of German Language in Mannheim[5], the corpus of the *Digitales Wörterbuch der Deutschen Sprache (DWDS)*[6], and Internet search with Google. The examples in CoDII-BW.en mainly come from dictionaries, from the Internet and from the *British National Corpus* (via the SARA software package[7]).

The last block, 'Classificatory Information', reports, for each item, classifications found in the literature. For English and German BWs, the classifications were taken from Dobrovol'skij (1988, 1989), Dobrovol'skij/Piirainen (1994), and Nunberg et al. (1994). Polarity items are classified as positive or negative, and are subdivided in three semantic classes according to the theory of Zwarts (1997) (see the discussion below). For citations from literature that does not use these semantic distinctions, we use the classification tag 'open'.

The five CoDII collections are encoded in XML with a uniform schema. Technical details for BWs are described in Sailer and Trawiński (2006) and Trawiński et al. (2008ab), and for PIs, in Trawiński/Soehn (2008). Design and data structure of CoDII are conceived in such a way that further types of distributionally idiosyncratic items, such as anaphora, can be modeled, and collections from various languages can easily be integrated using the existing schema.

CoDII not only compiles, documents and (alphabetically) lists distributionally idiosyncratic items. Due to the integration into the Open Source XML database *eXist*,[8] it also offers dynamic and flexible access. The design of the internal data structure and the annotation with syntactic and (partial) semantic information make it possible to query our resource with respect to particular lemmata, syntactic properties and linguistically interesting classifications. First statistical observations on the data in our collections which were obtained by using these database functionalities are reported in Trawiński et al. (2008a, 2008b) and Trawiński/Soehn (2008).

The user interface of CoDII displays all the linguistic information, including syntactic structure and licensing contexts together with the links to corresponding examples (see figure 1 and figure 2). Comments, information about the classification systems, licensing contexts, and examples of the usage of each item in context can be obtained by clicking on the links in the display. All bibliographic references in CoDII are linked to two electronic bibliographies, the 'Bound Words Bibliography'[9], and the 'Polarity Items Bibliography'[10].

2.3 Context Classification and Variation

Figure 1 and figure 2 show the web interface of CoDII for two entries in different subcollections: The German BW *Hehl* ('secret') and the German multi-word NPI *ein(en)*

5 *www.ids-mannheim.de/cosmas2/*
6 *www.dwds.de*
7 *www.natcorp.ox.ac.uk/sara*
8 *exist.sourceforge.net*
9 *www.sfb441.uni-tuebingen.de/a5/bwb*
10 *www.sfb441.uni-tuebingen.de/a5/pib/XML2HTML/list.html*

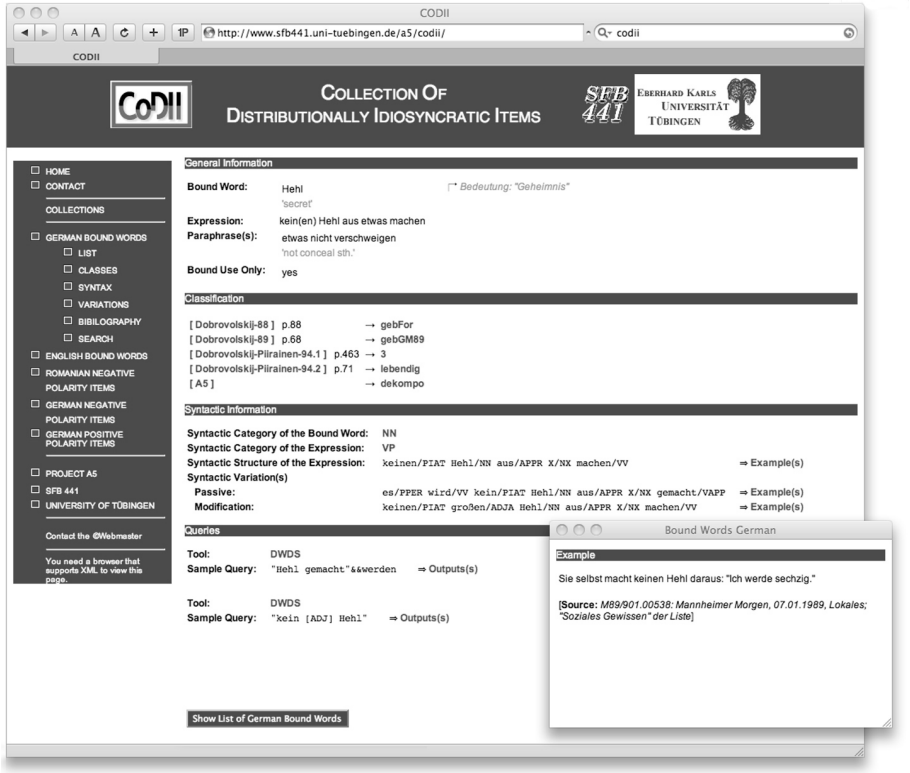

Figure 1: CoDII web interface for the German BW *Hehl* ('secret')

Hehl aus etw. machen ('to make a secret out of sth.').[11] In CoDII-BW.de, *Hehl* is recorded as a word without the usual free distribution of a noun; it may only occur as part of the multi-word expression *ein(en) Hehl aus etw. machen*. Figure 1 shows the information blocks that CoDII records for a bound word, including a window available through the 'Output(s)' link which illustrates one result to a given sample query. Note the links in this

11 With the idea of decomposing the meaning of the idiom *ein(en) Hehl aus etw. machen* and assigning *Hehl* a meaning contribution of its own, we follow the analysis of decomposable VP idioms in Sailer (2003) and Soehn (2006). This does not mean that we suggest to decompose every idiomatic phrase. For example, the words in the non-decomposable VP idiom *die Flinte ins Korn werfen* ('to give up') and the bound word *klipp* in the frozen expression *klipp und klar* ('point-blank') have a very different status with respect to the interpretation of the overall expression.

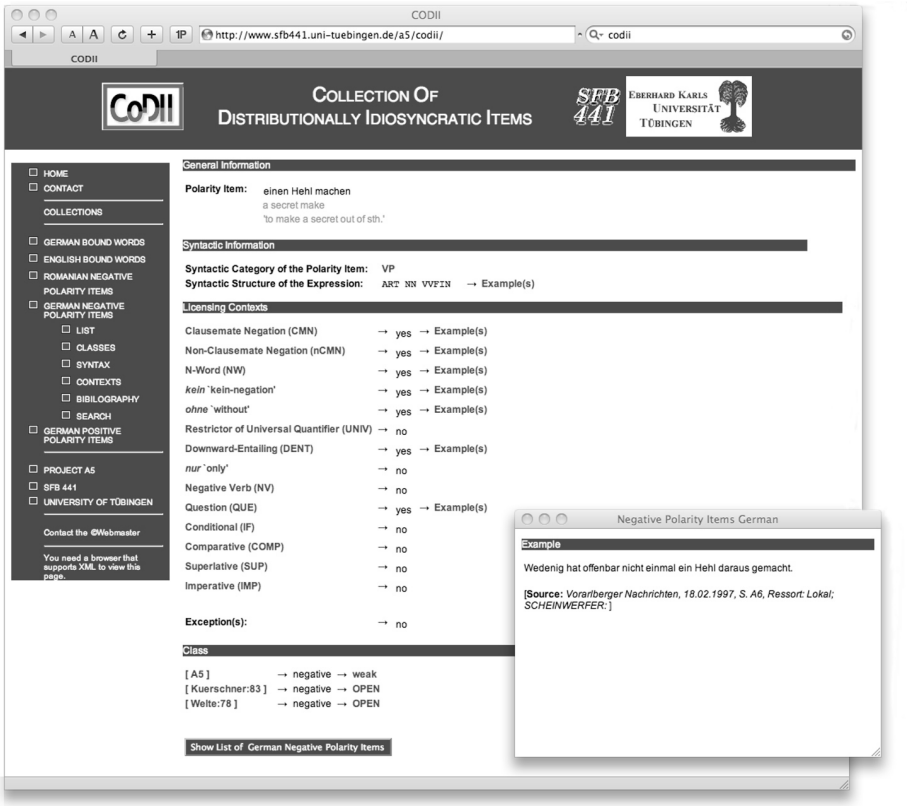

Figure 2: CoDII web interface for the German NPI *ein(en) Hehl aus etw. machen* ('to make a secret out of sth.')

CoDII entry, which provide background information about the various categorizations offered on this page.

Before looking at *Hehl* as item in CoDII-NPI.de, a more precise explanation of NPIs and their semantic subclasses is in order. The three classes of NPIs we distinguish in CoDII, weak NPIs, strong NPIs, and superstrong NPIs, were introduced by Zwarts (1997). In the formulation of the theory given by van der Wouden (1997) they are algebraically defined as follows: (1) NPIs are *superstrong* if they are licensed only by antimorphic contexts (overt

negation).[12] An example of an antimorphic operator is sentential negation. (2) NPIs are *strong* if they are licensed by antimorphic and anti-additive contexts.[13] Examples of anti-additive operators are the expressions *nobody* and *never*. The word *nobody* is shown to be anti-additive by checking that the sentence *Nobody complained or resisted* is true in exactly those situations in which *Nobody complained and nobody resisted* is true. (3) NPIs are *weak* if they are licensed by antimorphic, anti-additive, and downward-entailing contexts (and possibly some others).[14] An example of a plain downward-entailing operator is the phrase *few students*. This phrase is shown to be downward entailing by checking that *Few students complained or resisted* implies *Few students complained and few students resisted*. Moreover, *Few students complained or few students resisted* implies *Few students complained and resisted*. According to the definition of the three NPI classes, any NPI is licensed by sentential negation. Strong NPIs need to be in the scope of an operator that is at least strong. The German strong NPI *einen blassen Schimmer haben* ('to have the faintest idea') is thus licensed by sentential negation and *niemals* ('never') but not by *wenige Studenten* ('few students'). Weak NPIs are already satisfied in the presence of a weak licenser.

Hehl is recorded in CoDII-NPI.de because apart from being a bound word, it is also a lexeme which occurs in a multi-word expression that behaves like an NPI: Figure 2 shows the corresponding CoDII entry and a window with a corpus example reachable through the link 'Example(s)' of the licensing context 'Clausemate Negation (CMN)'. The reader will notice that the '[A5]' classification categorizes the item as a weak negative polarity item. This means that it is an item which only needs a logically weak form of negation as licenser. A reflex of this fact is the existence of corpus evidence in the category 'Downward-Entailing (DENT)'. These are licensing environments that are weaker than the antimorphic 'Clausemate Negation (CMN)' environment or the restrictor of the determiner *kein-*. The NPI classification in CoDII into weak, strong and superstrong is preliminary in the sense that it strictly follows the corpus evidence that we found: It can (and does) happen that an item which is generally considered a weak NPI is classified as strong in CoDII, because we only found corpus evidence for its occurrence with sentential negation, *kein-*, and *ohne* ('without'). It is important to realize that CoDII deliberately stays within the limited horizon of its database and leaves it to the user's judgment and research to revise this preliminary categorization where it is appropriate or necessary.

12 An operator f is antimorphic iff for each set X and for each set Y, $f(X \cup Y)$ equals $f(X) \cap f(Y)$ and $f(X \cap Y)$ equals $f(X) \cup f(Y)$.

13 An operator f is anti-additive iff for each set X and for each set Y, $f(X \cup Y)$ equals $f(X) \cap f(Y)$.

14 An operator f is downward-entailing iff for each set X and for each set Y, $f(X \cup Y)$ implies $f(X) \cap f(Y)$ and $f(X) \cup f(Y)$ implies $f(X \cap Y)$.

3 Theory

3.1 Grammar Theories

Idioms are treated very differently in different areas of linguistics. Two opposite extremes within the overall spectrum are the constructional (holistic) approach, and the collocational approach. The constructional perspective views idioms as syntactic and semantic units which are usually treated as fixed, stored chunks. They are basically conceived of as lexical items, differing from words primarily in that they may be syntactically complex. This perspective is common in the phraseological literature such as Fleischer (1997) and in formal linguistics, be it Generative Grammar (Chomsky 1981), or Construction Grammar (Fillmore et al. 1988). The collocational perspective originates from corpus linguistic research. Under this perspective, the co-occurrence patterns of individual words are studied. If a word co-occurs with a second word more often than expected on the basis of their syntactic category, the two words form a collocation. This perspective is common in computational corpus linguistic research on idioms, such as in computational lexicography (Sinclair 1991, Moon 1998), and in more general computational linguistic approaches such as Krenn (1999).

Interestingly for us, there is a natural area of overlap between these two perspectives: The constituents of what would traditionally be called an idiom may show high co-occurrence ratios in corpora. However, the two perspectives do not cover the same ground. Many idioms are very infrequent in corpora (see Moon 2007), which makes them invisible to the collocational method. On the other hand, many high-frequency co-occurrence pairs do not show any degree of syntactic irregularity or semantic idiomaticity, which makes them irrelevant from the constructional perspective. One of the important missions of CoDII is to demonstrate that this last point is not just an innocent blind spot of the constructional approach. CoDII sets out to contribute to the development of a theory that can overcome this shortcoming and supply a picture of the missing landscape.

Formal grammars usually strive to formulate linguistic generalizations, whereas collocations (and idioms) are by definition idiosyncratic and lexeme-specific. In formal grammars, context effects are occasionally encoded when they capture generalizations about syntactic structures or systematic differences between lexical items. The concept of *selection* plays an important role in this. Selection is responsible for binary combinations of a (syntactic or semantic) functor and its argument: A syntactic head imposes restrictions on its complement(s), and an adjunct imposes restrictions on the syntactic head it combines with. Formal grammars have developed sophisticated means to express these and only these relations and restrictions. They are primarily realized in subcategorization frames or valence specifications in lexical entries. Collocations, however, do not necessarily respect the directions of selection or other grammatical relations. A good example are light-verb constructions, i.e. verb-noun collocations such as *take a shower*, *do the dishes*, *make a mistake*. In these cases the noun is syntactically realized as the complement of the verb. Nonetheless it can be argued that it is the noun that determines which verb must be used in the combination.

In previous work (Richter/Sailer 2003, Sailer 2004) we argued that so-called *bound words* show that at least some collocations should be included within the empirical domain of formal grammar. The underlined words in (1) are *bound* in the sense that they can only occur in this particular context.

(1) a. wend one's way (= make one's way)
b. make headway (= make progress); take/ have a dekko (= take a look)
c. without fail (= fully predictable, with no exception or cause for doubt)
d. the whole caboodle (= the whole lot)
e. flotsam and jetsam (=pieces from a wrecked ship floating in the sea or scattered on the floor)

The data in (1) show that the relation between a bound word and its required context cannot be captured with the means of selection: In (1-b) we see the same pattern as in support verb constructions, i.e. the noun determines which verb has to be chosen. In (1-d) the noun requires the presence of a certain modifier, and in (1-c) the noun must occur as the complement of a particular preposition. In (1-e) there are two bound words occurring in a conjunction. Normally no mutual selection relation is assumed among conjuncts. The data in (2) add a crucial second dimension to the behavior of bound words:

(2) a. achieve progress/ * headway
b. with exceptions/ * fail
c. (i) *jetsam and flotsam (ii) collect *(flotsam and) jetsam

Native speakers of English can give grammaticality judgments about the distribution of bound words. In particular, combinations as in (2) are judged ungrammatical. If it is a central goal of formal grammars to capture the grammaticality judgments of native speakers, the distribution of bound words can certainly not be ignored. The theoretical significance of bound words, thus, lies in their property to exhibit a firm co-occurrence with a particular other word. Crucially, the necessity of this co-occurrence is observable in the grammaticality judgments of native speakers. Despite their clear grammatical relevance, these co-occurrence patterns are not captured by the theory of syntactic selection.

When we turn to polarity items, a different, but equally puzzling picture emerges. We saw in Section 2 that NPIs require the presence of a licensing element, which is prototypically — but not necessarily — sentential negation. Many NPIs are idioms, but negation is an abstract part of the idiom rather than a lexicalized component. For this reason a holistic view on idioms lacks the means to express the negation requirement correctly.

The majority of the formal and theoretical research on polarity items focus on a small number of expressions, primarily on English *any* and *ever*. The contexts in which these NPIs may occur are carefully characterized and categorized. The limitation of this line of research to very few selected items is acknowledged in important contributions to the field, such as Kadmon/Landman (1993), von Fintel (1999), and Chierchia (2004). It is unclear whether the distribution of polarity items in general is captured, or merely the idiosyncratic distribution of certain items. In addition, it remains an open question how the negation

requirement can be linked to the relevant lexical items (or multi-word expressions), let alone what the connection could be to a theory of idioms.

Within the collocational tradition, NPIs have been largely ignored. Sinclair (2004) discusses the verb *budge* (a weak NPI) and observes that it occurs in negative contexts. However, this insight is based on an inductive inspection of corpus data. No precise characterization of the term *negative context* is provided. Hoeksema (1997) presents corpus studies of individual polarity items that confirm the impression that the distribution of polarity items in their potential licensing contexts is not as homogeneous as many theoretical approaches suggest.

3.2 Distribution Profiles

Bearing in mind the general picture outlined in the previous section, we can now look at the contribution of CoDII to the field. The information on contexts and variation collected in CoDII is chosen in a way that makes it easy for researchers to check their theories against more data. Within phraseological research it has often been claimed that the syntactic flexibility of an idiom is related to semantic properties. While passivization as a semantically neutral operation is possible with many VP idioms, spreading an idiom over a main clause and a relative clause seems to be restricted to semantically decomposable idioms, i.e. to idioms whose parts can be assigned a meaning that contributes to the overall meaning of the idiom in a regular way (McCawley 1981, Schenk 1995). The study of bound words in relative clauses is, consequently, of great theoretical importance. If a bound word can occur in a relative clause constellation, it should be possible to assign this word a meaning.

The distributional profiles for polarity items can help us provide a better classification of NPIs. As we have seen earlier, the distinction between our three classes of NPIs is based on the entailment properties of their licensing contexts. Weak NPIs such as German *jemals* ('ever') may occur in all potentially NPI-licensing contexts. Strong NPIs such as German *eine Miene verziehen* ('show emotions') are restricted to sentences that contain the negation adverb *nicht* ('not') or a negative constituent such as *kein- N* ('no N') or *niemals* ('never'). They may also occur in the restrictor of a universal quantifier. Superstrong NPIs such as English *one bit* are claimed to occur only with *not*.

In CoDII we document polarity item data in exactly those contexts that have been looked at in the literature. We assign classifications to the items based on the distribution profiles found for these contexts. The resulting profiles do not confirm the predictions of Zwarts' tripartite theory. For instance, NPI modals such as German *brauchen* and English *need* may occur in many NPI contexts, but are banned from the restrictor of universal quantifiers (Hoeksema 1997). This distributional gap is not predicted in Zwarts' theory. The distribution profiles in CoDII can be used to check whether this unexpected behavior is idiosyncratic to *brauchen* or attested with other NPIs as well.

Pragmatic theories of NPI licensing such as Krifka (1995) and Israel (2004) assume that NPIs are admitted whenever certain pragmatic conditions are met. As a consequence, they predict the availability of NPIs in contexts which are not traditionally considered NPI

licensing contexts. CoDII also includes a field for unusual occurrences. If large numbers of examples can be found in this category, this may be taken as support for pragmatic theories.

In sum, CoDII attempts to provide reliable, qualitative profiles for NPIs. These profiles can be used to confirm or to challenge the predictions of NPI theories.

3.3 Collocations in a Formal Theory of Grammar

Let us finally look at a line of research which tries to encode the data collected in CoDII in a formal theory of grammar, Head-driven Phrase Structure Grammar (HPSG, Pollard/Sag 1994). HPSG is a framework that has its roots in context-free phrase structure grammars. For this reason, the original formulation of the theory in Pollard/Sag (1994) did not provide the means to encode idiosyncratic syntactic and semantic units that span more than a local tree. It was only in recent developments that a link from HPSG to Construction Grammar was established (Ginzburg/Sag 2000, Riehemann 2001), indicating that HPSG takes a constructional perspective on idioms.

In various publications (including Sailer 2003, Soehn 2006, and Richter/Soehn 2006) we developed a collocational module for HPSG that can model the data documented in CoDII. Let us illustrate this collocational module with the NPI *ein Hehl aus etw. machen* that contains the bound word *Hehl*. All distributional idiosyncrasies can be located in the lexical entry, sketched in figure 3.

HPSG is a constraint-based theory that employs feature structures (or similar appropriate mathematical structures) as linguistic representations (Richter 2004). These structures encode all linguistically relevant components of a sign, including the phonological representation, a semantic representation, the syntactic category, and valence information. A prominent part of this is indicated in figure 3. The value of the feature PHON(onology) specifies the phonological representation of the word. In SYNSEM HEAD the syntactic category of the noun is given. We also use a feature LISTEME which provides a unique identification label for each listeme. The relevant notion of a listeme is borrowed from Di Sciullo and Williams (1987), and is meant to subsume simple word lexemes as well as phrasal lexemes. The CONTENT value is the semantic representation of the sign. In figure 3 we simplify and only mention the logical semantic constant that belongs to the word.

We enrich this conventional HPSG architecture with a new feature, COLL (context of lexical licensing), whose value specifies the co-occurrence requirements of a lexical item. The word *Hehl* has two requirements. First, it must occur as the direct object to the support verb *machen* ('make'). Second, the semantics of *Hehl* must occur in the scope of an NPI-licensing operator. These two requirements are expressed in the two elements on the COLL list. Each element defines the syntactic domain within which the collocational restriction has to be met. The first one must hold within the minimal clause that contains the word *Hehl*. The second one only needs to be satisfied within the overall utterance.

The first COLL-element has a feature LOC-LIC. By means of its complex feature value, *Hehl* is collocationally restricted to co-occur with a particular lexeme. The collocating lexeme is identified on the basis of its LISTEME value. In figure 3 this is specified as $machen_2$, which we assume to be the LISTEME value of the required support verb *machen*. More

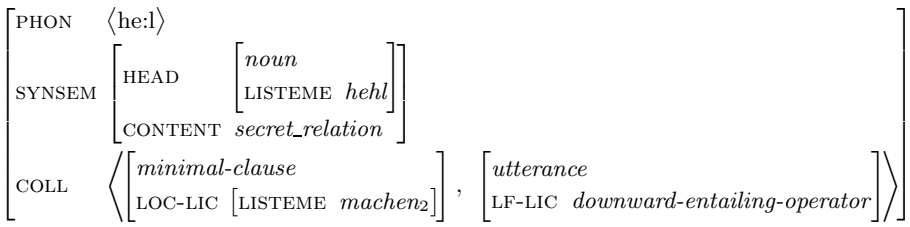

Figure 3: Sketch of the lexical entry of the bound word *Hehl*

examples of this kind of collocational restrictions are discussed in Soehn (2006), which focuses on decomposable and non-decomposable VP idioms.

The second element on the COLL list has a feature LF-LIC, which is short for LOGICAL-FORM-LICENSER. This indicates a semantic restriction. The value of this feature specifies that the semantic contribution of the word must be in the scope of an operator with a particular semantic property. In the figure we simply write *downward-entailing-operator* as a shorthand for a logical specification. A precise version for this type of collocational requirement can be found in Richter/Soehn (2006).

The lexical specifications of COLL values are complemented with a general Licensing Principle. This principle ensures that in each utterance, the collocational requirements of the lexical items that occur in the utterance are satisfied.

Our brief sketch of the collocational module shows that the information included in CoDII can be incorporated into the grammar architecture of HPSG. The better we understand the distributional patterns of bound words and polarity items the more adequately we can state the collocational constraints in a formal grammar framework. It should be clear that the analysis of bound words carries over to collocations with free words such as *take a shower*. A collocation module is a first step towards a formal theory that embodies both a constructional and a collocational perspective and thus, stands a chance to model the implicit linguistic knowledge of native speakers, including knowledge of idiosyncrasies.

4 Conclusion

We presented the Collection of Distributionally Idiosyncratic Items (CoDII), an electronic resource which collects and presents different types of lexical items that exhibit distributional idiosyncrasies. It is a characteristic feature of CoDII that it is open for the inclusion of new subcollections. It is also dynamic on the level of the items in the different collections: Not only may items be added but also new corpus evidence, which can broaden the empirical documentation and, as a consequence, change the theoretical categorization of the items. The flexibility of an electronic resource provides added value to linguists by the various search functions in the system which can be used when researching empirical

evidence or counter-evidence to theoretical claims about universal properties of idiomatic expressions or polarity items. We also illustrated how the corpus evidence collected in CoDII can highlight distribution patterns that went unnoticed before and might lead to new generalizations on the behavior of the recorded classes of items. Most importantly, we believe that CoDII emphasizes the need for a comprehensive study of collocational patterns in language between mere statistical tendencies and phenomena that have acquired the status of grammatical facts that are subject to categorical grammaticality judgments and should consequently also be subject to grammatical description in formal grammar frameworks.

Bibliography

Belletti, A. (ed.) (2004): *Structure and Beyond. The Cartography of Syntactic Structures. Vol. 3.* Oxford/New York.

Bernal, E./De Cesaris, J. (eds.) (2008): *Proceedings of the XIII Euralex International Congress.* Barcelona.

Beyssade, C./Bonami, O./Cabredo Hofherr, P. C./Corblin, F. (eds.) (2003): *Empirical Issues in Formal Syntax and Semantics.* Paris.

Burger, H./Dobrovol'skij, D./Kühn, P./Norrick, N. R. (eds.) (2007): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin/New York.

Chierchia, G. (2004): "Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface." In: Belletti (ed.); 39-103.

Chomsky, N. (1981): *Lectures on Government and Binding.* Dordrecht.

Di Sciullo, A.-M./Williams, E. (1987): *On the Definition of Word.* Cambridge, Mass.

Dobrovol'skij, D. (1988): *Phraseologie als Objekt der Universallinguistik.* Leipzig.

Dobrovol'skij, D. (1989): "Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse." In: Fleischer et al. (eds.) (1989); 57-78.

Dobrovol'skij, D./Piirainen, E. (1994): "Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative", in: *Folia Linguistica 27, 3-4*; 449-473.

Ernst, Th. (2005): *On Speaker-Oriented Adverbs as Positive Polarity Items*. Electronic Poster for the Workshop: Polarity From Different Perspectives, New York University, 11.–13.03.2005. *www.nyu.edu/gsas/dept/lingu/events/polarity/posters/ernst.pdf*.

Everaert, M./van der Linden, E.-J./Schenk, A./Schreuder, R. (eds.) (1995): *Idioms. Structural and Psychological Perspectives.* Hillsdale.

Featherston, S./Sternefeld, W. (eds.) (2007): *Roots: Linguistics in Search of its Evidential Base*. Berlin.

Fillmore, Ch./Kay, P./O'Connor, M. (1988): "Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone", in: *Language 64*; 501-538.

Fintel, K. (1999): "NPI-Licensing, Strawson-Entailment, and Context-Dependency", in: *Journal of Semantics 16*; 97-148.

Fleischer, W. (1997): *Phraseologie der deutschen Gegenwartssprache.* Tübingen.

Fleischer, W./Große, R./Lerchner, G. (eds.) (1989): *Beiträge zur Erforschung der deutschen Sprache. Vol. 9.* Leipzig.

Ginzburg, J./Sag, I. A. (2000): *Interrogative Investigations. The Form, Meaning, and Use of English Interrogatives.* CSLI Publications.

Hamm, F./Hinrichs, E. W. (eds.) (1997): *Plurality and Quantification.* Dordrecht.

Hoeksema, J. (1997): *Corpus Study of Negative Polarity Items*, in *IV-V Jornades de corpus linguistics 1996-1997. http://odur.let.rug.nl/$\sim$hoeksema/docs/barcelona.html*.
Horn, L./Ward, G. (eds.) (2004): *The Handbook of Pragmatics*. Oxford.
Israel, M. (2004): "The Pragmatics of Polarity." In: Horn/Ward (eds.) (2004); 701-723.
Kadmon, N./Landman, F. (1993): 'Any', in: *Linguistics and Philosophy 16*; 353-422.
Krenn, B. (1999): "The Usual Suspects. Data-Oriented Models for Identification and Representation of Lexical Collocations." *Saarbrücken Dissertations in Computational Linguistics and Language Technology, Vol 7*. Saarbrücken.
Krifka, M. (1995): "The Semantics and Pragmatics of Weak and Strong Polarity Items", in: *Linguistic Analysis 25*; 209-257.
Kuiper, K./McCann, H./Quinn, H./Aitchison, Th./van der Veer, K. (2003): *Syntactically Annotated Idiom Database (SAID) v.1.* Documentation to a LDC resource.
Kürschner, W. (1983): *Studien zur Negation im Deutschen.* Tübingen.
Lichte, T. (2005): *Korpusbasierte Acquirierung negativ-polärer Elemente.* Magisterarbeit, Universität Tübingen. Tübingen.
Lichte, T./Soehn, J.-Ph. (2007): "The Retrieval and Classification of Negative Polarity Items using Statistical Profiles". In: Featherston/Sternefeld (eds.) (2007); 249-266.
McCawley, J. D. (1981): "The Syntax and Semantics of English Relative Clauses", in: *Lingua 53*; 99-149.
Moon, R. (1998): *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford.
Moon, R. (2007): "Corpus Linguistic Approaches with German Corpora." In: Burger et al. (eds.) (2007); 1045-1059.
Müller, S. (ed.) (2006): *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*. Stanford. *http://cslipublications.stanford.edu/HPSG/7/toc.shtml*.
Nunberg, G./Sag, I.A./Wasow, Th. (1994): "Idioms", in: *Language 70*; 491-538.
Os, Ch. van (1989): *Aspekte der Intensivierung im Deutschen.* Tübingen.
Pollard, C./Sag, I. A. (1994): *Head-Driven Phrase Structure Grammar.* Chicago/London.
Richter, F. (2004): *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar.* Phil. Diss. Universität Tübingen, Tübingen.
Richter, F./Sailer, M. (2003): "Cranberry Words in Formal Grammar." In: Beyssade et al. (eds.) (2003); 155-171.
Richter, F./Soehn, J.-Ph. (2006): "Braucht niemanden zu scheren: A Survey of NPI Licensing in German." In: Müller (ed.) (2006); 421-440. *http://cslipublications.stanford.edu/HPSG/7/richter-soehn.pdf*.
Riehemann, S. Z. (2001): *A Constructional Approach to Idioms and Word Formation.* Phil. Diss. Stanford University, Stanford. *http://doors.stanford.edu/~sr/sr-diss.ps.gz*.
Sailer, M. (2003): *Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar.* Phil. Diss. Universität Tübingen.
Sailer, M. (2004): "Distributionsidiosynkrasien: Korpuslinguistische Erfassung und grammatiktheoretische Deutung." In: Steyer (ed.) (2004); 194-221
Sailer, M./Trawiński, B. (2006): "The Collection of Distributionally Idiosyncratic Items: A Multilingual Resource for Linguistic Research." In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy; 471-474.
Schenk, A. (1995): "The Syntactic Behavior of Idioms". In: Everaert et al. (eds.) (1995); 253-271.
Sinclair, J. (1991): *Corpus, Concordance, Collocation.* Oxford.
Sinclair, J. (2004): *Trust the Text. Language, Corpus and Discourse.* London/New York.
Soehn, J.-Ph. (2006): *Über Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG.* Frankfurt a. M..
Steyer, K. (ed.) (2004): *Wortverbindungen – mehr oder weniger fest.* Berlin/New York.

Trawiński, B./Soehn, J.-Ph. (2008): "A Multilingual Database of Polarity Items." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.*
Trawiński, B./Soehn, J.-Ph./Sailer, M./Richter, F. (2008a): "A Multilingual Electronic Database of Distributionally Idiosyncratic Items." In: Bernal/De Cesaris (eds.) (2008); 1445-1451.
Trawiński, B./Sailer, M./Soehn, J.-Ph./Lemnitzer, L./Richter, F. (2008b): "Cranberry Expressions in English and in German." In: *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco*; 35-38.
Welte, W. (1978): *Negationslinguistik. Ansätze zur Beschreibung und Erklärung von Aspekten der Negation im Englischen.* München.
van der Wouden, T. (1997): *Negative Contexts. Collocation, Polarity and Multiple Negation.* London.
Zwarts, F. (1997): "Three Types of Polarity." In: Hamm/Hinrichs (eds.) (1997); 177-237.

List of Abbreviations

BW	bound word
CoDII	Collection of Distributionally Idiosyncratic Items
DII	distributionally idiosyncratic item
HPSG	Head-driven Phrase Structure Grammar
iff	if and only if
MWE	multi-word expression
N	noun
NPI	negative polarity item
PI	polarity item
PPI	positive polarity item
VP	verb phrase
XML	Extensible Markup Language

Frank Richter
Seminar für Sprachwissenschaft
Abteilung Allgemeine Sprachwissenschaft / Computerlinguistik
Universität Tübingen
Wilhelmstraße 19
72074 Tübingen
Deutschland
fr@sfs.uni-tuebingen.de

Manfred Sailer
Seminar für Englische Philologie
Abteilung Linguistik
Universität Göttingen
Käte-Hamburger-Weg 3
37073 Göttingen
Deutschland
manfred.sailer@phil.uni-goettingen.de

Beata Trawiński
Institut für Slawistik
Universität Wien
Spitalgasse 2, Hof 3
1090 Wien
Österreich
beata.trawinski@univie.ac.at

Stichwortverzeichnis/Index