

Lexical Bundles. Eine korpuslinguistische Untersuchung

Lexical Bundles. A Corpus-linguistic Research

Masterarbeit

zur Erlangung des akademischen Grades Master of Arts (M.A.)
im Fach Linguistik

Humboldt-Universität zu Berlin
Philosophische Fakultät II
Institut für deutsche Sprache und Linguistik

eingereicht von: Carolin Odebrecht

ErstgutachterIn: Prof. Dr. Anke Lüdeling

ZweitgutachterIn: Prof. Dr. Bernd Pompino-Marschall

Berlin, den 29.09.2011

Inhaltsverzeichnis

1	Einleitung	8
2	Theoretische Grundlagen und Einordnung	10
3	Lexical Bundles	12
3.1	Übersicht der bereits untersuchten Register	12
3.2	Untersuchungsgegenstand	14
3.2.1	Definition	14
3.2.2	Festlegung der Frequenz von Lexical Bundles	16
3.3	Strukturelle Eigenschaften von Lexical Bundles	17
3.3.1	Taxonomien im Vergleich	17
3.3.2	Erkenntnisse aus der strukturellen Klassifikation	21
3.3.3	Zusammenhang zwischen Wortarten und struktureller Klassifikation	22
3.4	Funktionale Eigenschaften von Lexical Bundles	23
3.4.1	Taxonomien im Vergleich	23
3.4.2	Strittige Klassifikationen	26
4	Verwandte Phänomene	29
5	Korpusstudie	32
5.1	Korpus	32
5.2	Eigenschaften der Register Politikforum und Plenarprotokoll	33
5.3	Datenerhebung	35
5.3.1	Allgemeine Richtlinien zur Identifikation	35
5.3.2	Lokale Wiederholungen	36
5.3.3	Unterschiedliche Textebenen	37
6	Ergebnisse der Studie	39
6.1	Allgemeine Diskussion der Ergebnisse	39
6.2	Strukturelle Klassifikation	40
6.3	Funktionale Klassifikation	46
6.3.1	Mögliche Klassifikationen	46
6.3.2	Topik und Fokus	49
6.3.3	Redewendungen	53
7	Schlussfolgerung und Ausblick	55

8	Quellen- und Literaturverzeichnis	59
8.1	Quellen	59
8.2	Software	59
8.3	Literatur	59
A	Funktionale Klassifikation nach Biber, Conrad und Cortes (2004)	64
	Danksagung	66

Tabellenverzeichnis

3.1	Übersicht der Studien mit Angabe von Korpus und Register. Diese Übersicht gibt die jeweils neu in die Untersuchung integrierten Teilkorpora sowie deren Register an.	13
3.2	Vergleich der Frequenzgrenzen mit Angabe der jeweiligen Wortanzahl aller untersuchter Korpora nach den Studien von Biber und Kollegen.	16
3.3	Übersicht der strukturellen Klassifikation nach Biber et al. (1999, 996). . .	17
3.4	Übersicht der strukturellen Klassifikation nach Conrad und Biber (2004, 64).	19
3.5	Übersicht der strukturellen Klassifikation nach Biber (2004, 172).	19
3.6	Übersicht der strukturellen Klassifikation nach Cortes (2004, 405, 407). . .	20
3.7	Übersicht der funktionalen Klassifikation nach Biber, Conrad und Cortes (2003, 80).	24
5.1	Übersicht über die Anzahl der Wörter und Texte der Korpora <i>Politikforum</i> und <i>Plenarprotokoll</i>	33
5.2	Vergleich zwischen Videoaufzeichnung und der Bearbeitung der Plenarreden durch den Stenografischen Dienst des Deutschen Bundestages. Quelle: Stenografischer Bericht 17121 17. Legislaturperiode, 121. Sitzung, und Videoaufzeichnung der 17. Legislaturperiode 121. Sitzung Berlin.	34
5.3	Beispiel für die Art der Datenerhebung, Protokoll 17. Legislaturperiode, 34. Sitzung Berlin.	36
5.4	Beispiele für die Moderation der Plenarreden durch den Bundestagspräsidenten. Alle Zitate aus Protokoll 17. Wahlperiode 119. Sitzung Berlin. . .	38
6.1	Übersicht der Datengrundlage für beide Korpora mit Angabe der lokalen Wiederholung, der strukturellen Ausschlüsse und der Lexical Bundles. . .	39
6.2	Übersicht der gerundeten prozentualen Anteile der exemplarisch ausgesuchten Wortarten in beiden Korpora.	39
6.3	Beispiele verschiedener Strukturen von Lexical Bundles in beiden Korpora.	40
6.4	Verteilung der strukturellen Typen in beiden Registern. Hier werden die vollen Vorkommen der einzelnen Wortketten angegeben.	42
6.5	Darstellung der verschiedenen automatisch annotierten Ebenen POS und Lemma.	45
6.6	Übersicht über mögliche <i>Referential Bundles</i> im Deutschen aus beiden Korpora.	46
6.7	Übersicht über mögliche <i>Stance Bundles</i> im Deutschen aus beiden Korpora.	47

6.8	Übersicht über mögliche <i>Discourse Organizers</i> im Deutschen aus beiden Korpora.	49
6.9	Fokuskonstruktion im Spanischen. Beispiel aus Tracy-Ventura, Cortes und Biber (2007, 225).	53
A.1	Übersicht der funktionalen Klassifikation nach Biber, Conrad und Cortes (2004, 384-388).	64

Abkürzungsverzeichnis

ADJA	attribuierendes Adjektiv nach STTS
ADJD	adverbiales oder prädikatives Adjektiv nach STTS
ADV	Adverb nach STTS
AP	Adjektivphrase
APPR	Präposition nach STTS
ART	bestimmter oder unbestimmter Artikel nach STTS
Aux	Auxiliar, Hilfsverb
Konnek.	Konnektor
lex.V	lexikalisches Verb
NE	Nomen, Eigennamen nach STTS
NEG	Negationspartikel nach STTS
NN	Nomen nach STTS
PAV	Pronominaladverb nach STTS
PDAT	attribuierendes Demonstrativpronomen nach STTS
PDS	substituierendes Demonstrativpronomen nach STTS
PIS	substituierendes Indefinitpronomen nach STTS
POS	Part-of-Speech
PP	Präpositionalphrase
PPER	Personalpronomen, nach STTS
PPER	irreflexives Personalpronomen nach STTS
PPOSAT	attribuierendes Possessivpronomen nach STTS
PRELS	substituierendes Personalpronomen nach STTS
PRF	reflexives Personalpronomen nach STTS
PWAV	adverbiales Interrogativ- oder Relativpronomen nach STTS
STTS	Stuttgart Tübingen Tagset
V	Verb
VAFIN	finites Hilfsverb nach STTS
VAINF	infinites Hilfsverb nach STTS
VMFIN	finites Modelverb nach STTS
VMINF	infinites Modelverb nach STTS
VP	Verbalphrase
VVFIN	finites Vollverb nach STTS
VVIMP	Vollverb im Imperativ nach STTS
VVINFINF	infinites Vollverb nach STTS
VVIZU	Infinitiv mit <i>zu</i> nach STTS
VVPP	Partizip Perfekt nach STTS

1 Einleitung

Können mit Hilfe rekurrenter Wortketten Register hinsichtlich ihrer kommunikativen Funktionen charakterisiert werden? Welchen Erkenntnisse können hinsichtlich des Aufbaus und des Gebrauchs von Sprache aus einer frequenzbasierten Untersuchung dieser Ketten gewonnen werden?

Dem übergeordnet ist die Frage, wie Sprecher Sprache nutzen, wie sie produziert und rezipiert wird. Sind es feste, vorgefertigte Phrasen, auf deren Grundlage Sprecher Sprache planen oder verarbeiten?

Lexical Bundles¹ sind nach den Studien von Biber und Kollegen solche rekurrenten Wortketten, durch deren unterschiedliche, strukturelle und funktionale Eigenschaften spezifische Gebrauchsmuster zwischen Registern akademischer, englischsprachiger Texte aufgezeigt werden. Untersucht werden soll, ob dieses Phänomen auch in deutschsprachigen Texten in gleicher Weise identifiziert werden kann und in wie weit die strukturellen und funktionalen Systematisierungen auf diese Daten übertragbar sind und somit ebenfalls Register im Deutschen genauer untersucht werden können. Diese Arbeit versucht demnach durch eine frequenzbasierte Untersuchung von deutschsprachigen Daten mit Hilfe korpuslinguistischer Methoden das Konzept *Lexical Bundles* zu überprüfen. Untersucht werden soll weiterhin, ob der hochfrequente Wiedergebrauch von Ketten lexikalischer Einheiten in bestimmten Registern variiert. Weisen dann diese Ketten bestimmte strukturelle und funktionale Eigenschaften ganz allgemein und für jeweils ein bestimmtes Register auf? Biber et al. (1999) und deren Folgestudien führen den Begriff *Lexical Bundle* für diese rekurrenten Wortsequenzen ein. Der definitorisch-theoretische Ansatz und die Methode der Auswertung sollen für diese Arbeit übernommen und an deutschen Texten - Einträgen des Forums *abgeordnetenwatch.de* und den *Stenografischen Berichten des Bundestages* - überprüft werden. Folgende Fragen werden in dieser Arbeit dazu beantwortet: Können nach der Methode von Biber und Kollegen Lexical Bundles in deutschen Texten nachgewiesen werden? Sind die Methode und das Konzept auf deutsche Daten übertragbar? Besitzen diese Daten auch das gleiche Erklärungspotenzial hinsichtlich der Eigenschaften bestimmter Register? Sind die Ergebnisse mit den Vorgängerstudien vergleichbar?

Dazu werden im ersten Kapitel der Forschungsbereich und die Methode eingeordnet. Darauf folgt ein Überblick über die Lexical Bundles-Studien von Biber und Kollegen und deren ausführliche Diskussion hinsichtlich der Definition sowie deren struktureller und funktionaler Taxonomien. Ein kurzer Abriss über verwandte, linguistische Phänomene in Kapitel Vier wird dem Konzept der Lexical Bundles gegenübergestellt. Im fünften Kapitel wird das für diese Studie verwendete Korpus, dessen Eigenschaften und Beson-

¹Dieser Begriff wird hier weiterhin nach Biber et al. (1999) verwendet, da noch keine gängige deutsche Entsprechung beziehungsweise Übersetzung in der Forschung etabliert ist.

derheiten sowie die konkrete Datenerhebung vorgestellt. Weiterhin werden die Ergebnisse allgemein präsentiert und hinsichtlich ihrer linguistischen Auswertung in Bezug auf die strukturelle und funktionale Klassifikation diskutiert. Am Schluss sollen die Ergebnisse und Schlussfolgerungen zusammengefasst werden.

2 Theoretische Grundlagen und Einordnung

Lexical Bundles tragen nach Biber et al. (1999) zur Unterscheidung und Klassifikation von Registern bei (Biber et al. 1999, 992). *Register* ist ein linguistischer Begriff, der in vielen wissenschaftlichen Disziplinen Anwendung findet. In dieser Arbeit wird er nach Biber (1994) wie folgt verstanden:

'I use the term *register* in this paper, as it is used in this book, as a general cover term for all language varieties associated with different situations and purposes (Biber 1994, 32).'

Der Begriff *Register* umfasst demnach den situativen Kontext einer Äußerung und die genutzten linguistischen Mittel sowie deren Beziehung untereinander (Biber und Conrad 2009, 5-6). Und weiter ist wichtig:

'[...] that linguistic features are always functional when considered from a register perspective. That is, linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register (Biber und Conrad 2009, 6).'

So sind Registerstudien gebrauchsbasiert und untersuchen realsprachliche Daten aller Art. Herausgefunden werden soll mit Hilfe einer korpusbasierten Methode, welche typischen Merkmale ein bestimmtes Register besitzt und wie es sich so von anderen unterscheidet (vgl. Biber, Conrad und Reppen 1998, 3). So gibt es mehrere Komponenten, die in einer Registerstudie zu betrachten sind: Die situativen Eigenschaften eines Registers, dessen linguistische Merkmale sowie die funktionale oder konventionale Beziehung zwischen diesen beiden (Biber 1994, 33). Registerstudien fokussieren meist auf die lexikogrammatistischen Merkmale einer Sprache, die systematisch unter anderem je nach Modus, Kommunikationszielen und Kontext in den einzelnen Registern variieren (Biber, Connor und Upton 2007, 8). Weiterhin wird deutlich, wie methodologisch Register und deren Eigenschaften untersucht werden: Der situative Gebrauch von linguistischen Merkmalen soll empirisch untersucht werden. Lexical Bundles sind nach Biber (et al. 1999) solche linguistischen Merkmale, deren funktionale Eigenschaften den kommunikativen Zweck des jeweiligen Registers realisieren können (vgl. ebd., 992). Mit Hilfe korpuslinguistischer Methoden lässt sich dieser Forschungsansatz umsetzen:

'Als Korpuslinguistik bezeichnet man die Beschreibung von Äußerungen natürlicher Sprachen, ihre Elemente und Strukturen, und darauf aufbauende Theorienbildung auf Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind [...] (Lemnitzer und Zinsmeister 2006, 9).'

Die Ergebnisse dieser Methode sollen überprüfbar und reproduzierbar sein, indem unter anderem statistische Aussagen über Regularitäten und Tendenzen getroffen werden (Lemnitzer und Zinsmeister 2006, 10). Die Lexical Bundles-Forschung setzt einen rein quantitativen Ansatz zur Datenerhebung ohne linguistische Theorienbildung vor. Das Korpus beziehungsweise die Korpora, die jeweils die digitalisierten Texte der Register beinhalten, sind bei diesem Ansatz nicht weiter aufbereitet oder interpretiert beziehungsweise annotiert (Biber et al. 1999, 990-993; vgl. Lemnitzer und Zinsmeister 2006, 33). Daher gilt:

'[...], a large amount of effort in corpus-based studies is devoted to explaining and exemplifying quantitative patterns (Biber, Conrad und Reppen 1998, 9).'

Durch die Analyse des situativen Gebrauchs werden Register in Hinblick auf den spezifischen Diskurs und die Variation von Sprache untersucht (Biber, Connor und Upton 2007, 3). Linguistische Theorien sind demnach nicht unabhängig von den Ergebnissen: Beobachtungen führen zu Hypothesen, Hypothesen führen zu Generalisierung und dies führt zu theoretischen Aussagen/Schlussfolgerungen (Tognini-Bonelli 2001, 75). Die Lexical Bundles-Studien sind eine Möglichkeit der Registerforschung, die hier an deutschen Daten reproduziert und überprüft werden soll.

3 Lexical Bundles

3.1 Übersicht der bereits untersuchten Register

Der Begriff *Lexical Bundle* und die spezielle Methode zur Erhebung der Daten werden von Biber et al. (1999) mit der *Longman Grammar of Spoken and Written English*-Studie eingeführt. Mehrere weitere Untersuchungen an vorwiegend englischsprachigen, akademischen Registern nach gleicher Methode folgen. Hier soll ein Überblick über die relevanten Studien zu diesem Thema hinsichtlich der untersuchten Korpora und Fragestellungen gegeben werden, um im Weiteren gezielt darauf referieren zu können.

Untersucht werden in Biber et al. (1999) zwei Teilkorpora des LSWE-Korpus, die *Gesprächsaufzeichnungen* (CONV) und die *akademischen Prosa* (ACAD), deren Texte beide in einem akademischen Rahmen geäußert beziehungsweise veröffentlicht wurden (Biber et al. 1999, 990-993).² Mit dieser Studie sind erstmals Richtlinien zur Identifikation des Phänomens und eine strukturelle sowie in Ansätzen eine funktionale Klassifikation entwickelt worden. Die Studie kann zeigen, dass die Gebrauchsmuster bestimmter struktureller wie funktionaler Bundles Register unterscheiden und charakterisieren können (vgl. Biber et al. 1999, 996-997, 1003, 1007, 1015, 1018).³

Biber (2004) und Biber, Conrad und Cortes (2004) integrieren die Studien zu den akademischen Registern CONV und ACAD in ihre Untersuchungen zu Lexical Bundles der akademischen Register *Textbuch* (TEXTB) und *Unterrichtsgespräch* (CLASS), die aus dem T2K-SWAL-Korpus zusammengestellt worden sind, und stellen ebenfalls für jedes Register unterschiedliche Gebrauchsmuster der verschiedenen Typen von Bundles, die funktional zu erklären sind, fest (Biber 2004, 167,171, 175; Biber, Conrad und Cortes 2004, 374-375).

Biber (2006) und Biber und Barbieri (2007) diskutieren detaillierter die Ergebnisse aus Biber, Conrad und Cortes (2004) und erweitern das Korpus um mehrere mündliche Register aus dem T2K-SWAL-Korpus, die aus einem universitären Umfeld erhoben wurden - *Klassenmanagement*, *Sprechstundenaufzeichnungen*, *Studiengruppengesprächen* und *Dienstleistung* - sowie um mehrere schriftliche Register - *institutionelle Texte* der Universitäten, eine Aufschlüsselung der Lehrbuchtexte nach akademischen Disziplinen und *schriftliches Klassenmanagement*. Hier werden alle Ergebnisse aus den diversen mündli-

²Die von Biber und Kollegen durchgängig genutzten Abkürzungen der einzelnen Register werden hier ebenfalls übernommen.

³Biber und Conrad (1999), Biber, Conrad und Cortes (2003) und Conrad und Biber (2004) greifen die *Longman*-Studie mit den Registern *Gesprächsaufzeichnungen* und *akademischer Prosa* und deren Ergebnisse nochmals auf und diskutieren unter anderem vertieft die möglichen Klassifikationen und theoretischen Schlüsse, die es ermöglichen sollen, Lexical Bundles in allen möglichen Registern zu analysieren und zu klassifizieren.

chen und schriftlichen Registern verglichen und diskutiert (vgl. Biber 2006, 135, 148, 165; Biber und Barbieri 2007, 266-267, 279-282). Cortes (2004) untersucht mit derselben Methode die Verteilung von Lexical Bundles in akademischer Prosa von Studierenden und Akademikern aus *Biologie* (BIO) und *Geschichtswissenschaft* (HIST) (ebd., 401-402). Dieser Ansatz betrachtet zum Teil den Erwerb von Lexical Bundles in der akademischen Prosa und erweitert somit die Fragestellung und den Untersuchungsansatz zu den Lexical Bundles. Im Vergleich zu den oben genannten Studien, denen Sprachdaten von Sprechern des amerikanischen und britischen Englischs zugrunde liegen, wenden Tracy-Ventura, Cortes und Biber (2007) die Methode an unterschiedlichen Registern in Spanisch an - *soziolinguistische Interviews* (INTERV) und *akademische Prosa* (ebd., 218). Nachfolgend werden alle Studien mit ihren Korpora sowie den untersuchten Registern kurz zusammengefasst:

Register	Korpus	Studien
ACAD, CONV	LSWE	Biber et al. 1999, Biber und Conrad 1999 Biber, Conrad und Cortes 2003, Conrad und Biber 2004
CLASS, TEXTB	T2K-SWAL	Biber 2004, Biber, Conrad und Cortes 2004
BIO, HIST	CC, Studierende	Cortes 2004
FULL RANGE	T2K-SWAL	Biber 2006, Biber und Barbieri 2007
INTERV, ACAD	Habla, Zeitschrift	Tracy-Ventura, Cortes und Biber 2007

Tabelle 3.1: Übersicht der Studien mit Angabe von Korpus und Register. Diese Übersicht gibt die jeweils neu in die Untersuchung integrierten Teilkorpora sowie deren Register an.

Aus der Tabelle 3.1 geht hervor, dass die Studien vorwiegend auf mündlichen sowie schriftlichen Registern aus einem akademischen/universitären Umfeld basieren. Weiterhin werden die Ergebnisse aus Biber et al. (1999) in die nachfolgenden Studien für einen Vergleich und/oder einer weitumfassenderen Diskussion integriert.⁴ Ganz allgemein bindet jede Studie jeweils die Vorgängerstudien mit ihren Ergebnissen in ihre Diskussion mit ein. Hierbei sind die immer wieder neu erstellten, strukturellen sowie funktionalen Klassifikationen interessant. Dieser wissenschaftliche Diskurs - angefangen mit einer definitorischen Grundlage über mögliche Taxonomien bis hin zu Schlussfolgerungen über Register unterscheidende Funktionen der Lexical Bundles - wird in dem nachfolgenden Kapiteln dargestellt und diskutiert.

⁴Weitere Studien zu Lexical Bundles in akademischen Registern sind unter anderem von Hyland (2008), Strunkyte und Jurkunaite (2008), Pang (2010) und Jalali und Ghayoomi (2010) durchgeführt worden. Alle Studien beziehen sich grundlegend auf Biber et al. (1999) und deren Nachfolgestudien. Untersucht werden in allen Fällen akademische Texte von L1 und/oder L2 Sprechern des Englischen, dabei werden die grundlegenden Taxonomien - zum Teil mit Änderungen - übernommen. Deshalb wird in dieser Arbeit die weitere Diskussion auf die Studien von Biber und Kollegen beschränkt.

3.2 Untersuchungsgegenstand

3.2.1 Definition

In Biber et al. (1999) wird der Begriff und der konkrete Phänomenbereich der Lexical Bundles erstmals eingeführt:

'Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse (Biber et al. 1999, 990).'

Diese Definition⁵ beinhaltet keine Einschränkungen bezüglich Struktur oder Funktion eines Lexical Bundle. Im Gegenteil, sie gibt in erster Linie keine weiteren definitorischen Vorgaben außer, dass es sich bei einem Lexical Bundle um mehrere, hintereinander folgende 'word forms' handeln muss, die immer wieder im Diskurs in exakt der gleichen Reihenfolge vorkommen. Darüber hinaus ist ein gebrauchsbasierter Ansatz zur Untersuchung der Lexical Bundles in der Definition mit 'in natural discourse' verankert. So ist die Methode zur Datenerhebung rein quantitativ und ohne theoretische Vorannahmen. Die Analyse der Lexical Bundles erfolgt *bottom-up*, indem erst die N-Wortketten erstellt werden, danach folgt die linguistische Auswertung mit einer strukturellen und funktionalen Klassifikation der Daten (vgl. Biber, Connor und Upton 2007, 14).

Mit 'word forms' ist das orthografische Wort als eine linguistische Einheit eines Lexical Bundles gemeint (vgl. Biber et al. 1999, 990). Klitische Verbindungen wie *don't* oder *it's* werden hier ebenfalls als ein orthografisches Wort gewertet (vgl. Biber et al. 1999, 990). Die Untersuchung setzt an der Textebene des jeweiligen Korpus an. So kann potentiell jedes Wort beziehungsweise jede Wortform Bestandteil eines Lexical Bundles sein. Auf Grundlage der Textebene des Korpus wird errechnet, welche seriell festen Mehr-Wortketten immer wieder beziehungsweise entsprechend n-mal vorkommen. Conrad und Biber (2004) geben hier ein Beispiel:

'the frequency analysis for
frequency analysis for the
analysis for the lexical
(Conrad und Biber 2004, 60)'

So werden alle realsprachlichen Texte Wortform für Wortform in ihrer dort festen Reihenfolge nach N-Wortketten durchsucht und die entsprechend häufigen Sequenzen herausgerechnet. Die Frequenz ist demnach ausschlaggebendes Kriterium für dieses linguistische Phänomen (Conrad und Biber 2004, 58). Die Anzahl der in einem Bundle enthaltenden Wörter wird nicht definitorisch festgelegt. Biber et al. (1999) beziehen in ihre Analyse Drei- bis Sechs-Wortketten ein, wobei die Vier-Wortketten vordergründlich

⁵Die hier dargelegte Definition wird in den nachfolgenden Studien übernommen, vgl. u.a. Biber und Conrad (1999), Biber Conrad und Cortes (2003), Conrad und Biber (2004), Biber, Conrad und Cortes (2004), Biber (2004), Biber (2006), Biber und Barbieri (2007), Tracy-Ventura, Cortes und Biber (2007).

betrachtet werden (ebd., 990). Motiviert wird dieses Vorgehen dadurch, dass viele Zwei-Wortketten in Drei-Wortketten enthalten sind und diese zu mehreren, unterschiedlichen Vier-Wortketten erweitert werden können (ebd., 990-995). Die darauf folgenden Studien legen den Schwerpunkt der Untersuchung ebenfalls auf Vier-Wortketten. Damit wurde eine Art Konvention hinsichtlich eines Untersuchungsschwerpunktes geschaffen, die aber die Möglichkeit einer Betrachtung von Drei-, Fünf- oder allgemein Mehr-Wortketten nicht ausschließt.

Lexical Bundles müssen nicht zwingend komplette Phrasen oder Sätze bilden, da die Identifikation strukturell *blind* erfolgt. So werden die Wortformketten ohne Berücksichtigung von strukturellen Abhängigkeiten untersucht, mit allen möglichen Linearisierungen und allen der Sprache entsprechenden Flexionsendungen.

Weiterhin gibt diese Definition keinen direkten Aufschluss über Status, Aufgaben und Funktionen dieser linguistischen Einheiten, sondern legt vorwiegend die Identifikationsrichtlinien fest und gibt einen methodologischen Rahmen vor. Die konkreten Aufgaben und Funktionen ergeben sich erst aus der Analyse der so erhobenen Daten. Damit ist für eine weitere Untersuchung und Diskussion der Daten ein induktiver Ansatz zwingend erforderlich (vgl. Biber und Barbieri 2007, 270). Implizit wird folgendes deutlich: Durch die auf Frequenz basierende Erhebung sind Lexical Bundles nicht idiomatisch im Sinne von festen, nicht kompositionellen Redewendungen wie *kick the bucket*, da diese eher rar gebraucht und so nicht mit den Richtlinien erfasst werden (Biber, Conrad und Cortes 2004, 376-378). Gebräuchlichere Redewendungen sind dennoch erfassbar.

Nach der quantitativen Datenerhebung geben Biber et al. (1999) eine qualitative Einschränkung zur Identifizierung von Lexical Bundles vor: Wortketten, die direkt mit dem Thema eines Textes im Zusammenhang stehen, sogenannte *lokale Wiederholungen*, werden nicht weiter betrachtet (ebd., 991). Diese spiegeln thematische Schwerpunkte eines Textes, aber nicht den allgemeinen Gebrauch in einem Register wieder, da ein Sprecher/Autor seinem Thema und dem Kontext verbunden ist. Aus diesem Grund würde er sich wiederholen. *Lokale Wiederholungen*, die aus einer Interaktion im Diskurs resultieren, werden ebenfalls nicht als Lexical Bundle definiert (vgl. Biber et al. 1999, 998). Diese ausgeschlossenen Daten stellen eher individuelle oder auf bestimmte Themen beziehungsweise Gesprächskontexte spezialisierte Wortketten dar, die keinen Erkenntnisgewinn für den Gebrauch von Lexical Bundles in bestimmten Registern versprechen (vgl. u.a. Biber et al. 1999, 990-991).⁶ Dennoch wird nicht ganz klar, wie nah oder wie weit ein potenzielles Bundle vom Thema entfernt sein kann beziehungsweise was hier genau unter *lokal* verstanden wird. Diese Richtlinie erscheint nicht ausreichend konkret und es folgen keine illustrierenden Beispiele.⁷ Eine strukturelle Restriktion bei der Identifizierung der Lexical Bundles legen Biber et al. (1999) mit dem Ausschluss satzübergreifender Sequenzen fest (ebd., 993), was die zu untersuchende Datenmenge weiter eingrenzt. Diese kurze Diskussion der Definition und der Einschränkungen macht deutlich, dass sich die Richtlinien zur Identifikation zum Teil nach den zur Verfügung stehenden Sprachdaten richten.

⁶Im Kapitel 3.2 werden alle Identifikationsrichtlinien vorgestellt und ihre Varianten verglichen und diskutiert.

⁷In wie weit dies auf Grundlage deutscher Daten reproduzierbar ist, siehe Kapitel 6.

Allgemein werden die Lexical Bundles als *building blocks of discourse* beschrieben, bei denen es herauszufinden gilt, in welchen Kontexten und Registern sie welche Aufgaben erfüllen (Conrad und Biber 2004, 58):

‘Our hypothesis is that extremely common, fixed sequences of words are used as unanalyzed chunks by speakers and writers, and therefore will have identifiable discourse functions in texts (Conrad und Biber 2004, 58).’

Der psycholinguistische Status dieser ‘unanalyzed chunks’ kann mit dieser Methode nicht direkt diskutiert werden. Allein die hohe Frequenz dieser Phänomene lässt diese Annahme zu (vgl. Biber und Barbieri 2007, 283-284).

Die durch eine rein frequenzielle Erhebung gewonnenen Daten können allein keine Erklärungsansätze oder einen Erkenntnisgewinn aufzeigen. Diese Daten spiegeln den eigentlichen Gebrauch von bestimmten Wortketten wieder, woraus nicht im selben Schritt die jeweilige Funktion ablesbar ist (vgl. Biber, Conrad und Cortes 2004, 376). Daher werden die strukturellen und funktionalen Eigenschaften von Lexical Bundles erst in einem nächsten Schritt betrachtet. Diese induktive Methode zur Erstellung einer Systematik wird in den nachfolgenden Kapiteln noch genauer diskutiert.

3.2.2 Festlegung der Frequenz von Lexical Bundles

Das Prinzip der Datenerhebung wurde schon im vorherigen Kapitel angesprochen. Zu Vergleichszwecken sollen nun die einzelnen konkreten Frequenzen für die Lexical Bundles mit Blick auf die Datenmenge herausgestellt werden:⁸

Studie	Korpusgröße	minimale Frequenz
Biber et al. 1999, Biber und Conrad 1999	12,3 Mio. Wörter	10/1 Mio. Token
Biber 2004	2 Mio. Wörter	20/1 Mio. Token
Biber, Conrad und Cortes 2003	12,3 Mio. Wörter	20/1 Mio. Token
Biber, Conrad und Cortes 2004	2 Mio. Wörter	40/1 Mio. Token
Conrad und Biber 2004	9,2 Mio. Wörter	10/1 Mio. Token
Cortes 2004	2 Mio. Wörter	20/1 Mio. Token
Biber 2006	7,6 Mio. Wörter	40/1 Mio. Token
Biber und Barbieri (2007)	7,6 Mio. Wörter	40/1 Mio. Token
Tracy-Ventura, Cortes und Biber 2007	3,2 Mio. Wörter	30/1 Mio. Token

Tabelle 3.2: Vergleich der Frequenzgrenzen mit Angabe der jeweiligen Wortanzahl aller untersuchter Korpora nach den Studien von Biber und Kollegen.

Einen ersten Richtwert zur Identifikation geben Biber et al. (1999): Ein Vorkommen von mindestens zehnmal pro eine Million Wörter bei einer Korpusgröße von rund sieben Millionen Wörtern und in mindestens fünf verschiedenen Texten eines Registers wird als

⁸Conrad und Biber (2004) diskutieren in ihrer Studie den Britischen Englisch Teil des Korpus *Gesprächsaufzeichnungen* aus Biber et al. (1999) (Conrad und Biber 2004, 59). Die Anzahl der Wörter pro Korpus wurde aus der Summe aller für die jeweilige Untersuchung genutzten Teilkorpora errechnet. Die Angabe ist jeweils gerundet angegeben.

Richtlinie festgelegt (ebd., 992-993). Dieses Vorgehen ist auf eine Art arbiträr und Biber et al. (1999) geben praktische Gründe für diese Festlegung, sowie für die Eingrenzung der zu untersuchenden Datenmenge an (Biber et al. 1999, 992-994). Aus Tabelle 3.2 ist ersichtlich, dass die konkreten Richtlinien zur Identifikation und die Größe der Korpora zwischen den einzelnen Studien variieren. Die Betrachtung der kleineren Teilkorpora wie *Sprechstundenaufzeichnung* mit rund 50.000 Wörtern aus dem T2K-SWAL-Korpus ist im Vergleich zu den anderen Teilkorpora mit weit mehr als 100.000 Wörtern problematisch. Hier soll eine Normalisierung der Zahlen eine Gegenüberstellung der Daten aus den verschiedenen großen Teilkorpora ermöglichen (vgl. Biber 2006, 148; Biber und Barbieri 2007, 268-269). Trotzdem ist eine Streuung der Daten so nicht komplett ausgeschlossen (vgl. dazu Baroni 2009). Die Auswertung der Daten erfolgt statistisch beschreibend. Auf Grundlage der errechneten Frequenzlisten werden Vergleiche zwischen den einzelnen Registern in Bezug auf Verteilung und Vorkommen gezogen. In jedem Fall werden die dargestellten Verhältnisse als Richtwert für die Korpusgröße und die Höhe der Frequenz der nachfolgenden Untersuchung fungieren.

3.3 Strukturelle Eigenschaften von Lexical Bundles

3.3.1 Taxonomien im Vergleich

Die so durchgeführte Datenerhebung erfolgt, wie bereits erwähnt, strukturell *blind*. Die eigentliche linguistische Untersuchung dieser Daten erfolgt erst im Nachhinein. Alle strukturellen Klassifikationen sind in jeder Studie auf Grundlage der Vier-Wortketten erstellt. Im Folgenden werden nicht alle Studien im Detail diskutiert, es wird eine Auswahl der wichtigsten und unterschiedlichsten Studien getroffen.

Nr.	Struktur	Beispiel
1	PPER+lex.VP (+Komplementsatz)	<i>I don't know what</i>
2	PPER/NP (+Aux)+Kopula <i>be</i> (+)	<i>it was in the</i>
3	(Aux+) Verb im Aktiv (+)	<i>have a look at</i>
4	<i>yes-no+wh</i> -Fragenfragment	<i>can I have a</i>
5	(Verb+) <i>wh</i> -Satzfragment	<i>know what I mean</i>
6	NP mit Postmodifiziererfragment	<i>the nature of the</i>
7	Präposition+NP-Fragment	<i>as a result of</i>
8	antizipatorisches <i>it</i> +VP/Adjektiv (+Komplementsatz)	<i>it is possible to</i>
9	Verb im Passiv+PP-Fragment	<i>is based on the</i>
10	(Verb+) <i>that</i> -Satzfragment	<i>should be noted that</i>
11	(Verb/Adjektiv+) <i>to</i> -Satzfragment	<i>are likely to be</i>
12	andere Ausdrücke	

Tabelle 3.3: Übersicht der strukturellen Klassifikation nach Biber et al. (1999, 996).

Biber et al. (1999), Tabelle 3.3⁹, arbeiten eine detaillierte Gliederung zu den unterschiedlichen Strukturen von Lexical Bundles heraus. Diese Klassifikation beschreibt die vorhandenen Daten bezüglich ihrer Linearisierung und der enthaltenden Phrasen und Phrasenfragmenten. In Tabelle 3.3 sind typische Strukturen für beide Register CONV und ACAD aufgelistet. Ein (+) oder (X+)/(+X) zeigt eine mögliche Erweiterung eines Lexical Bundle in die eine oder andere Richtung an. Diese können Bestandteile bereits bestehender Phrasen sein, so wie ein mögliches Hilfsverb in Nr. 2, oder über diese hinaus reichen, wie die Verben in Nr. 5 und Nr. 10, die in einem übergeordneten, syntaktischen Verhältnis zu den Nebensatzfragmenten stehen. Pronomen, Artikel, Adjektive, Verben, Nomen, Präpositionen und Konjunktionen sind mit ihren jeweiligen Phrasenstrukturen, wie aus Tabelle 3.3 hervorgeht, in den Bundles enthalten. Jedes Lexical Bundle wird einer dieser Kategorien in Tabelle 3.3 zugeordnet, wobei in einer Sequenz noch mehrere strukturelle Typen enthalten sein können. So bildet sich Nr. 1 *I don't know what* aus Personalpronomen (PPER)+lex.VP mit einem *wh*-Fragment (vgl. auch Biber et al. 1999, 1001). Ein Teil der Daten kann keiner dieser Strukturen zugeordnet werden und ist in einer Restgruppe - *andere Ausdrücke* - zusammengefasst, für die keine Beispiele angegeben werden. Hier wäre interessant, welche Strukturen in diese Restmenge aufgenommen worden sind.

Etwa 60 Prozent aller Lexical Bundles, die in ACAD gefunden worden sind, lassen sich den Typen Nr. 6 und 7 aus Tabelle 3.3 zuordnen, hingegen sind es nur rund sieben Prozent in CONV, die zu diesen Typen zählen (vgl. Biber et al. 1999, 996). Es zeichnet sich also ein klares Gefälle der Verteilung in den verschiedenen Registern ab. In CONV werden mehr unterschiedliche Typen zu den jeweiligen strukturellen Einteilungen gefunden - 424 verschiedene Typen. ACAD weist dazu nur 277 unterschiedliche Typen in den verschiedenen strukturellen Klassifikationen auf (vgl. Biber et al. 1999). Generell werden in den ACAD mehr NP/PP basierte Lexical Bundles genutzt als in CONV, in dem sich der Großteil aus verbalen Bestandteilen zusammensetzt (vgl. dazu auch Conrad und Biber 2004, 63-64). Damit kann gezeigt werden, dass mit Hilfe dieser Taxonomie Unterschiede im Gebrauch der jeweiligen strukturellen Typen zwischen den Registern festgemacht werden können. Die Klassifikationen variieren dennoch in den einzelnen Studien und legen jeweils andere Schwerpunkte für die Klassifikation fest. Conrad und Biber (2004) stellen nur einen Teil der möglichen Strukturen und ihre Verteilung zwischen den Registern ACAD und CONV vor.¹⁰

⁹Für alle Tabellen, die eine Übersicht über die strukturellen und funktionalen Klassifikationen vorgeben gilt: Für je eine Klassifikationsart wurde ein anschauliches Beispiel ausgewählt.

¹⁰Biber, Conrad und Cortes (2003) fokussieren ihre Diskussion auf den funktionalen Aspekt von Lexical Bundles nach der *Longman*-Studie und diskutieren keine neue Systematisierung der strukturellen Typen (ebd., 75).

Nr.	Struktur	Beispiel
1	PPER+lex.VP (+Komplement Satz)	<i>I don't know what</i>
2	(Aux+) mit Verb im Aktiv (+)	<i>have a look at</i>
3	<i>yes-no</i> and <i>wh</i> -Fragenfragment	<i>can I have a</i>
4	NP mit Postmodifiziererfragment	<i>the nature of the</i>
5	Präposition+NP-Fragment	<i>as a result of</i>

Tabelle 3.4: Übersicht der strukturellen Klassifikation nach Conrad und Biber (2004, 64).

Wie Tabelle 3.4 verdeutlicht, scheint es prototypische, strukturelle Konstruktionen in diesen Registern zu geben, die eine Raffung der Systematisierung von Biber et al. (1999) ermöglichen. Es zeigt sich dennoch, dass bei der strukturellen Klassifikation nicht ganz konsistente Zuordnungen beziehungsweise Änderungen der Zuordnung zu finden sind, deren Motivation ohne weitere Erklärungen nicht ersichtlich ist.

Nr.	Struktur	Beispiel
1	(Konnek.+) 1./2. Person PPER+ Satzfragment	<i>I want you to</i>
2	(Konnek.+) 3. Person PPER + Satzfragment	<i>it's going to be</i>
3	VP mit Verb im Aktiv	<i>take a look at</i>
4	VP mit Verb im Passiv	<i>is based on the</i>
5	<i>yes-no</i> -Fragenfragment	<i>do you want to</i>
6	<i>wh</i> -Fragenfragment	<i>what do you think</i>
7	<i>wh</i> -Satzfragment	<i>what I want to</i>
8	<i>if</i> -Satzfragment	<i>if you want to</i>
9	(Verb/Adjektiv+) <i>to</i> -Satzfragment	<i>to be able to</i>
10	<i>that</i> -Satzfragment	<i>that there is a</i>
11	(Konnek.+) NP mit <i>of</i> -Phrasenfragment	<i>one thing of the</i>
12	NP mit anderen Postmodifiziererfragmenten	<i>the way in which</i>
13	andere NP-Fragmente	<i>a little bit more</i>
14	PP-Ausdrücke	<i>at the end of</i>
15	komparativer Ausdruck	<i>as far as the</i>

Tabelle 3.5: Übersicht der strukturellen Klassifikation nach Biber (2004, 172).

Beispielsweise wird das Bundle *is based on the* in Biber et al. (1999), Nr. 9 in Tabelle 3.3, als *Verb im Passiv+PP-Fragment* klassifiziert. In der Klassifikation von Biber (2004), Tabelle 3.5, wird dieses als *VP mit Verb im Passiv* kategorisiert, womit das PP-Fragment oder mögliche weitere Ergänzungen nicht direkt aufgenommen werden. Auch die Unterscheidung zwischen Erster und Dritter Person fehlt bei Biber et al. (1999). Biber (2004) nimmt für die verschiedenen interrogativen Lexical Bundles je einzelne Kategorien an: Nr. 5 bis Nr. 7 Tabelle 3.5. Eine genauere Klassifikation der *Satzfragmente* - beispielsweise Nr. 8 *if*- oder Nr. 10 *that*-*Satzfragmente* Tabelle 3.5 - wird bei Biber (2004) vorgegeben, die in der Übersicht bei Biber et al. (1999) nicht berücksichtigt werden. In Tabelle 3.5 sind Nr. 12 bis 15 phrasal und werden damit detaillierter klassifiziert als bei Biber et al.

(1999), dessen Klassifikation für die phrasalen Typen nur zwei Kategorien zulässt; Nr. 6 *NP mit Postmodifiziererfragment* und Nr. 7 *Präposition+NP-Fragment*, Tabelle 3.3. Dieser kurze Vergleich zeigt, dass die Klassifizierung der Daten nicht unbedingt in allen Studien gleich erfolgt. Für solche Änderungen oder neue Schwerpunktsetzungen in den Taxonomien werden keine expliziten Motivationen oder Erklärungen gegeben. Wie variabel und Register spezifisch die Strukturen sein können, zeigt sich auch bei Cortes (2004), die auf Grundlage von zwei Haupttypen wiederum näher auf die unterschiedlichen Varianten der NP/PP- oder VP-basierenden Bundles eingeht, die oft in schriftlichen, akademischen Registern gefunden werden (vgl. Cortes 2004, 405, 407):

Nr.	Struktur	Beispiel
1	NP+ <i>of</i> -Phrasenfragment	<i>the nature of the</i>
2	NP+postnominales Satzfragment	<i>the fact that the</i>
3	PP+eingebettete <i>of</i> -Phrase	<i>as a result of</i>
4	andere präpositionale Fragmente	<i>at a time when</i>
5	Verb (<i>be</i>)+Komplement (NP)	<i>was one of the</i>
6	andere Ausdrücke	<i>as well as the</i>
7	andere NP	<i>an increase in the</i>
8	andere PP	<i>in addition to the</i>
9	<i>it</i> +V <i>be</i> +Adjektiv+(Satzfragment)	<i>it is possible to</i>
10	(modal) V <i>be</i> +(Komplement NP/AP)	<i>is a function of</i>
11	(NP/PPER)+V+(Komplement)	<i>not appear to be</i>

Tabelle 3.6: Übersicht der strukturellen Klassifikation nach Cortes (2004, 405, 407).

Aus Tabelle 3.6 wird deutlich, dass der dritte Haupttyp *Satzfragment* aus Biber, Conrad und Cortes (2004) in die beiden verbliebenen Gruppen integriert wird. Hier scheinen die Phrasentypen NP/PP/VP grundlegende Kriterien zur Klassifikation zu sein. Damit zeigt sich, dass die induktive Methode Raum für eine variable Betrachtung der Daten zulässt. Cortes (2004) untersucht zwei geschriebene, akademische Register, die sich generell durch die Vielzahl der NP/PP-basierten Bundles und untereinander abgrenzen. So folgt eine feinere Unterscheidung dieser Gruppen.

Tracy-Ventura, Cortes und Biber (2007) unterteilen strukturell in nur zwei Haupttypen - VP- und NP/PP-basiert. Diese werden nach ihren einzelnen Ergänzungen/Modifizierern weiter unterteilt (vgl. Tracy-Ventura, Cortes und Biber 2007, 221). Auf eine verkürzte strukturelle Klassifikation beziehen sich auch Biber, Conrad und Cortes (2004), in dem sie von drei Grundtypen ausgehen: Bundles mit VP-Fragmenten, Bundles mit NP-Fragmenten und Bundles mit Fragmenten abhängiger Sätze (ebd., 380). Diese gröbere Einteilung wird auch in Biber (2006) weiter angewandt. Biber und Barbieri (2007) hingegen geben keine detaillierte, strukturelle Klassifikation (vgl. Biber 2006, 136; Biber und Barbieri 2007, 260-271).

In der Summe konzentrieren sich demnach viele der strukturellen Klassifikationen auf die allgemeinen Phrasentypen NP, PP und VP. Nur wenige der Lexical Bundles, die beispielsweise in den Registern *Gesprächsaufzeichnungen* und *akademischer Prosa* aus dem

LSWE-Korpus identifiziert wurden, sind in sich strukturell geschlossen.¹¹ Dennoch zeigen diese Klassifizierungen, dass Lexical Bundles starke grammatische Korrelate besitzen, auf Grundlage dessen sie differenziert werden können (vgl. Biber 2004, 77). Weiterhin wird deutlich, dass die meisten Sequenzen zwei Phrasen oder Sätze überbrücken. Beispielsweise ist bei dem Bundle *the nature of the*, siehe Tabelle 3.6 Nr.1, die PP, die der NP *the nature* strukturell untergeordnet ist, nicht komplett und hat je nach Kontext eine andere lexikalische Vervollständigung. In gleicher Weise wird ein abhängiger Satz mit *to* in *I want you to*, siehe Tabelle 3.5 Nr.1, eingeleitet, bleibt aber strukturell unvollständig (vgl. Biber et al. 1999, 995, Biber 2004, 173-174).

Die Typologien von Tracy-Ventura, Cortes und Biber (2007), Cortes (2004), Biber (2004) und auch Biber et al. (1999) können potenziell auf die zwei bis drei strukturellen Haupttypen nach Biber (2006) und Biber, Conrad und Cortes (2004) zusammengefasst werden. Wie eingangs schon hervorgehoben, folgt die strukturelle Klassifikation nach der Datenerhebung und ist demnach reaktiv auf die spezifisch gewonnenen Daten aus den einzelnen Registern. Wenn Arten und Verteilungsmuster von Lexical Bundles Register typisch sind, dann kann in einer Studie die Notwendigkeit bestehen, einen Typen zu klassifizieren, der in einer anderen Studie mit anderer Textgrundlage nicht auftreten wird. Dieser Schritt oder auch abweichende Klassifikationen werden kaum weiter kommentiert oder motiviert. Eine genaue und konventionalisierte Klassifikation hinsichtlich der Struktur von Lexical Bundles scheint bislang noch nicht möglich.

Eine frequenzbasierte, strukturell blinde Analyse kann demnach linguistisch interpretierbare Daten, die auf grammatisch vollständige Einheiten zurückführbar sind, erzeugen. Die sich zum Teil widersprechenden Klassifizierungen und beispielsweise die strukturellen Fragmente aus Tracy-Ventura, Cortes und Biber (2007) sind nicht ausreichend diskutiert. Fraglich ist, in wie weit diese Studien nachvollziehbar und reproduzierbar sind. Gerade in Hinblick auf andere Sprachen oder stark differierende Register ist ein Transport der Taxonomien nicht ohne weiteres möglich.

3.3.2 Erkenntnisse aus der strukturellen Klassifikation

Bereits mit der strukturellen Klassifikation können einige Erkenntnisse hinsichtlich des linguistischen Status von Lexical Bundles gewonnen werden. Bei einem ersten Vergleich der Bundles zeigt sich, dass Lexical Bundles scheinbar aus linguistischen Rahmen bestehen, die *Lücken* für auswechselbare und/oder feste Elemente beinhalten (Biber, Conrad und Cortes 2003, 78):

- (1)
'I/you + don't/didn't + know/think/want +
(Biber, Conrad und Cortes 2003, 78)'

Der Rahmen *PPER + NEG + lex.V + (x)* bleibt bestehen, die einzelnen lexikalischen Realisierungen variieren. Das Verb wird je nach Aussagevorhaben und Ausdrucksart der

¹¹Beispielsweise sind in CONV nur 15 Prozent der Lexical Bundles strukturell komplett und weniger als fünf Prozent in ACAD bilden vollständige Phrasen (Biber et al 1999, 995).

Äußerung gewechselt, das Subjektpronomen wird dem Agens angepasst und *NEG* erhält zum *lexV* eine zeitliche Markierung. Alle diese Varianten wurden im Korpus gefunden (Biber, Conrad und Cortes 2003, 78). Davon grenzen sich weniger variantenreiche Bundles, indem sie einen sehr festen Rahmen bieten, ab:

- (2)
'I want to + do/get/see/go/know
Do you want + to/a/me/some/any/it/the
(Biber, Conrad und Cortes 2003, 78)'

Diese Sequenzen werden zu einem Großteil genauso gebraucht: Das Subjektpronomen und das Verb variieren im Gebrauch nicht, lediglich das Brückenelement, das die ersten lexikalischen Informationen des subordinierten Satzes liefert, ist variabel. Die Lexical Bundles besitzen demnach *strukturelle Rahmen*, die bestimmte *Lücken* beinhalten, die dann wiederum mit lexikalischen Elementen gefüllt werden können. Dieser Rahmen funktioniert nach Biber (2006):

- '[...] as a kind of discourse anchor for the 'new' information in the slot, telling the listener/reader how to interpret that information with respect to stance, discourse organization, or referential status.' (Biber 2006, 172).'

Diese Diskursfunktionen werden in Kapitel 3.4 näher betrachtet.

3.3.3 Zusammenhang zwischen Wortarten und struktureller Klassifikation

Ein wichtiger Aspekt im Zusammenhang mit den strukturellen Systematisierungen ist die Untersuchung der Verteilung der Wortarten in den verschiedenen Registern. Für die Register CONV und ACAD aus dem LSWE-Korpus und den Registern CLASS und TEXTB aus dem T2K-SWAL-Korpus erstellen Biber, Conrad und Cortes (2004) eine Übersicht über die Verteilung der Wortarten *Verb*, *Personalpronomen* und *Nomen* (vgl. ebd., 378). Hier zeigt sich, dass die *Gesprächsaufzeichnungen* etwa gleiche Anteile bei allen Wortarten aufweisen und die *Unterrichtsgespräche* eine ähnlich gleichmäßige Verteilung zeigen, wobei Nomen etwas mehr als Verben und Pronomen vorkommen. Bei den schriftlichen Registern hingegen wird deutlich, dass es überproportional mehr Nomen als Verben oder Pronomen gibt (vgl. ebd., 378). Auch Conrad und Biber (2004) resümieren dieses Verhältnis so: *Gesprächsaufzeichnungen* bestehen mehr aus Verben, Personalpronomen und Fragen, die *akademische Prosa* beinhaltet hingegen mehr NPn und PPn (vgl. Conrad und Biber 2004, 64).

Biber (2004) erstellt eine Übersicht über die Wortartenverteilung von Nomen, Verben, Adverbien und Adjektiven für die akademischen Register des T2K-SWAL-Korpus (ebd., 169). Hieraus ergibt sich ebenfalls, dass in den schriftlichen Registern überproportional mehr Nomen als Verben, Adverbien und Adjektive gebraucht werden. Die mündlichen Register beinhalten mehr Verben, wobei dieses Verhältnis nicht so klar wie bei den schriftlichen Registern ist. Die Übersicht zeigt auch einen Unterschied zwischen konventionellen

Registern wie *Dienstleistung* und informationellen Registern wie *Unterrichtsgespräch* an. Erstere ähneln in dieser Hinsicht dem Register CONV aus der *Longman*-Studie und letztere eher ACAD (vgl. Biber 2004, 169-170). Gerade *Unterrichtsgespräch* fällt aus der sonst klaren Unterscheidung zwischen mündlichen und schriftlichen Registern heraus, da aus beiden Bereichen bestimmte Charakteristika wiederzufinden sind.

Die Verteilung der strukturellen Typen verhält sich gerade in Hinblick auf die schriftlichen Register tendenziell wie die der Wortarten. Im Register *Gesprächsaufzeichnung* enthalten über 90 Prozent der Wortketten Teile einer VP, in der *akademischen Prosa* besteht der größte Anteil, um 60 Prozent, aus vollständigen oder fragmentalen NPn oder PPn (vgl. Conrad und Biber 2004, 63-64). Auch die Daten aus dem Spanischen zeigen ein gleiches Verteilungsmuster. So wurden nur sehr wenige VP-basierte Bundles im Vergleich mit den NP-basierten Bundles in der akademischen Prosa gefunden, etwas mehr als die Hälfte der Bundles in *Soziolinguistischen Interviews* sind hingegen VP-basiert (vgl. Tracy-Ventura, Cortes und Biber 2007, 222). Damit zeigt sich ein Zusammenhang zwischen Verteilung der Wortarten und Verteilung der verschiedenen strukturellen Bundles.

3.4 Funktionale Eigenschaften von Lexical Bundles

3.4.1 Taxonomien im Vergleich

Die funktionale Klassifikation erfolgt ebenfalls nach der Datenerhebung. Generell werden alle Bundles mit ähnlichen Funktionen basierend auf den typischen Bedeutungen und deren Gebrauch zusammengefasst. Hier werden nicht alle Ergebnisse und Diskussionen aller Studien ausführlich betrachtet, vielmehr soll ein Überblick über die verschiedenen Taxonomien und ein knappes Resümee der wichtigsten Ergebnisse gegeben werden. Eine ausgearbeitete Systematisierung wie von Biber, Conrad und Cortes (2003) Tabelle 3.7 wird bei Biber et al. (1999) nicht gegeben, jedoch beispielorientierte, funktionale Beschreibungen der Lexical Bundles (vgl. ebd., 1003-1005). Wie aus Tabelle 3.7 hervorgeht, unterteilen Biber, Conrad und Cortes (2003) die Lexical Bundles in vier funktionale Haupttypen, die unterschiedliche Subklassifikationen erlauben.¹² Für jede dieser Kategorien geben Biber, Conrad und Cortes (2003) eine kurze Definition:

'Referential Bundles make direct reference to elements in the physical world or the textual context. Text organizers reflect relationships between prior and coming discourse. Stance bundles express attitudes or assessments of certainty towards the following proposition. Finally, interactional bundles are usually situational formulas associated with a specific situation, or conversational expressions used as strategies for conversational interactions (Biber, Conrad und Cortes 2003, 79).'

Auf Grundlage dieser Einordnung wird jedem Bundle eine bestimmte Funktion zugeordnet, vielen der Bundles können dennoch mehrere Funktionen übernehmen: Das Bundle *at*

¹²Auf eine Übersetzung der Terminologie wird hier verzichtet, da noch keine deutschen Entsprechungen herausgearbeitet worden sind. Begriffe wie *Stance Bundles* oder *Referential Bundles* werden übernommen.

Kategorie	Subkategorie	Beispiel
<i>Referential Bundles</i>	time markers	<i>at the same time</i>
	time/place/text deixis	<i>at the end of</i>
<i>Text Organizers</i>	contrast/comparison	<i>on the other hand</i>
	inferential	<i>on the basis of</i>
	framing	<i>in the case of</i>
<i>Stance Bundles</i>	epistemic-personal/certain	<i>you know what I</i>
	epistemic-impersonal/certain	<i>the fact that the</i>
	epistemic-personal/uncertain	<i>I don't know if</i>
	epistemic-personal/probable	<i>I think it was</i>
	epistemic-impersonal/probable	<i>it is possible to</i>
	desire	<i>do you want a</i>
	obligation	<i>it is necessary to</i>
<i>Interactional Bundles</i>	intention	<i>I was going to</i>
	true inquiry	<i>what do you think</i>
	reporting	<i>I said to him</i>
	imprecision tags	<i>or something like that</i>
	politeness markers	<i>thank you very much</i>

Tabelle 3.7: Übersicht der funktionalen Klassifikation nach Biber, Conrad und Cortes (2003, 80).

the end of the kann entweder eine zeitliche, lokale oder textuelle Referenz ausdrücken (Biber, Conrad und Cortes 2003, 79). Mögliche lexikalische Füllungen für die verschiedenen Referenzen könnten sein:

- (3) *at the end of the week* - zeitlich
- (4) *at the end of the corridor* - lokal
- (5) *at the end of the section* - textuell

Diese Multifunktionalität wird nicht an allen Beispielen diskutiert. So wird nicht klar, warum das Bundle *on the basis of* als Subkategorie *inferential* der *Text Organizers* geführt wird, siehe Tabelle 3.7, aber in den darauf gegebenen Beispielen als *framing bundle* diskutiert wird (vgl. Biber, Conrad und Cortes 2003, 80-81). Biber, Conrad und Cortes (2003) schließen ihre Ausführungen damit, dass die erarbeitete Taxonomie auf andere Register anwendbar ist, aber kein Anspruch auf Vollständigkeit besitzt (ebd., 82-83). Mit Blick auf die nachfolgenden Studien ist dies nicht ohne Einschränkungen der Fall. Cortes (2004) setzt für die *Text Organizers* noch eine weitere Subkategorie *Focus* an (ebd., 408-409). Alle der hier betrachteten Analysen¹³ diskutieren die *Text Organizers* größtenteils wie *Discourse Organizers* und umgekehrt. Weiterhin wird mit *Topik* und

¹³Diskutiert werden hier ausführlicher Conrad und Biber (2004), Biber, Conrad und Cortes (2004), Biber (2006) sowie Biber und Barbieri (2007).

Fokus operiert, um diese Gruppe weiter zu klassifizieren (vgl. u.a. Conrad und Biber 2004, 66; Biber und Barbieri 2007, 271).

Die Subkategorien für die *Discourse Organizers* werden bei Conrad und Biber (2004) wie folgt definiert:

'They serve two major functions: topic introduction/focus and topic elaboration/clarification. Topic introduction/focus bundles provide overt signals that a new topic (or subtopic) is being introduced or is becoming the focus of attention (e.g. *do you know what, I tell you what*). Topic elaboration/clarification bundles serve to add more information to a topic (e.g. *nothing to do with*) (Conrad und Biber 2004, 76).'

Unter *Elaboration/Clarification Bundles* fallen auch jene, die eine Erklärung/Klärung der vorangegangenen Aussage beinhalten/erfragen oder die Beziehungen zwischen Diskurseinheiten herstellen (Conrad und Biber 2004, 76). Eine ähnliche Taxonomie liegt auch bei Biber, Conrad und Cortes (2004) vor (ebd., 392-393), vgl. Tabelle A.1. Biber (2006) erweitert diesen funktionalen Typus um *Identification/Focus Bundles*, denen er die zuvor als *Referential Bundles* eingeordneten Bundles wie *those of you who* zuordnet, und um die Subkategorie *Conditions*, in der Lexical Bundles eingeordnet werden, die Bedingungen ausdrücken (Biber 2006, 144, 159). Die Phrase, die nach diesen *Identification/Focus Bundles* benutzt wird, wird als besonders wichtig hervorgehoben (Biber 2006, 144). Biber und Barbieri (2007) fügen hinzu, dass diese *Identification/Focus Bundles* oft nach längeren Erklärungen folgen, um deren Inhalt zusammenzufassen oder den Schwerpunkt herauszustellen (Biber und Barbieri 2007, 271).

Der Fall der *Discourse Organizers* zeigt, dass eine funktionale Taxonomie noch nicht ausreichend entwickelt worden ist und entweder je nach Textgrundlage oder nach einer erneuten Prüfung der Daten von anderen abweicht. Einen Grund dafür stellt die Methode der Studie dar. Die Zuordnung der *Discourse Organizers* ist in allen Fällen definiert, dennoch wird weder Fokus noch Topik hinreichend in diese Untersuchung eingeführt. Beide Begriffe sind in der Forschung zu Informationsstruktur viel diskutiert und besitzen mehrere Entsprechungen und definitorische Grundlagen (vgl. Musan 2010, 59-61).¹⁴

Conrad und Biber (2004) stellen darüber hinaus bei ihrer Analyse fest, dass

'[...] the common 4-word-bundles in conversation and academic prose have almost entirely non-overlapping distributions with respect to the functional categories (Conrad und Biber 2004, 67).'

Lexical Bundles können so die funktionalen Eigenschaften, die kommunikativen Ziele und Kontexte jedes Registers widerspiegeln. Die *Gesprächsaufzeichnungen* sind dialogisch, interaktiv und die Sprecherhaltung sowie die Interpretation der Propositionen werden in diesen Registern hauptsächlich mittels der *Stance Expressions* und bestimmter *Discourse Organizers* ausgedrückt (vgl. Conrad und Biber 2004, 66-67). Im Gegenteil dazu werden in der *akademischen Prosa* fast ausschließlich *Referential Bundles* identifiziert, mit deren

¹⁴Für eine genauere Diskussion siehe Kapitel 6.3.2.

Hilfe dem Rezipient präzise Informationen ohne persönliche Einbindung vermittelt werden. Die wenigen enthaltenden *Stance Bundles* sind im Gegensatz zu den *Gesprächsaufzeichnungen* immer unpersönlich (vgl. Conrad und Biber 2004, 67-68).

Die funktionalen Verteilungsmuster in den verschiedenen Texten lassen nach den hier vorgestellten Studien Rückschlüsse auf bestimmte Kernfunktionen eines einzelnen Registers beziehungsweise über Aufgaben des Sprechers in dem jeweiligen Register zu. Die schriftlichen Register weisen generell weniger Bundles auf als die mündlichen, letztere beinhalten zu einem großen Teil *Stance Bundles*, die wiederum in den *Textbüchern* und der *akademischen Prosa* seltener Verwendung finden (vgl. Biber, Conrad und Cortes (2004, 396)). *Stance Bundles* sind zu einem großen Teil persönlich, sprecher- und hörerbezogen, was unter anderem durch Personalpronomen realisiert wird (vgl. Biber, Conrad und Cortes 2004, 389). Die wenigen unpersönlichen *Stance Bundles* finden Anwendung in den schriftlichen Registern, wo der persönliche Bezug in den Hintergrund tritt. Auch die *Discourse Organizers* sind in den mündlichen Registern gebräuchlicher als in den schriftlichen (vgl. Biber, Conrad und Cortes 2004, 396). Diese lassen teilweise ebenfalls eine Referenz auf den Sprecher/Hörer zu und haben interaktiv-kommunikative Zwecke wie *I want to talk about* oder *I want to do*. Die *Referential Bundles* sind dagegen eher phrasal, was das vorhergehende Kapitel bereits aufgezeigt hat, und ohne Sprecher- oder Hörerbezug. Sie sind ein Mittel, um präzise Informationen und Relationen darzustellen. Das untersuchte Register *Unterrichtsgespräch* hingegen nutzt alle funktionalen Typen gleichermaßen (vgl. Biber, Conrad und Cortes 2004, 396), wodurch eine klare Zuordnung, welche Funktionen eher im mündlichen und welche im schriftlichen Bereich benötigt werden, nicht klar abzugrenzen ist. Ein etwas gegensätzliches Ergebnis weisen Tracy-Ventura, Cortes und Biber (2007) auf. Ganz allgemein werden mehr Lexical Bundles in der *akademischen Prosa* genutzt als in den *soziolinguistischen Interviews*, was im direkten Widerspruch zu den Ergebnissen im Englischen steht, wo die akademische Prosa generell sehr wenig Gebrauch von Lexical Bundles im Vergleich zu allen anderen Register macht (ebd., 230).

Grundsätzlich können auch Tendenzen in Bezug auf die Beziehung zwischen Form und Funktion aus den Daten abgeleitet werden. Aus der Tabelle A.1 und Biber, Conrad und Cortes (2004, 384-388) sowie aus der funktionalen Taxonomie von Conrad und Biber (2004, 65-66) geht hervor, dass die *Referential Bundles* zum Großteil aus phrasalen Bundles bestehen, während die *Stance Bundles* oft VP-basiert sind und strukturelle Anchlüsse für Subordinierer enthalten. Lediglich die *Discourse Organizers* weisen kein so klares Verhältnis auf (vgl. Biber, Conrad und Cortes 2004, 397-398). So scheinen Form und Bedeutung miteinander in Beziehung zu stehen: Eine komplexe Interaktion zwischen Form, Diskursfunktion, typischen Zielen/Zwecken eines Registers und der situativen Eigenschaften von Lexical Bundles scheint dadurch plausibel (vgl. Biber, Conrad und Cortes 2004, 398; Biber 2006, 162-172; Biber und Barbieri 2007, 273, 282).

3.4.2 Strittige Klassifikationen

Innerhalb der Taxonomien selbst sowie zwischen den einzelnen Studien zeigen sich inkonsistente oder nicht schlüssige Zuordnungen. Wie bereits erwähnt, ist eine komplette

Übernahme der Systematisierungen nicht unbedingt von Vorteil. Beispielsweise ergeben sich aus den Daten nicht immer Beispiele für jeden Typus: Cortes (2004) ordnet in dem Register *Geschichte* den *Referential Bundle time-markers* keine Beispiele zu (vgl. Cortes 2004, 406). Hier zeigt sich, dass sich die funktionalen Zuordnungen nach den Daten richten und in der Auswertung darauf reagiert werden muss. Aus dem Vergleich mit den Klassifikationen von Conrad und Biber (2004) und Biber, Conrad und Cortes (2003, 80), siehe Tabelle 3.7 wird deutlich, dass die Taxonomie von Conrad und Biber (2004, 65-66) eine feinere Unterscheidung bei den referentiellen Wortketten vorsieht, die Wortketten der *Discourse Organizers* nach informationsstrukturellen Kriterien neu ordnet und Unterscheidungen zwischen Sicherheit/ Unsicherheit bei den *Stance Bundles* hingegen nicht wieder aufgreift. Beide Studien gehen von den aus dem LSWE-Korpus erhobenen Daten aus und kommen dennoch zu unterschiedlichen Interpretationen.

Dies zeigt sich ebenfalls bei Biber (2006). Diese Studie umfasst viele verschiedene Register, deren Bundles eine spezifische Erweiterung der Systematisierung bedürfen. Beispielsweise wird die Systematisierung um eine weitere Subkategorie für die *Discourse Organizers*, die *Imperatives* erweitert. Längere Bundles mit bestimmten Funktionen für ein Register erhalten eine neue Kategorie, die *longer expressions in service encounters* (vgl. Biber 2006, 151-153). Diese Gruppierung umfasst Mehr-Wortketten, die über vier Token hinausgehen (Biber 2006, 156; Biber und Barbieri 2007, 277-278). Hier wird ein Register dadurch hervorgehoben, dass Sprecher vorwiegend immer wieder dieselben Sätze für die jeden Tag gleichen Anfragen benutzen. Diese Wiederholungen sind nach Biber (2006) keine typischen Lexical Bundles, da sie meist strukturell komplett und mit ihrem Satzstatus keine eigentlichen *building blocks of discourse* darstellen (Biber 2006, 156). Wenn die *longer expressions in service encounters*-Bundles keine eigentlichen Bundles darstellen, ist die Motivation, eine solche Kategorie dennoch zu diskutieren, nicht klar. Der theoretische Status der Bundles und damit auch die Übertragbarkeit des Konzeptes auf andere Register und Sprachen ist vor dem Hintergrund solcher Klassifikationen schwer greifbar.

Der ebenfalls von Biber et. al (1999) eingeführte Begriff *Lokale Wiederholung* sowie dessen Anwendung ist nicht in allen Studien in gleicher Weise übernommen worden. Cortes (2004) erstellt für die *Referential Bundles* beider Korpora eine neue Subkategorie, die sie *subject-bound-bundles* nennt (vgl. ebd., 406-405, 409):

'It was previously mentioned that the corpus of published academic writing in history contained several bundles that had not been identified as frequent word combinations in the related literature (Biber et al., 1999). Some of these bundles reflect an inherent connection to the social disciplines, especially those related to social events or issues (*the power of the, the creation of the, in the context of, the image of the*) (Cortes 2004, 406).'

Cortes (2004) bezieht sich dabei nur auf 'Some of these', so wird nicht deutlich, wie genau diese Bundles identifiziert werden. Mit dieser Eingrenzung sind nicht alle möglichen lokalen Wiederholungen gemeint, sodass die Ergebnisse nicht auf ein Nichtbeachten dieser Richtlinie zurückzuführen ist.

Aus der Untersuchung spanischer Korpora gehen Bundles hervor, die nur Funktionswör-

ter enthalten, unter der Kategorie *Structural Only* zusammengefasst und als strukturelle Artefakte nicht weiter diskutiert werden (vgl. Tracy-Ventura, Cortes und Biber 2007, 225, 228). Diese strukturellen Artefakte kommen hingegen in beiden Teilkorpora mit 20 bis 39-mal pro eine Million Wörter und mit 40 bis 99-mal pro eine Million Wörter vor (vgl. Tracy-Ventura, Cortes und Biber 2007, 225). Damit sind sie kein unwesentliches Nebenprodukt der Datenerhebung und auch nicht per Definition ausgeschlossen, weder strukturell noch funktional. Mit dem konkreten Begriff *Lexical Bundle* wird vorweggenommen, dass lexikalische Einheiten im Fokus der Untersuchung stehen. Dennoch umfasst die Definition dieses Begriffes diesen Aspekt nicht direkt. In diesen Bundles sind auch pronominale Konstruktionen wie *mi/me* enthalten, deren lexikalischer Bezug über die Koreferenz mit einem Diskursreferenten besteht. Eine klare Einordnung und Diskussion auch hinsichtlich der Definition von Lexical Bundles steht hier aus.

Ebenfalls nicht per Definition ausgeschlossen sind Bundles wie *in the United States*. Diese werden bei Biber, Conrad und Cortes (2004) als lokale *Referential Bundles* eingeordnet (ebd., 288, siehe auch Tabelle A.1). Eine strenge Auslegung der Restriktion über die lokalen Wiederholungen könnte diese Bundles als themenspezifisch klassifizieren und so ausschließen. Das Bundle stellt eine sehr spezifische Ortsangabe beziehungsweise einen Eigennamen dar. Eigennamen, die mehrere Token umfassen, werden so als Lexical Bundle bei entsprechender Frequenz analysiert. Weiterhin fällt hier auf, dass Eigennamen nicht in den Studien diskutiert sind und deren Handhabung somit nicht eindeutig festgelegt ist. Eine Zusammenfassung der Token bei Eigennamen mit mehreren Wörtern erscheint hier in jedem Fall sinnvoll.

Aus diesen Studien wird deutlich, dass auch die funktionalen Klassifikationen in Anbetracht der Daten variieren. Dennoch wird wenig darauf hingewiesen und die Änderungen werden kaum diskutiert. Somit erscheinen die Richtlinien wenig transparent und eine Übernahme dieser für weitere Studien erscheint problematisch. Trotz allem können aus diesen Ergebnissen Tendenzen und bestimmte Präferenzen der einzelnen Register hinsichtlich Modus und kommunikativem Zweck aufgezeigt werden.

4 Verwandte Phänomene

Neben den Studien zu Lexical Bundles gibt es weitere, die einen ähnlichen, frequenzialen Ansatz zur Untersuchung sprachlicher Phänomene verfolgen. Diese sind zumeist theoretisch in die Phraseologieforschung eingebettet. Altenberg (1998) untersucht Phraseologien gesprochener Sprache mit Hilfe des *London-Lund*-Korpus. Die linguistischen Phänomene, die er dazu untersucht, werden wie folgt definiert:

‘By ‘recurrent word-combinations‘ I here mean any continuous string of words occurring more than once in identical form (Altenberg 1998, 101).’

Wie bei der Definition von Biber et al. (1999) liegt hier der Fokus ebenfalls auf sich wiederholenden Mehr-Wortketten, deren Identifikation frequenziell erfolgt. So werden Frequenzlisten errechnet und ausgewertet. Der empirisch-frequenzielle Ansatz ist bei Altenberg (1998) und den Studien von Biber und Kollegen der gleiche. Im Gegensatz zu den Lexical Bundles werden hier keine weiteren strukturellen Beschränkungen bei der Identifikation auferlegt und keine Vorkommengrenze für die Relevanz einer Wortkette bestimmt. Dennoch sind keine satzübergreifenden Phrasen in seinen Beispielen verzeichnet. Die Datenerhebung erfolgt ebenfalls an nicht interpretierten Primärdaten (Altenberg 1998, 104). Danach folgt eine strukturelle und funktionale Charakterisierung der Wortketten. Ersteres wird allgemein zwischen abhängigen und unabhängigen Sätzen und Satzgliedern unterschieden. Letzteres in allgemeine Funktionen wie Antworten, Metafragen, indirekte Bedingungen oder Interrogativa eingeteilt (Altenberg 1998, 104, 109, 114, 115-117). Darüber hinaus betrachtet Altenberg (1998) hochfrequente Wortkombinationen und Kollokationen, die aus den Sequenzen ablesbar sind (ebd., 108, 120). Altenberg (1998) unterstreicht, dass alle dieser Sequenzen fest - ‘frozen’ - aber dennoch semantisch transparent sind. Er stellt ebenfalls die Rahmenstruktur, die einen Kern mit bestimmten Erweiterungen beinhaltet, in den Vordergrund der Diskussion (Altenberg 1998, 104-105). Somit zeigen sich Parallelen mit den Studien von Biber und Kollegen zu den Annahmen und Schlussfolgerungen in Bezug auf die Lexical Bundles. Altenbergs Methode wird auch unter anderem in Biber, Conrad und Cortes (2004) als Vorgängerstudie zu den Lexical Bundles erwähnt (Biber, Conrad und Cortes 2004, 372). Der so verfolgte Ansatz von Biber und Kollegen wird bei der Diskussion der unvollständigen Phrasen deutlich (Altenberg 1998, 119). Die Annahme, dass Sprecher leicht zugängliche, feste Phrasen innerhalb der Sprachproduktion nutzen, ist in beiden Studien zentral, auch wenn sie in unterschiedlich tiefer struktureller und funktionaler Weise untersucht wird (vgl. Altenberg 1998, 121). Diese Schlussfolgerung ziehen auch Teliya et al. (1998, 62). In ihrer Untersuchung ist der kulturelle Einfluss auf Sprache selbst im Vordergrund, daraus ableitbar wären feste Phrasen und Idiome, die in unterschiedlichen Diskursen auftreten. Den Ansatz, mit Hilfe der Sequenzen Register zu untersuchen, verfolgt Altenberg (1998) nicht.

Untersuchungen zu rekurrenten Wortketten und deren Register unterscheidenden Funktionen führen unter anderem Partington und Morley (2004) durch. Sie erstellen ebenfalls Frequenzlisten rekurrenter Vier-Wortketten - hier *Cluster* genannt - von verschiedenen Registern (Partington und Morley 2004, 179-181). Wie Altenberg (1998) diskutieren Partington und Morley (2004) keine vollständigen Datensätze und analysieren deren Verteilung, sondern beziehen sich anscheinend auf eine Auswahl der häufigsten Sequenzen. Bei Partington und Morley (2004) liegt der Schwerpunkt auf dem Kontext der Untersuchung und den rhetorischen Strategien, die daraus ablesbar sind (ebd., 183-189). Nach Partington und Morley (2004) sind Cluster:

'[...] at the first level of abstraction from this morass - just like collocations with which, in fact they have a great deal in common (Partington und Morley 2004, 190).'

Lexical Bundles liegen demnach auf einem höheren Level der Abstraktion. In jedem Fall gehen Partington und Morley (2004) auch davon aus, dass diese rekurrenten Wortketten kein Zufallsprodukt sind und viel über den Gebrauch von Sprache zeigen können (ebd., 190).

All diesen Ansätzen liegt die Annahme zugrunde, dass Sprache zu einem großen Teil aus festen, hochfrequenten Phrasen/Clustern/Bundles besteht und diese auch Funktionen im Diskurs erfüllen. Hinsichtlich der Anzahl- und Frequenzbeschränkungen beziehungsweise der jeweiligen Auswertung der Daten differieren die einzelnen Studien. Die Herauslösung dieses Ansatzes aus der Phraseologie in die Registerforschung ist bezeichnend.

Wie innerhalb der Phraseologieforschung eine frequenzbasierte Untersuchung auf Web-Korpora angewendet werden kann, zeigen Quasthoff, Schmidt und Hallsteinsdóttir (2010). Sie diskutieren die Häufigkeit, den Rang/die Wichtigkeit und die Struktur von Phraseologismen aus einer vorgefertigten Listen in unter anderem Zeitungstexten und Wikipedia-Texten aus dem Internet (ebd., 38).

Wie nahe sich die Konzepte von frequenzbasierten Phänomenen sind, wird auch bei der Definition der Lexical Bundles deutlich:

'Lexical Bundles are extended collocations, sequences of three and more words that statistically co-occur in a register (Cortes 2004, 400).'

Firth (1957) stellt folgende Funktionen hinsichtlich der Wortbedeutung von Kollokationen heraus:

'Meaning by collocations is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night* (Firth 1957, 196).'

Diese Definition von Kollokationen wird bei Partington (1998) als textuell im Vergleich zu psycholinguistischen Konzepten, die von Assoziationsverbindungen eines Wortes zum anderen bezeichnet. Ein dritter Schwerpunkt wird mit dem statistischen Konzept von

Kollokationen ausgewiesen (Partington 1998, 15-16). Nach Evert (2009) muss zwischen dem empirischen und dem theoretischen Konzept von Kollokationen unterschieden werden. Unter dem empirischen Konzept versteht Evert (2009) rekurrente und vorhersagbare Wortkombinationen, die in natürlicher Sprache gefunden werden können. Das theoretische Konzept von Kollokationen wird durch linguistische Tests und Sprecherurteile gewonnen (ebd., 1213-1214). Und weiter:

'First and foremost, collocations are observable facts about language, i.e. primary data. From a strictly data-driven perspective, they can be interpreted as empirical predictions about the neighborhood of a word (Evert 2009, 1218).'

Kollokationen sind typischerweise Wortpaare, die zusammen in einem Kontext auftreten. Partington (1998) schlägt einen weiteren Anwendungsrahmen für Kollokationen - Wörter, Phrasen und Sätze - vor (ebd., 16-17). In jedem Fall unterscheiden sich Kollokationen von den rekurrenten Wortketten, die eine Serialisierung von mehreren Wörtern darstellen, in dem sie diese so nicht immer aufweisen müssen. Von einem theoretischen Standpunkt aus werden Kollokationen wie Idiome oder kulturelle Stereotypen als ein Epiphänomen innerhalb der Sprache betrachtet (Evert 2009, 1218). Damit besteht eine konzeptionelle Verbindung zwischen der Kollokations- und der Idiomforschung. Gerade Kollokationen wie *kick* und *bucket* werden mit Idiomem gleichgesetzt. Wulff (2008) weist hingegen darauf hin, dass das Konzept *Idiom* kein binäres Phänomen - Idiom vs. Nicht-Idiom - ist, sondern durchaus als ein Kontinuum hinsichtlich Flexibilität, Funktionalität und Frequenz gedacht werden kann (Wulff 2008, 8).

Allen Forschungsrichtungen ist die auf Frequenz basierte Datenerhebung gemein. Der jeweils konkret untersuchte, linguistische Gegenstand variiert, es lassen sich dennoch Beziehungen zwischen ihnen erkennen. Grundsätzlich können die hier kurz dargestellten linguistischen Phänomene und die Lexical Bundles in die Forschung zur Kombinatorik von Sprache eingebettet werden.

5 Korpusstudie

5.1 Korpus

Für diese Studie wurden zwei Korpora aus einer Auswahl an Texten aus dem Forum *abgeordnetenwatch.de* sowie aus den *Stenografischen Berichten des Deutschen Bundestages* ausgewählt.¹⁵ Die *Plenarprotokolle* wurden in einem Zeitraum vom 18.10.2005 bis 08.07.2011 und damit aus der abgeschlossenen 16. Legislaturperiode und der derzeitigen 17. Legislaturperiode zufällig ausgewählt.¹⁶ Weiterhin wurden nur die *Redetexte* jedes Protokolls in die Untersuchung eingebunden.¹⁷

Das Korpus *Politikforum* besteht aus insgesamt 1429 Einträgen aller 110 Abgeordneten der SPD, die in der 16. und 17. Legislaturperiode ein Mandat besaßen beziehungsweise besitzen. Die Anzahl der Beiträge der einzelnen Abgeordneten differenziert stark, sodass eine Obergrenze von 15 Beiträgen je Abgeordneter gesetzt wurde. Der jeweilige Mittelwert der Anzahl der Beiträge aus einer Legislaturperiode wurde als Faktor für die Auswahl jeder n-ten Antwort benutzt, um so die einzelnen Beiträge der Abgeordneten auszuwählen.¹⁸

¹⁵Es kann online auf die Originaldaten des jeweiligen Korpus unter <http://www.abgeordnetenwatch.de/> und <http://www.bundestag.de/dokumente/protokolle/plenarprotokolle/index.html> zugegriffen werden. Die Adressierungen der Websites sind grafisch hervorgehoben, um sie von Eigennamen wie *abgeordnetenwatch.de* zu unterscheiden.

Das Korpus der Beiträge auf *abgeordnetenwatch.de* ist bereits über das Such- und Visualisierungssystem ANNIS2 unter <http://www.sfb632.uni-potsdam.de/~d1/annis/> über einen beantragbaren Login verfügbar. Das Korpus der Plenarprotokolle wird auf ANNIS2 mit freiem Zugang zur Verfügung gestellt werden.

¹⁶Die Begriffe *Stenografischer Bericht des Deutschen Bundestages* und *Plenarprotokolle* werden hier synonym gebraucht.

¹⁷*Plenarprotokolle* beinhalten ein Inhaltsverzeichnis, den Redetext selbst und verschiedene für die jeweilige Sitzung relevante Anlagen wie zum Beispiel eine Liste der entschuldigten Abgeordneten, Amtliche Mitteilungen oder zu Protokoll gegebene Reden. Untersucht werden sollen hier die eigentlichen Reden der Abgeordneten, so dass alle weiteren Bestandteile der Protokolle nicht mit in die Datenerfassung integriert worden sind. Auch eventuelle Nachträge zu den Redetexten wie Berichtigungen werden nicht weiter berücksichtigt.

¹⁸Dieses Korpus ist auf Grundlage einer anderen wissenschaftlichen Fragestellung konzipiert und erstellt worden. Für die hier durchgeführte Registerstudie ist die spezielle Erstellung des Korpus grundsätzlich nicht vor- oder nachteilig. Die so gewonnenen Daten und Erkenntnisse können nur nicht umfassend für jede Fraktion aber dennoch für das Register *Politikforum* ausgewertet und diskutiert werden. Dies gelingt unter der Annahme, dass Foreneinträge als ein Register bestimmte Eigenschaften besitzen, und unter der Voraussetzung, dass auch die Fragen der Bürger an die Abgeordneten in dieser Studie betrachtet werden.

Daraus resultiert folgende Datengrundlage für die hier untersuchten Register *Plenarprotokoll* und *Politikforum*:

Register	Anzahl Texte	Anzahl Wörter
<i>Politikforum</i>	1.429	513.759
<i>Plenarprotokoll</i>	35	2.644.527
Gesamt	1.464	3.158.286

Tabelle 5.1: Übersicht über die Anzahl der Wörter und Texte der Korpora *Politikforum* und *Plenarprotokoll*.

Wie aus der Tabelle 5.1 ersichtlich, ist das Korpus der Plenarprotokolle wesentlich größer als das der Foreneinträge. Das Verhältnis von Anzahl der Texte zu Anzahl der Wörter verhält sich in den beiden Korpora genau gegenteilig. Im Hinblick auf die Vielfältigkeit und die variierende Größe der Korpora in den Vorgängerstudien zu Lexical Bundles (siehe Kapitel 3.1 und 3.2.2) erscheinen die Größe und auch das Text-Wörter-Verhältnis zumindest ausreichend.

5.2 Eigenschaften der Register *Politikforum* und *Plenarprotokoll*

Die Foreneinträge aus *abgeordnetenwatch.de* und die Redetexte der *Plenarprotokolle des Deutschen Bundestages* lassen sich thematisch in einen politischen Rahmen fassen. Teile der Sprecher - die Abgeordneten - und der Zeitraum der Datenerhebung - jeweils die 16. und 17. Legislaturperiode - sind identisch und in beiden Fällen umfassen die Korpora eine große Gruppe von Sprechern. Dennoch sind bei beiden Registern die Sprachdaten nicht direkt den jeweiligen Sprechern zuzuordnen. Beispielsweise können Foreneinträge (F) im Auftrag eines Abgeordneten durch zum Beispiel sein Sekretariat verfasst (6) oder als eine Art von Standard-Antwort (7) gesetzt werden.

(6)

Sehr geehrte Frau ...,

wegen Ihrer Frage zum Beamtenrecht haben wir uns mit dem zuständigen Referenten in Verbindung gesetzt, um Ihnen keine falschen Angaben zu machen.

Sobald uns von dort Informationen zugehen, werden wir Sie informieren. Haben Sie daher bitte noch etwas Geduld. Vielen Dank!

Mit freundlichem Gruß

Barbara Wesner

Sekretariat Landtagsbüro Clemens Lammerskitten MdL

Beitrag 18.03.2011, Thema Inneres und Justiz,

http://www.abgeordnetenwatch.de/clemens_lammerskitten-462-40286.html.

Zuletzt angeschaut am 06.09.2011.

(7)

Sehr geehrter Herr,

Fragen zum Themenbereich Finanzen und Steuern stellen Sie bitte auch den entsprechenden Fachpolitikerinnen meiner Fraktion. Dies wären z. B. Nicolette Kressl oder Ingrid Arndt-Brauer.

Beste Grüße

Andrea Nahles

Beitrag 06.01.2011, Thema *Finanzen*,

http://www.abgeordnetenwatch.de/andrea_nahles-575-37835.html.

Zuletzt angeschaut am 06.09.2011.

Solche Antwortentypen sind zu einem geringen Teil im Korpus enthalten.

Die tatsächlichen Reden der Abgeordneten vor dem Plenum liegen in den Protokollen nicht als Transkription vor, sondern sind in bestimmter Hinsicht bearbeitet worden. Der Stenografische Dienst des Deutschen Bundestages veröffentlicht am nächsten Werktag die Plenarprotokolle (P), die so der Öffentlichkeit zugänglich gemacht werden. Ebenfalls öffentlich zugänglich sind auch die Videoaufzeichnungen (V) der einzelnen Sitzungen. Mit Hilfe der Videoaufzeichnungen ist ein Vergleich zwischen tatsächlich geäußelter Sprache und den Protokollen möglich. Um die Bearbeitung der Redetexte durch den Stenografischen Dienst zu verbildlichen, werden hier Beispiele gegeben:

Modus	Beispiel
V1	< wir haben die niedrigste Arbeitslosigkeit [...] >
P1	<i>Wir haben die niedrigste Arbeitslosenquote [...]</i>
V2	< wir können also sehr zufrieden sein mit den Ergebnissen [...] >
P2	<i>Wir können mit den Ergebnissen der deutschen Wirtschaft also [...]</i>
V3	< da können wir als Politiker auch mal Danke sagen >
P3	<i>Wir sollten als Politiker an dieser Stelle einmal Danke sagen.</i>

Tabelle 5.2: Vergleich zwischen Videoaufzeichnung und der Bearbeitung der Plenarreden durch den Stenografischen Dienst des Deutschen Bundestages. Quelle: Stenografischer Bericht 17121 17. Legislaturperiode, 121. Sitzung, und Videoaufzeichnung der 17. Legislaturperiode 121. Sitzung Berlin.

Die in Tabelle 5.2 gegebenen Beispiele sind alle aus dem ersten Abschnitt der ersten Rede des Abgeordneten Philipp Rösler in der 121. Sitzung der 17. Legislaturperiode entnommen.¹⁹ In Tabelle 5.2 zeigt sich, dass die Bearbeitung der Reden von punktuellen Wortersetzungen V1/P1 über Veränderung des Satzaufbaus V2/P2 bishin zu Standar-

¹⁹<http://dipbt.bundestag.de/dip21/btp/17/17121.pdf>. Zuletzt angeschaut am 06.09.2011; <http://www.bundestag.de/Mediathek/index.jsp?isLinkCallPlenar=1&action=search&contentArea=details&ids=1220007&instance=m187&categorie=Plenarsitzung&destination=search&mask=search>. Zuletzt angeschaut am 06.09.2011.

disierungen von Äußerungen V3/P3 reichen.²⁰ Darüber hinaus sind weiter keine phonetisch/phonologischen Markierungen außer Pausenzeichen oder Gedankenstriche in den Protokollen enthalten beziehungsweise vermerkt. Reaktionen und Zwischenrufe des Plenums sind im Protokoll mit aufgenommen und grafisch hervorgehoben. Der Redetext des Abgeordneten wird unterbrochen und die entsprechende Beschreibung wird eingefügt, wie zum Beispiel in der gleichen Rede der 121. Sitzung von Philipp Rösler, gleicher Abschnitt:

(8)
(*Beifall bei der FDP und der CDU/CSU - Christine Scheel [BÜNDNIS 90/DIE GRÜNEN]: Dieses Wachstum gibt es trotz Ihrer Politik!*) (Plenarprotokoll)

Beide Register bewegen sich im politischen Rahmen und können daher als öffentlich und in gewisser Hinsicht als formal betrachtet werden.²¹ Die Redetexte stellen eher eine Verschriftlichung von mündlichen Äußerungen, die zu einem großen Teil auf Grundlage eines Manuskriptes geäußert werden, dar. Die Forenbeiträge sind rein schriftsprachlicher Natur.²² Beide Register sind immer mit Bezug auf die Adressaten, die Fragen der Bürger im Forum und den Abgeordneten im Plenum sowie der allgemeinen Öffentlichkeit, argumentativ. Im Plenum wie im Forum stellen sich die Abgeordneten Fragen beziehungsweise Problemen und beziehen dazu öffentlich Stellung.²³

5.3 Datenerhebung

5.3.1 Allgemeine Richtlinien zur Identifikation

Die ausgewählten Texte für die beiden Register werden tokenisiert (tok), hier gelten das orthografische Wort sowie Interpunktionszeichen als jeweils ein Token.²⁴ Darüber hinaus werden keine weiteren Interpretationen an den Primärdaten vorgenommen. Die Erhebung der Daten erfolgt im ersten Schritt rein quantitativ, indem Vier-Gramme aus den beiden Korpora erstellt und deren Frequenzen errechnet werden.²⁵

Alle Vier-Gramme (LB) wurden auf Grundlage der Textebene (txt) jedes Korpus errechnet. Die Interpunktion wird dabei nicht berücksichtigt. Damit werden Frequenzlisten der N-Gramme erstellt, die die einzelnen Wortformketten und deren Frequenz im Korpus enthalten (vgl. dazu Gries 2008, 12-13). Um die zu untersuchende Datenmenge einzugrenzen, wurden aufgrund der differierenden Größen der Korpora bei den Protokollen nur alle Bundles, die mindestens zehnmal vorkommen, und bei den Forumseinträgen nur

²⁰Eine detaillierte Analyse dieser redaktionellen Bearbeitung der Plenarreden durch den Stenografischen Dienst des Deutschen Bundestages gibt Friesdorf (2011).

²¹Innerhalb der Soziolinguistik könnte hier *standardnah* als allgemein einordnender Begriff im Vergleich zu dialektalen Varietäten des Deutschen und einem umgangssprachlichen Kontinuum verstanden werden. Eine differenzierte Diskussion dieser Problematik erfolgt hier nicht. Vgl. dafür z.B. Barbour und Stevenson (1998).

²²Auf eine grundlegende Diskussion von Mündlichkeit und Schriftlichkeit wird hier verzichtet. Für eine erste Diskussion vgl. u.a. Nerius (2005).

²³Für eine detaillierte Einordnung und Diskussion vgl. Borsch (2008).

²⁴Dazu wurde der Treetagger genutzt.

²⁵Mit Hilfe des Perl Scripts Text NSP, wurden die N-Gramme und ihre Frequenz errechnet.

alle Bundles, die mindestens fünfmal vorkommen, in die Untersuchung eingebunden. Die vollständigen Klassifikationen werden an Vier-Wortketten durchgeführt:

Ebene	Beispiel
txt	<i>Das wäre insbesondere vor dem Hintergrund der Tatsache [...]</i>
tok	Das wäre insbesondere vor dem Hintergrund der Tatsache
LB	<i>vor dem Hintergrund der</i> (59 P)

Tabelle 5.3: Beispiel für die Art der Datenerhebung. Protokoll 17. Legislaturperiode, 34. Sitzung Berlin.

Es werden keine NGramme berücksichtigt, die (9) und (12) von Satzgrenzen oder Textgrenzen unterbrochen sind, die (10) und (13) Eigennamen oder Adressen von zum Beispiel Abgeordneten beinhalten oder die (11) und (14) wie im Falle der Protokolle redaktionelle Anmerkungen enthalten:

- | | | | |
|------|--------------------------------------|------|--------------------------------------|
| (9) | <i>werden Auch in diesem</i> (7 F) | (12) | <i>MdB Sehr geehrte Frau</i> (100 F) |
| (10) | <i>bei der FDP Elke</i> (11 P) | (13) | <i>Republik 1 11011 Berlin</i> (9 F) |
| (11) | <i>Heiterkeit bei der FDP</i> (13 P) | (14) | <i>Zuruf von der SPD</i> (140 P) |

In (9) sind Satzgrenzen leicht mit Hilfe der Großschreibung am Satzanfang identifizierbar. Auch der inhaltliche Zusammenhang kann Aufschluss darüber geben. In jedem Fall werden solche Sequenzen von Biber und Kollegen ausgeschlossen, was in dieser Studie übernommen wird. Das Beispiel (12) zeigt den Schluss eines Forumseintrages eines Abgeordneten und den Anfang eines neuen Beitrages, der mit einer Begrüßung *Sehr geehrte Frau* beginnt. Die Beispiele in (10) und (13) illustrieren, dass Eigennamen von Personen oder direkte Adressenangaben in den Texten Verwendung finden. Beides ist für die Registerstudie nicht relevant und wird deshalb ausgeschlossen. Wichtig bei den Plenarprotokollen ist, dass die redaktionellen Anmerkungen von dem reinen Redetext der Abgeordneten getrennt werden. Hier sind in den Protokollen möglicherweise zwei unterschiedliche Register - die Abgeordnetenreden und die begleitenden Kommentare des Stenografischen Dienstes des Deutschen Bundestages - enthalten. Ersteres soll hier weiter untersucht werden. Eine Klassifikation der so erhobenen Daten soll nach den Taxonomien von Biber und Kollegen erfolgen.

5.3.2 Lokale Wiederholungen

Qualitativ ausgeschlossen sind bei Biber und Kollegen *lokale Wiederholungen*. Wie bereits gezeigt wurde, ist diese Richtlinie auch bei den bisherigen Studien nicht klar eingegrenzt. In jedem Fall wurden alle Sequenzen, die in einem Satz eingebettet mehrfach verwendet worden sind, als lokale Wiederholungen identifiziert, sodass Zitate und direkte Rückbezüge nicht in die Untersuchung mit einfließen. Sicher als lokale Wiederholungen identifizierbar wären Bundles wie *welche Daten das Unternehmen* (8 F). Hier wird kein allgemeiner Bezug deutlich, darüber hinaus wird diese Sequenz immer in demselben Satz

direkt wieder verwendet, wodurch ebenfalls keine allgemeinen diskursaufbauenden Funktionen abgeleitet werden können.

Weitere Beispiele für lokale Wiederholungen wären:

- | | |
|---|---|
| (15) <i>Vereinbarkeit von Familie und</i> (5 F) | (18) <i>in der großen Koalition</i> (10 P) |
| (16) <i>Europa und eine instabile</i> (6 F) | (19) <i>eines Gesetzes über die</i> (20 P) |
| (17) <i>von Bund und Ländern</i> (8 F) | (20) <i>der Europäischen Union die</i> (12 P) |

Diese Sequenzen können thematisch politischen Diskussionen oder den gängigen Begriffen der Politik zugeordnet werden. Konkrete Datenangaben wie *ab dem 1. Januar* (18 P) werden ebenfalls ausgeschlossen. Dennoch ist die Zuordnung nicht immer eindeutig zu treffen. Problematisch sind Sequenzen wie *über die Zusammenarbeit von* (17 P), die insgesamt 15-mal in folgenden, stark thematisch bezogenen Kontexten Verwendung finden:

- (21)
*das Gesetz über die Zusammenarbeit von Bundesregierung
und Bundestag*
(Plenarprotokoll)

Das Bundle selbst besitzt keinen stark thematischen Bezug, es wird dennoch 15-mal in genau diesem Kontext gebraucht. So könnte die Vier-Wortkette zu einer Neun-Wortkette erweitert werden. Nicht ganz klar ist, ob dieses Bundle eine lokale Wiederholung darstellt.

Cortes (2004) führt in der funktionalen Systematisierung Bundles ein, die einen direkten thematischen Bezug besitzen. Dazu wird keine genauere Richtlinie für deren Abgrenzung zu lokalen Wiederholungen gegeben, die wiederum keine Erwähnung in dieser Studie finden. Offen bleibt, ob Bundles wie *zwischen Ost und West* (18 P), *in diesem Hohen Haus* (15 P) oder *zu Protokoll zu geben* (21 P) ebenfalls solche Themen gebundenen Bundles sind. Im Vergleich zu den *Referential Bundles in the United States* (Biber, Conrad und Cortes 2004, 288) und *in diesem Hohen Haus* (15 P) stellt sich die Frage erneut, wie weit *lokal* ist.

Grundsätzlich bleibt das Problem der fehlenden konkreten Richtlinien für den Ausschluss einiger Wortketten aus der zu untersuchenden Menge und das Problem der fehlenden Vergleichbarkeit, da komplette Datensätze mit den so entstandenen Restmengen/Ausschlüssen nicht einsehbar sind.

5.3.3 Unterschiedliche Textebenen

Wie bereits erwähnt, beinhalten die Redetexte der Plenarprotokolle verschiedene Ebenen sprachlicher Äußerungen. Grundsätzlich sind redaktionelle Anmerkungen, die eigentlichen Reden der Abgeordneten und die Einberufung und Moderation der Plenarsitzung durch den Bundestagspräsidenten oder einen Stellvertreter enthalten. Die Aufgaben des

Plenarvorsitz sind durch die Geschäftsordnung des Deutschen Bundestages geregelt und erfolgen immer in gleicher Weise.²⁶

- (22) *Präsident Dr. Norbert Lammert: Die Sitzung ist eröffnet. Nehmen Sie bitte Platz. Platz gibt es ja reichlich. [...]*
- (23) *Ich rufe den Tagesordnungspunkt 1 auf: [...]*
- (24) *Sind Sie damit einverstanden? - Das ist offenkundig der Fall.*
- (25) *Präsident Dr. Norbert Lammert: Vielen Dank. - Die erste Nachfrage ist von der Kollegin Haßelmann.*

Tabelle 5.4: Beispiele für die Moderation der Plenarreden durch den Bundestagspräsidenten. Alle Zitate aus Protokoll 17. Wahlperiode 119. Sitzung Berlin.

Bundles dieser Äußerungen werden hier von den Reden der Abgeordneten, die einen wesentlich freieren Charakter besitzen, getrennt und nicht weiter betrachtet. Dadurch, dass die Moderation und die eigentlichen Reden der Abgeordneten unterschiedlich Funktionen haben, ist nicht ganz klar, in wie weit diese Äußerungen zum selben Register zählen. Da die Moderation der Plenarsitzungen einen durchaus starken Kontext- und Themenbezug aufweist, werden sie hier als lokale Wiederholungen gewertet. So werden diese nicht in der Auswertung näher betrachtet, wenngleich eine genauere Betrachtung auch dieser Äußerungen nicht grundsätzlich für Register- und Lexical Bundles-Studien ausgeschlossen ist.

²⁶Die Geschäftsordnung ist unter http://www.bundestag.de/dokumente/rechtsgrundlagen/go_btg/go06.html frei verfügbar.

6 Ergebnisse der Studie

6.1 Allgemeine Diskussion der Ergebnisse

Die NGramm-Analyse²⁷ ergibt, bei einem *cut off* der Vorkommen von fünf bei den Foreneinträgen und von zehn bei den Protokollen, folgende Zahlenverhältnisse:

Teilmengen NGramme	Korpus Politikforum	Korpus Plenarprotokolle
Alle Vier-Gramme	1705	7403
Lokale Wiederholung	544	1787
Struktureller Ausschluss	702	4439
Lexical Bundles	459	1177

Tabelle 6.1: Übersicht der Datengrundlage für beide Korpora mit Angabe der lokalen Wiederholung, der strukturellen Ausschlüsse und der Lexical Bundles.

Rund 32 Prozent der NGramme bei den Foreneinträgen und rund 15 Prozent der Sequenzen in den Protokollen werden nach allen Ausschlüssen in die inhaltliche Auswertung übernommen. Interessant ist hierbei, dass wesentlich mehr Sequenzen in den Einträgen auf *abgeordnetenwatch.de* als Lexical Bundles identifiziert werden als in denen der *Plenarprotokolle*. Die Verteilung einer Auswahl an Wortarten in beiden Korpora soll ebenfalls betrachtet werden:

Wortartengruppen	Anteile in Politikforum	Anteile in Plenarprotokolle
Nomen	26,8 Prozent	31,7 Prozent
Verben	14,5 Prozent	14 Prozent
Funktionswörter	31,6 Prozent	29,2 Prozent
Adjektive/Adverbien	23,9 Prozent	15 Prozent
Konjunktion	5,5 Prozent	5,4 Prozent

Tabelle 6.2: Übersicht der gerundeten prozentualen Anteile der exemplarisch ausgesuchten Wortarten in beiden Korpora.

Jedes Korpus ist mit POS-Annotationen²⁸ versehen. Dadurch lassen sich die Anteile der einzelnen Wortarten in beiden Korpora ermitteln. Die Tabelle 6.2 zeigt den prozentualen Anteil der verbalen und der nominalen Wortarten sowie einiger Funktionswörter, der

²⁷Die Datensets der Vier-Wortketten beider Korpora sowie die Korpora selbst werden über ANNIS2 veröffentlicht.

²⁸POS steht für *Part-of-Speech* und meint hier die Wortarten-Annotation aller Token, die automatisch mit Hilfe des TreeTaggers ermittelt werden.

Adverbien und Adjektive.²⁹ Die Anteile der Gruppe *Verben* sind in beiden Registern geringer als die der Nomen und der Funktionswörter, die jeweils etwa ein Drittel aller Wortarten stellen. Es zeigt sich eine sehr ähnliche Verteilung der Wortarten in beiden Registern. Damit ist für die weitere Diskussion klar, dass beide Korpora etwa die gleichen Anteile an allen untersuchten Wortarten besitzen und somit eventuelle strukturelle Unterschiede zwischen diesen beiden nicht auf Grund von komplett divergierenden Wortartenverteilungen beruhen können.

6.2 Strukturelle Klassifikation

Wie bei der Gegenüberstellung der Vorgängerstudien deutlich wurde, differieren die strukturellen Klassifikationen stark. Hier wird der auf Phrasenköpfen basierende Ansatz, hinsichtlich NP-, PP- und VP-basierten Bundles verfolgt. Dieses Vorgehen wird durch die Menge der Daten motiviert, wodurch eine feinere Unterscheidung problematisch erscheint. Darüber hinaus sind die deutschen Sequenzen wesentlich heterogener in Bezug auf ihre Struktur als die in den Vorgängerstudien gegebenen Beispiele. Der alleinige Bezug auf Korrelate zu NPn, PPn oder VPn scheint bei den deutschen Daten allerdings nicht auszureichen:

NP/PP-basierte Bundles		VP-basierte Bundles	
(26)	<i>auch eine Frage der</i> (12 P)	(32)	<i>Wir gehen davon aus</i> (23 P)
(27)	<i>vielen Dank für Ihre</i> (575 F)	(33)	<i>ich danke Ihnen für</i> (25 F)
(28)	<i>den Druck auf die</i> (11 P)	(34)	<i>die ich gerne beantworte</i> (6 F)
(29)	<i>für den Fall dass</i> (6 F)	(35)	<i>Das ist der Punkt</i> (22 P)
(30)	<i>mit dem Hinweis auf</i> (11 P)	(36)	<i>stellen wir fest dass</i> (12 P)
(31)	<i>vor allen Dingen auch</i> (17 P)	(37)	<i>Ich kann nur sagen</i> (38 P)

Tabelle 6.3: Beispiele verschiedener Strukturen von Lexical Bundles in beiden Korpora.

Die prototypischen NP-basierten Bundles im Englisch beginnen jeweils mit einem Determinier, bei *auch eine Frage der* (26) ist das nicht der Fall. Hier ist das Adverb *auch* an die NP adjungiert und es ergibt sich eine neue Struktur der Phrase ähnlich wie bei der Sequenz *vor allen Dingen auch*, der das Adverb *auch*, welches wiederum einer andern Phrase zuzuordnen ist, folgt.

Beide Fälle beinhalten so mehr als ein nur NP- oder PP-Fragment. Weiterhin sind solche Bundles generell im Mittelfeld des Satzes vorzufinden:

²⁹Die Gruppe *Nomen* umfasst alle Nomen und Eigennamen mit NN und NE, die Gruppe *Verben* alle finiten und infiniten Hilfs-, Modal- und Vollverben mit VAFIN, VVFIN, VMFIN, VVINFIN, VVPP, VVIZU, VMINF, VAINF und VVIMP, die Gruppe *Funktionswörter* viele pronominale Wortarten, Konjunktionen und Artikel mit ART, PPER, APPR, PRF, PRELS, PAV, PDAT, PDS, PIS, PPOSAT und PWAV und die Gruppe *Adjektive/Adverbien* alle Adjektive und Adverbien mit ADJA, ADJD und ADV.

(38)

Es ist auch eine Frage der Fairness, liebe Kolleginnen und Kollegen der Opposition, dass Kinder im Steuerrecht nicht weiter wie Bürger zweiter Klasse behandelt werden. (Plenarprotokoll)

(39)

Außerdem ist es für mich in diesem Zusammenhang und vor allen Dingen auch vor dem Hintergrund der Debatte am heutigen Vormittag völlig unverständlich [...] (Plenarprotokoll)

Dadurch werden Bundles, die nicht eindeutig den Anfang einer NP oder PP oder/und deren phrasale Modifikatoren beinhalten, als Mittelfeld-Bundles (MF) klassifiziert. Darunter werden auch Bundles wie

(40) *vor allen Dingen die* (26 P)

(42) *in erster Linie die* (20 P)

(41) *wie zum Beispiel die* (20 P)

(43) *in erster Linie um* (11 P)

eingeorordnet. Die Maxime, Bundles werden 'regardless of their structural status' (Biber et al. 1999, 990) klassifiziert, wird so erfüllt, wenn auch keine solchen Beispiele in den englischen Bundles vorzufinden sind.

Die Postmodifikatoren einiger Phrasen sind ebenfalls sehr unterschiedlich. Bei dem Bundle *für den Fall dass* (29) ist der Anschluss satzartig, bei *mit dem Hinweis auf* (30) hingegen phrasal. Um diesen Unterschied zu kennzeichnen wird ersterer Fall aus der Menge der PP-basierten Bundles herausgenommen und mit einer eigenen Kategorie (CP/NP,PP) zusammen mit NP-basierten Bundles, die einen ähnlichen Anschluss besitzen, versehen. Die Bundles, die beispielsweise direkt mit einer Konjunktion oder einem Relativpronomen anfangen und/oder die keine nominalen Bestandteile beinhalten, sind in einer eigenen Gruppe (CP) zusammengefasst. Eine Unterscheidung bezüglich enthaltener Konjunktionen oder Nebensatz einleitenden Pronomen wird auch bei den verbalen Bundles getroffen: Die Unterscheidung zwischen VP-Bundles, die subordinierende Elemente wie Konjunktionen, Relativpronomen oder Teile von Haupt- und Nebensatz enthalten (NESVP), und solchen, die dies nicht tun (HSVP), gilt als Richtlinie zur Identifikation. Infinitive mit *zu* (INFVP) werden separat aufgenommen.

Wie aus Tabelle 6.4 hervorgeht, sind typische NPn im Gegensatz zu den typischen PPn in beiden Registern nicht viel gebraucht. Der größte Teil der Bundles enthält verbale Bestandteile ohne Nebensatzbestandteile oder Nebensatz einleitende Wörter (HSVP), die in beiden Registern etwa ein Drittel entsprechen. Das Verhältnis zwischen VP-basierten und nicht VP-basierten Bundles ist ausgewogen: Bei einer Gegenüberstellung von verbalen (NESVP, INFVP und HSVP) und nicht verbalen Bundles (NP, PP, CP/NP,PP und MF) zeigt sich, dass sich in den Registern beide Gruppen in etwa gleich verhalten: Etwa 54 Prozent der Bundles im Forum-Korpus und 59 Prozent in den Protokoll-Korpus enthalten verbale Bestandteile. In jedem Fall unterscheiden sich die Verteilungen der strukturellen Typen zwischen den beiden Registern nicht wesentlich.

Struktur	Politikforum	Plenarprotokolle
NP	52	33
PP	86	130
CP/NP,PP	22	36
CP	22	99
HSVP	149	431
NESVP	60	204
INFVP	4	38
MF	64	178

Tabelle 6.4: Verteilung der strukturellen Typen in beiden Registern. Hier werden die vollen Vorkommen der einzelnen Wortketten angegeben.

Weiterhin ist der Großteil der Bundles wie in den Vorgängerstudien strukturell unvollständig. Hier einige Beispiele:

Vollständig	Unvollständig
<i>Ich bin der Auffassung</i>	<i>Das ist ein ganz</i>
<i>Das ist die Situation</i>	<i>dann kann ich nur</i>
<i>auf der anderen Seite</i>	<i>auch im Hinblick auf</i>
<i>in Anspruch zu nehmen</i>	<i>ein guter Tag für</i>

Als strukturell vollständige Bundles sind komplette Infinitive mit *zu* wie *in Anspruch zu nehmen* und vollständige Satzeinheiten wie *Das ist die Situation* gewertet. Weiterhin werden auch Strukturen wie *Ich bin der Auffassung* als strukturell vollständig betrachtet. In beiden Korpora sind nur etwas über 20 Prozent der Bundles auf diese Art als strukturell komplett eingeordnet.

Die einzelnen Gruppen, trotz der feineren Klassifizierung, sind strukturell nicht so homogen wie die der Studien von Biber und Kollegen. Das liegt zum größten Teil an der variablen Wortstellung des Deutschen, mit einem flexiblen Mittelfeld, einer Satzklammerstruktur und einer V2/Vletzt-Asymmetrie bei den Haupt- und Nebensätzen:

- | | |
|---|---|
| (44a) <i>es nicht möglich ist</i> (6 F) | (44b) <i>ist es nicht möglich</i> (6 F) |
| (45a) <i>halte ich es für</i> (13 F) | (45b) <i>Ich halte es für</i> (9 F) |
| (46a) <i>geht es um die</i> (9 F) | (46b) <i>Es geht um die</i> (8 F) |

Diese Bundles zeigen diese Variation deutlich. Die verschiedene Linearisierung in Haupt- und Nebensätzen zeigt sich unter anderem so:

- (47)
*Mir stellt sich dabei die Frage, weshalb **es nicht möglich ist** im Rahmen der KfW-Bank einen Kredit mit sehr niedrigen Zinsen für Studenten anzubieten, der ein komplettes Studium abdeckt und eine möglichst lange Rückzahlungszeit, 10-30 Jahre, einplant?* (Politikforum)

(48)

*Allerdings **ist es nicht möglich**, dass ich mich als deutscher Bundestagsabgeordneter im europäischen Ausland für Änderungen derer Visa Bestimmungen einsetze.* (Politikforum)

Nicht nur zwischen Haupt- und Nebensätzen lässt sich eine solche Varianz beobachten:

(49)

*Insgesamt **halte ich es für** begrüßenswert, dass sich die Bundeswehr als Berufsfeld vorstellt.* (Politikforum)

(50)

***Ich halte es für** ein Gebot der Höflichkeit und des Respekts voreinander, dass man von Menschen, mit denen man in Kontakt steht, mehr weiß als einen Namen.* (Politikforum)

In (49) und (50) ist die Vorfelddbesetzung entscheidend für die Art der Linearisierung der Bundles. Bei dem Bundle *halte ich es für* ist das Vorfeld mit einem Adverb besetzt, sodass das Subjekt *ich* im Mittelfeld des Satzes realisiert wird. In gleicher Weise sind die Bundles *geht es um die* und *Es geht um die* zu erklären. So erscheinen die deutschen Daten heterogener und die Klassifikation nach bestimmter Serialisierung wie bei Biber et al. (1999) wäre problematisch und würde keine übersichtliche Systematisierung ermöglichen. Durch die variable Wortstellung im Deutschen ist noch ein anderes Phänomen in den deutschen Daten zu erkennen. Einige Bundles scheinen Kollokationsbeziehungen zu besitzen:

(51)

Bundle: *dass es sich um* (8 F)

*[...] **dass es sich um** zwei unterschiedliche Alterssicherungssysteme **handelt**.*

*[...] **dass es sich um** ein reines Präventionsgesetz in einem besonders gelagerten Fall **handelt**, [...]*

*[...] **dass es sich um** einen 'Einzelfall' **handelt**?*

(Politikforum)

(52)

Bundle: *es sich bei der* (11 P)

*[...] weil **es sich bei der** Fotovoltaik **um** eine junge Technologie **handelt**, [...]*

*Selbst nach Auffassung ihrer Befürworter **handelt es sich bei der** PID **um** eine Methode, [...]*

*Zudem **handelt es sich bei der** Asse **um** ein Forschungsbergwerk des Bundes.*

(Plenarprotokoll)

Beide Bundles in (51) und (52) werden immer mit dem Verb *handeln* verwendet, gleich welche syntaktische Struktur der Äußerung zugrunde liegt. Nach Evert (2009) können Kollokationen als empirische Vorhersagen über die Nachbarschaft eines Wortes bezeichnet werden (ebd., 1218), wie sich es hier zeigt. Beide Bundles besitzen eine unterschiedliche Struktur, ersteres *dass es sich um* leitet in jedem Fall einen Objektnebensatz ein, das Zweite *es sich bei der* lässt eine Nutzung in Haupt- und Nebensätzen zu. Weiterhin beinhalten sie selbst nur Artikel, Pronomen, Präpositionen oder/und Konjunktionen, die von sich aus keine lexikalischen Informationen beinhalten. Ähnliche kollokative Beziehungen scheinen zwischen einigen Bundles und ihrer Umgebung zu bestehen:

(53)

Bundle: *dass es zu einer* (15 P)

[...] **dass es zu einer** deutlichen Absenkung des Rentenbeitrags kommt, [...]

[...] **dass es zu einer** Lösung kommt [...]

[...] **dass es zu einer** Erhöhung der Regelsätze kommen wird.

(Plenarprotokoll)

(54)

Bundle: *ist für mich nicht* (5 F)

Dieser Sachverhalt **ist für mich nicht nachzuvollziehen**, wenn man [...]

Worauf sich hingegen Ihre Zahlen beziehen, **ist für mich nicht nachvollziehbar**.

[...] es **ist für mich nicht nachvollziehbar**, dass [...]

(Politikforum)

Die Bundles treten immer mit denselben lexikalischen Einheiten auf. Das Bundle *ist für mich nicht* tritt in allen Kontexten mit einer Form von *nachvollziehen* in unterschiedlichen syntaktischen Strukturen auf. Eine solch eindeutige Zuordnung ist aus dem Bundle selbst nicht ableitbar, die ganze Phrase *ist für mich nicht nachvollziehbar/nachzuvollziehen* ist in keiner Weise idiomatisch. Bundles mit wenig lexikalischem Inhalt scheinen demnach Kollokationsbeziehungen in Form von lexikalischen Anschlüssen/Erweiterungen aufzuweisen. Dies müsste genauer und an allen Bundles gleichermaßen untersucht werden. In den Studien von Biber und Kollegen werden solche Bundles, die nur Funktionswörter enthalten, nicht thematisiert oder als nicht relevant von der Diskussion ausgeschlossen. Weiterhin ist nicht nur die variable Wortstellung des Deutschen Ursache für diese heterogenen Daten, auch die Morphologie verhindert, dass einige Bundles nicht zusammengefasst werden können und somit die Anzahl der verschiedenen Typen ansteigt:

- | | | | |
|------|---------------------------------------|------|------------------------------------|
| (55) | <i>im Zusammenhang mit dem</i> (81 P) | (56) | <i>im Hinblick auf die</i> (187 P) |
| | <i>im Zusammenhang mit der</i> (77 P) | | <i>im Hinblick auf den</i> (38 P) |
| | <i>im Zusammenhang mit den</i> (24 P) | | <i>im Hinblick auf das</i> (26 P) |

Hier enden die präpositionalen Postmodifikatoren mit einem Artikel, der je nach Kasus, Numerus und Genus des Nomen in der nachfolgenden NP variiert. Diese Varianten

weist das Englische so nicht auf, weshalb die Bundles eine einheitlichere Struktur besitzen. Eine Lösung für dieses Problem könnte eine Datenerhebung auf Grundlage der Lemma-Ebene der Korpora sein. Diese kann wie die POS-Ebene ebenfalls mit dem Tree-Tagger automatisch erzeugt werden. Die jeweiligen Artikel werden dann einheitlich ohne Berücksichtigung der verschiedenen paradigmatischen Varianten annotiert:

Ebene	Beispiel	Beispiel
txt	<i>Im Zusammenhang mit dem</i>	<i>im Hinblick auf die</i>
POS	APPRART NN APPR ART	APPRART NN APPR ART
Lemma	im Zusammenhang mit d	im Hinblick auf d
LB	im Zusammenhang mit d (205 P)	im Hinblick auf d (265 P)

Tabelle 6.5: Darstellung der verschiedenen automatisch annotierten Ebenen POS und Lemma.

Ein weiterer Vorteil dieser Vorgehensweise ist, dass auch die Großschreibung keine Berücksichtigung findet. So werden Bundles, die zum Beispiel eine Großschreibung von Präpositionen aufgrund der Position im Satz - Satzanfang - beinhalten, mit Bundles, die solche Varianten nicht aufweisen, zusammengefasst. Die Struktur bleibt in jedem Fall die gleiche (vgl. dazu Tognini-Bonelli 2001, 82-83) Damit werden alle Vorkommen zusammengefasst, deren Zahlen dann wesentlich aussagekräftiger sind. Diese Problematik wird in allen Lexical Bundles Studien wenig oder nicht thematisiert. Altenberg (1998) bemerkt für das Englische Formvarianzen wie zum Beispiel durch das Tempus, die unterschiedliche Typenanzahlen produzieren (vgl. Altenberg 1998, 104). Diese Varianz besteht ebenfalls im Deutschen. Bei *im Zusammenhang mit d* wären das insgesamt 205 und bei *im Hinblick auf d* insgesamt 265 Vorkommen. Die Typenzahl an verschiedenen Bundles reduziert sich und lässt ein genaueres Zahlenbild zu.

Mit diesem Schritt werden bereits interpretierte Daten untersucht. Lemmata sind lexikalische Einheiten, dennoch wird die Ebene der Primäraußerungen verlassen. Fraglich ist, ob die Lexical Bundles dann grundsätzlich einen abstrakten Status erhalten. Sie wären dann keine Wortformketten, sondern Lemmasequenzen, die keinen so starken Bezug zum Kontext aufweisen.

So sind die eingangs in Tabelle 6.4 dargestellten Zahlen nicht so aussagekräftig wie die der Englischen Daten. Auf Grundlage dieser weitere Aussagen zu treffen erscheint in Hinblick auf die genannten Probleme schwierig. Wenn die Ergebnisse vergleichbar sein sollen, dann müsste die Datenerhebung auf der Lemma-Ebene basieren. Ein detaillierter Vergleich der Strukturen zu den englischen Lexical Bundles ist auch aufgrund der heterogenen Zusammensetzung der deutschen Daten schwierig. Die Darstellung der unterschiedlichen Satzstellungen sind hingegen nicht ohne eine feinkörnige Systematisierung möglich, eine Reduzierung auf drei Typen - NP-, PP- und VP-basiert - ermöglicht wiederum einen leichteren Vergleich zu den Vorgängerstudien.

6.3 Funktionale Klassifikation

6.3.1 Mögliche Klassifikationen

Wie aus der bisherigen Diskussion hervorgeht, ist die linguistische Auswertung der Lexical Bundles zumindest im Deutschen problematisch und somit nicht immer eindeutig. Die funktionalen Systematisierungen der Vorgängerstudien sind vielfältig und besitzen zum Teil den Registern angepasste Schwerpunkte. Wie gezeigt wird, erweist sich die Übertragung der Konzepte auf die deutschen Bundles grundsätzlich als schwierig.

Nach Biber, Conrad und Cortes (2004) identifizieren *Referential Bundles*

' [...] an entity or single out some particular attribute of an entity as especially important (Biber, Conrad und Cortes 2004, 393).'

Viele der Bundles aus beiden Korpora können analog zu den in den Studien gegebenen Beispielen eingeordnet werden, siehe Tabelle 6.6:

Politikforum	Subkategorie <i>Referential Bundles</i>
<i>vor dem Hintergrund der</i> (17 F)	multifunktionale Referenz
<i>im Hinblick auf die</i> (15 F)	Referenz auf eine nicht greifbare Entität
<i>in Bezug auf die</i> (13 F)	Referenz auf eine nicht greifbare Entität
<i>bis Ende des Jahres</i> (6 F)	zeitliche Referenz
<i>in der vergangenen Woche</i> (5 F)	zeitliche Referenz
<i>In den letzten Jahren</i> (5 F)	zeitliche Referenz
Plenarprotokoll	Subkategorie <i>Referential Bundles</i>
<i>auf der Ebene der</i> (13 P)	multifunktionale Referenz
<i>ein hohes Maß an</i> (15 P)	quantitative Referenz
<i>für die Zukunft der</i> (12 P)	Referenz auf eine nicht greifbare Entität
<i>mit Blick auf den</i> (23 P)	Referenz auf eine nicht greifbare Entität
<i>auf dem Gebiet des</i> (12 P)	lokale Referenz
<i>in den letzten Jahren</i> (236 P)	zeitliche Referenz

Tabelle 6.6: Übersicht über mögliche *Referential Bundles* im Deutschen aus beiden Korpora.

Bundles wie *auf dem Gebiet des* und *vor dem Hintergrund der* referieren lokal oder in abstrakter, nichtgreifbarer Weise auf die folgende Entität, Bundles wie *In den letzten Jahren* auf zeitliche Zusammenhänge. Weitere multifunktionale Bundles könnten *gerade in den letzten* (10 P) und *schon in der letzten* (10 P) sein. Hier greifen zwei Funktionen, indem ein mit *schon* beziehungsweise *gerade* skaliertes, zeitlicher Bezug kommuniziert wird. Die Definition der multifunktionalen Bundles schließt dies grundsätzlich nicht aus, eine solche Doppelfunktionalität wird so nicht thematisiert.

Die Bundles sollen keinen direkten Bezug zu Themen der jeweiligen Texte besitzen. Beispielsweise bei Biber, Conrad und Cortes (2004) wird das Bundles *in the United States*

als lokales Referential Bundle eingeordnet (ebd., 388).³⁰ Bundles wie *in Ost und West* (7 F), *in der Bundesrepublik Deutschland* (7 F) und *in der ehemaligen DDR* (6 F) müssten dann ebenfalls als solche betrachtet werden. Dennoch sind gerade die Angaben *in Ost und West* und *in der ehemaligen DDR* durchaus Referenzen, die dem politischen Rahmen entsprechen. Auch hier ist das Konzept der lokalen Wiederholungen nicht klar und wann *lokal* lokal ist, scheint einer Einzelfallentscheidung zu unterliegen.

Stance Bundles drücken unterschiedliche Haltungen des Sprechers aus. Diese Zuordnung kann auch in den deutschen Daten realisiert werden.

Plenarprotokoll	Subkategorie <i>Stance Bundles</i>
<i>Es ist völlig klar</i> (11 P)	epistemisch
<i>bin der festen Überzeugung</i> (25 P)	epistemisch
<i>hätte mir gewünscht dass</i> (16 P)	Wunsch
<i>will ich gar nicht</i> (11 P)	Wunsch
<i>Ich möchte noch auf</i> (11 P)	Wunsch/Intention
<i>Es wäre schön wenn</i> (15 P)	epistemisch
Politikforum	Subkategorie <i>Stance Bundles</i>
<i>Ich bin der Meinung</i> (11 F)	epistemisch
<i>Ich bin mir sicher</i> (8 F)	epistemisch
<i>Es ist richtig dass</i> (17 F)	epistemisch
<i>ist meiner Meinung nach</i> (8 F)	epistemisch
<i>Ich möchte Sie bitten</i> (5 F)	direktive Obligation
<i>es nicht möglich ist</i> (6 F)	Fähigkeit/Möglichkeit

Tabelle 6.7: Übersicht über mögliche *Stance Bundles* im Deutschen aus beiden Korpora.

Wie die Tabelle 6.7 zeigt, lassen sich deutsche Entsprechungen in Hinblick auf diese Klassifikation finden. Dennoch ist nicht immer eindeutig, welche Zuordnung wie im Falle von *Ich möchte noch auf* (11 P) gewählt werden soll. Nicht klar ist, in wie weit die Einteilung zu den Wunschaussagen denen der Intensionsaussagen übergeordnet ist, da durch Verben wie *möchten* auch Vorhaben/Intentionen vermittelt werden können:

(57)

Ich möchte noch auf das Thema Integrationskurse zu sprechen kommen.

(Politikforum)

Hier stellt sich die Frage, ob die lexikalische Ebene, mit dem Schwerpunkt auf der grundsätzlichen Bedeutung von *möchten* als Wunsch, vor der kontextuellen, situativen Ebene, mit der Funktion von *möchten* als Ausdruck der Intention, steht. Biber, Conrad und Cortes (2004) stellen heraus, dass die Bundles je nach Kontext, verschiedene Funktionen übernehmen können (ebd., 383-384). In Anbetracht dieser Annahme wären solche Bundles multifunktional. Gerade die Verben, die Wunsch/Willen des Sprechers ausdrücken,

³⁰Die Problematik um Eigennamen und Tokenanzahl ist bereits im Kapitel 3.4.2 diskutiert.

werden vielfältig genutzt. Biber und Barbieri (2007) kategorisieren *What I want to do is* als *Discourse Organizer Topic Introduction/Focus* (ebd., 271), *I want to do* wird bei Biber, Conrad und Cortes (2004) ebenfalls als *Discourse Organizer* klassifiziert, weil dies die spezialisierte Funktion des Bundles sei (ebd., 386, 392). Die negierte Form dieses Bundles *I don't want to* hingegen wird als *Desire Bundle* bei beiden Studien identifiziert (vgl. Biber, Conrad und Cortes 2004, 386; Biber und Barbieri 2007, 271). Bundles wie *I want you to* fallen ebenfalls unter die *Stance Bundles Desire* (Biber und Barbieri 2007, 271), in denen wiederum eine direktive Aufforderung unabhängig vom Kontext enthalten ist. Somit sind die Kategorien nicht unbedingt voneinander zu trennen. Generell leitet sich hieraus die Frage ab, welche Einheit ein Lexical Bundle darstellt. Per Definition ist keine Regelung getroffen, es scheinen aber zum Einem Vererbungsbeziehungen wie bei *I want to do* und *What I want to do* zu bestehen, zum anderen ist damit ein Konzept von Einheiten verbunden, deren minimaler Umfang indirekt erfragt ist. Beispielsweise ist das Bundle *im Zusammenhang mit dem* (81 P) ein prototypisches *Rerefential Bundle*, dessen Postmodifizierer noch eine lexikalische Einheit offen lässt, die dann als *Intangible Framing Attributes* klassifiziert wird. Bundles wie *ich in diesem Zusammenhang* (14 P), *In diesem Zusammenhang ist* (5 F) oder *von dieser Stelle aus* (15 P) besitzen keine offene Postmodifiziererphrase. Hier ist das, worauf in bestimmter Art und Weise referiert wird, pronominal anaphorisch mit *diesem/dieser* ausgedrückt. Ist die Funktionszuweisung *Referential Bundle* in dieser Form ebenfalls gegeben? Und mit Rückbezug auf die Frage nach dem Umfang von Lexical Bundles: Die Strukturen der beiden Bundles unterscheiden sich untereinander und von denen der englischen *Referential Bundles*. Wenn vor diesem Hintergrund diese als *Referential Bundles* klassifiziert werden können, nehmen diese Strukturen +PPER oder +VAFIN die funktionale Zuweisung? Oder muss davon ausgegangen werden, dass hier eigentlich eine Drei-Wortkette vorliegt?

Wenn Bundles solche Beziehungen untereinander besitzen, so müsste die Klassifikation dies zum Ausdruck bringen. Bisher wurde ohne Berücksichtigung eventueller hierarchischer Beziehungen direkt auf Grundlage des Kontextes klassifiziert. Dennoch erscheint hier eine abstrahierende Klassifikation möglich, wenn nicht notwendig.

Discourse Organizers zeigen nach Biber und Barbieri (2007) den informationellen Status der Aussage an. Diese Zuordnungen gelingen nur an Teilen der deutschen Daten:

Die Beispiele der *Discourse Organizers* aus den Studien von Biber und Kollegen scheinen eher auf die Kommunikation des jeweiligen Themas der Äußerung zu zielen. *Purpose of this course, our goal is to, want to talk about* oder *in this chapter we* können grundsätzlich Topik angeben und tun dies auch in direkter, ausformulierter Weise, sie scheinen intuitiv auch die globalen Themenangaben des situativen Kontexts zu umreißen. Topik kann ebenfalls einen viel kleineren Sachverhalt innerhalb *dieses Kapitels*, ausgehend von *in this chapter we*, darstellen.

Plenarprotokoll	Subkategorie <i>Discourse Organizers</i>
<i>um die es geht</i> (11 P)	Topik Einführung/Fokus
<i>ist ein Thema das</i> (10 P)	Topik Elaboration
<i>hat deutlich gemacht dass</i> (10 P)	Topik Elaboration
Politikforum	Subkategorie <i>Discourse Organizers</i>
<i>Es geht um die</i> (8 F)	Topik Einführung/Fokus
<i>auf der anderen Seite</i> (16 F)	Topik Elaboration
<i>Auf der anderen Seite</i> (9 F)	Topik Elaboration
<i>Zu Ihrer zweiten Frage</i> (7 F)	Topik Einführung/Fokus
<i>Meine Frage an Sie</i> (6 F)	Topik Einführung/Fokus

Tabelle 6.8: Übersicht über mögliche *Discourse Organizers* im Deutschen aus beiden Korpora.

In den Korpora *Plenarprotokoll* und *Politikforum* finden sich Bundles, die auf eine Interaktion von Sprecher zu Rezipient aufweisen:

- | | |
|--|--|
| (58) <i>Lassen Sie uns das</i> (11 P) | (62) <i>Deshalb bitte ich Sie</i> (11 P) |
| (59) <i>Bitte erklären Sie mir</i> (6 F) | (63) <i>Das ist Ihre Politik</i> (10 P) |
| (60) <i>im Gegensatz zu Ihnen</i> (13 P) | (64) <i>wende mich an Sie</i> (9 F) |
| (61) <i>Ich werde Ihnen gerne</i> (6 F) | (65) <i>Ich bitte Sie daher</i> (5 F) |

Hyland (2008) verfolgt, wenn auch mit anderem Schwerpunkt, eine Klassifikation, die grundsätzlich zwischen *Research-oriented*, *Text-oriented* und *Participant-oriented Bundles* unterscheidet (vgl. ebd., 13-14). Auch die Studien von Biber und Kollegen greifen diesen Aspekt teilweise mit der Kategorie *Special Conversational Functions* auf. Zumindest der Ansatz, Bundles auch danach zu klassifizieren wird durch die oben gegebenen Beispiele motiviert. Alle Sequenzen beinhalten einen direkten Rezipientenbezug wie *Sie/Ihnen*. Eine feinere Kategorisierung bezüglich dieser Bundles erscheint sinnvoll, da ihre Struktur und Funktion nicht nur auf *Politeness Bundles* wie *thank you very much* zurückzuführen sind (vgl. Conrad und Biber 2004, 66). Beide Register zeichnen sich durch diesen Bezug auch aus. Die einzelnen Problemfälle wie Topik, Fokus und Multifunktionalität werden nachfolgend ausführlicher diskutiert.

6.3.2 Topik und Fokus

Die Studien von Biber und Kollegen scheinen mit dem Topikbegriff ein weitgefasstes, in einer hohen Textebene angesiedeltes Thema zu verbinden. *[D]as, worüber man etwas sagt* ist hingegen in verschiedenen Textebenen kommuniziert (Musan 2010, 26). Möglicherweise umfasst der Begriff *Topik* bei Biber und Kollegen das *Diskurstopik*, welches sich auf größere Texteinheiten bezieht. In der Forschung zur Informationsstruktur hingegen wird Topik zwar sehr verschieden aber dennoch auch als Satztopik diskutiert. In der Forschung wird allgemein davon ausgegangen, dass Topik in einer Topik-Kommentar-Struktur oder auch -Dimension fest verankert ist (vgl. Musan 2010, 25; Jacobs 2001, 641). Und weiter:

Topik kann in lexikalischer, syntaktischer und/oder intonaler Weise ausgedrückt werden. Jacobs (2001) geht darüber hinaus davon aus, dass es prototypische Topik-Kommentar-Konstruktionen gibt, die in verschiedener Weise zusammenspielen können und so Topik ausdrücken (ebd., 642-643). Eine prototypische Konstruktion für Topik ist nach Jacobs (2001) das Frame-Setting (ebd., 655). Die Klassifikation nach Biber und Kollegen gibt bei den *Referential Bundles framing attributes* an. Ein kontextueller Bezug wird ebenfalls genannt und mit *shown in figure N/as shown in figure* auch mit Beispielen belegt, umfasst aber dennoch nicht den Begriff *Topik*. Dieser wird nur in den *Discourse Organizers* aufgenommen (vgl. Biber, Conrad und Cortes 2004, 384-388, Tabelle A.1). Diese Bundles wie *Vor dem Hintergrund der* (18 P) können dennoch Topik ausdrücken:

(66)

Vor dem Hintergrund der Anfragen und der Antworten ist mir heute etwas unverständlich, warum Sie diesen Antrag vorlegen; denn Ihre Forderungen in den Anfragen sind jeweils weitestgehend erfüllt worden. (Plenarprotokoll)

Jabocs (2001) definiert 'Frame-setting' wie folgt:

'In (X Y), X ist the frame for Y iff X specifies a domain of (possible) reality to which the proposition expressed by Y is restricted (Jacobs 2001, 656).'

Jacobs argumentiert weiter, dass Topik nicht begrenzt auf *Aboutness* ist (ebd., 643). Somit gibt X *Vor dem Hintergrund der Anfragen und der Antworten* die Domäne an, in der Y *ist mir heute etwas unverständlich* gilt. Damit ist X 'crucial for the truth conditions of the whole sentence', in der Domäne, die das Topik beschreibt (Jacobs 2001, 656). Das Topik ist meist, wie in diesem Fall, bekannt (vgl. Musan 2010, 27). Hier besetzt Topik die Vorfeld-Position des Satzes, welche nach Musan (2010) eine typische Position des Topiks ist (ebd., 35,65). Damit zeigt sich, dass die *Framing Attributes* sich nicht nur auf den lexikalischen Anschluss des Bundle beziehen, sondern auch Topik ausdrücken können. So erscheint die Klassifikation aufgrund der wenig klar formulierten Richtlinien und der dadurch fehlenden Transparenz auch in Hinblick auf die kompletten Datensets nicht konsistent und widersprüchlich.

Referential Bundles können nach den Taxonomien von Biber und Kollegen Fokus identifizieren. Die *Discourse Organizers* werden in den meisten Fällen zwischen *Topic Introduction/Focus* und *Topic Elaboration/Clarification* unterschieden. Bei Biber (2006) und Biber und Barbieri (2007) werden auch *Identification/Focus Bundles* diskutiert (Biber 2006, 159; Biber und Barbieri 2007, 271). Nun ist es linguistisch nicht ganz klar motiviert, warum Kategorien wie Topik nur bei einigen Bundles diskutiert werden. Und dennoch andere, verwandte Kategorien wie Fokus in den *Discourse Organizers* und den *Referential Bundles* die Beziehung zwischen den vorangegangenen und folgenden Diskurs beschreiben, und so eine gleiche Funktion zugewiesen bekommen (vgl. Biber, Conrad und Cortes 2004, 384, 386-387).

Diese parallele Zuteilung beziehungsweise der Ausschluss von Funktionen wird in keiner der Studien motiviert oder diskutiert. Es ist auffällig, dass der größte Teil der *Discourse Organizers* einen hohen verbalen Anteil beinhaltet (Biber 2006 158-159; Biber, Conrad

und Cortes 2004, 386). Gegenteilig verhält es sich mit den *Referential Bundles*, die zu einem großen Teil aus PP- oder NP-Fragmenten bestehen. Die Studien geben weiterhin keine Definition hinsichtlich *Topik* und *Fokus*. Beide Begriffe sind, wie bereits erwähnt, vielfältig definiert und genutzt (vgl. Musan 2010, 59-61), sodass eine Zuordnung ohne eine Diskussion der Terminologie kaum möglich erscheint.

Beide Begriffe beschreiben Phänomene, die aus der Informationsstruktur der jeweiligen Äußerungen entstehen. Beide sind insofern kontextsensitiv, dass sie in Beziehungsstrukturen, Topik-Kommentar und Fokus-Hintergrund, eingebettet sind (vgl. Altmann 1993, 5, 7; Musan 2010, 44) und werden durch verschiedene Mittel wie Prosodie oder Satzstellung realisiert (Musan 2010, 2). Zur Beziehung der beiden Begriffe untereinander:

'Obwohl die anderen Dimensionen die Realisierung eines Fokusakzents beeinflussen können, ist Fokus grundsätzlich unabhängig von den Dimensionen Bekanntheit/Belebtheit und Topik/Kommentar (Musan 2010, 58).'

Die nichtgreifbaren *Referential Bundles* geben für das nachfolgende Nomen einen interpretativen Rahmen, wie zum Beispiel das Hervorheben einer einzelnen Eigenschaft, vor (Biber 2006, 171). In vielen Studien wird als Beispiel dafür das Bundle *on the basis of* gegeben (vgl. Biber, Conrad und Cortes 2004, 288). Das Bundle *auch in Hinblick auf* (37 P) würde ebenfalls in diese Einordnung passen. Es ist nicht ausgeschlossen, dass diese Art Bundles bestimmte informationsstrukturelle Eigenschaften besitzen:

(67)

*Die Befürchtungen, es könne in diesem Bereich zu Missbrauch kommen, können wir ganz einfach beseitigen, indem wir eine entsprechende Höhe der Pauschalierung festlegen. Ziel dieser Pauschalierung ist es ja nicht, keine oder nur geringe Sozialversicherungsbeiträge abzuführen. Durch diese Pauschalierung soll vielmehr eine Vereinfachung bei der Berechnung herbeigeführt werden. [...] Wir werden deshalb in den nächsten Monaten gemeinsam noch einmal versuchen, für diese Fälle eine Pauschalierung **auch im Hinblick auf** die Sozialversicherungsbeiträge hinzubekommen. (Plenarprotokoll).*

Nach Musan (2010) sind *nur*, *schon* und *auch* fokussensitive Gradpartikeln, die 'sich aufgrund ihrer speziellen Bedeutung auf den fokussierten Ausdruck' beziehen (ebd., 43). Der fokussierte Diskursreferent ist hier nicht neu eingeführt, womit gezeigt wird, dass Fokus nicht nur auf neuen Diskursreferenten liegen kann. So kann hier die Kategorie Identification/Focus bei den *Referential Bundles* nach Biber, Conrad und Cortes (2004) nicht greifen, auch wenn davon ausgegangen wird, dass diese Bundles auch diskursorganisierende Funktionen übernehmen können (ebd., 387, 394). Demnach könnte Fokus in diesem Kontext auch mit Hilfe dieses Bundles realisiert werden, obwohl es diese Funktion nicht explizit mit der zu den Studien von Biber und Kollegen analogen Einordnung besitzt. Die Definition der *Referential Bundles* schließt dies nicht unbedingt aus, es erscheint aber unplausibel, Funktionen, die die Informationsstruktur eines Satzes realisieren, nur in einer funktionalen Kategorie, die der *Discourse Organizers*, in einem großen Umfang zu diskutieren. Weiterhin beschränkt sich Fokus nicht nur auf das Hervorheben neuer

Diskursreferenten sondern auch auf verschiedene Arten Kontrast auszudrücken (vgl. u.a. Molnár 2002, 147-148). Dies wird in den Klassifikationen nach Biber und Kollegen nicht berücksichtigt. Wie bereits erwähnt, werden die Funktionen innerhalb des Kontextes identifiziert. Somit ist nicht ausgeschlossen, dass das Bundle in einer anderen Äußerung diese Funktion nicht mehr so ausführt. Wenn alle Lexical Bundles 'building blocks of discourse' sind, dann können sie auch grundsätzlich die Informationsstruktur eines Satzes beeinflussen.

Damit erfasst die Klassifikation der *Referential Bundles* mit *Identification/Focus* nur einen möglichen Teil der Funktionen. Das Bundle *auch im Hinblick auf* ist mit 37 Vorkommen im Protokoll-Korpus kein marginales Phänomen. Dessen Einordnung gelingt nach den Taxonomien von Biber und Kollegen nur schwer. Andere, vergleichbare Bundles sind *Gerade in Zeiten in* (5 F) und *auch im Bereich der* (9 F). Diese Bundles können ebenfalls Fokus einer neu eingeführten oder bereits bekannten Entität ausdrücken. Gerade im Deutschen wird Fokus auch durch Akzentsetzung oder Wortstellung beeinflusst. Für jedes konkrete Bundle müsste eine Einzelfallanalyse durchgeführt werden, um zu entscheiden, wie und ob Fokus mit Hilfe dieses Bundles ausgedrückt wird. So müssten die Funktionen Topik und Fokus, wenn man sie annimmt, unabhängig von der Struktur bei allen Bundles im Rahmen der Diskursorganisation diskutiert werden. Was grundsätzlich nicht in den Klassifikationen nach Biber und Kollegen diskutiert ist, dass Fokus und Topik auch zusammenfallen können. Und Bundles wie *on the other hand*, die bei Biber, Conrad und Cortes (2004) als *Topic Elaboration/Clarification Bundle* interpretiert werden, könnten kontrastiv auf etwas davor genanntes referieren, Alternativen zum Nachgenannten aufzeigen und somit das Nachfolgende in den Fokus setzen - im Sinne einer Fokusprojektion (vgl. Musan 2010, 42-43, 58)³¹.

Die englischen Daten, soweit sie in den Publikationen einsehbar sind, weisen keine solche Strukturen wie ADV+PP oder PP+ADV auf, womit dieses Zuordnungsproblem dort nicht in dieser Form besteht. In jedem Fall geben die Bundles einen bestimmten, interpretativen Rahmen für das nachfolgende Element vor, dennoch sind damit ihre Funktionen mit einer Einordnung in den *Referential Bundles* nicht ausreichend beschrieben. Grundsätzlich erzeugen alle Bundles mit der Struktur *the X of the, in the X of* oder *for the x of* Modifikationen, Begrenzungen oder Mengen der eingebetteten *of*-Phrase. Dies erscheint strukturinhärent und weniger funktional im Diskurs selbst. Möglicherweise sollten hier zwei unterschiedliche funktionale Ebenen - eine eng an die Bundles gekoppelte und eine an den Kontext orientierte Ebene - angenommen werden, wobei Topik und Fokus, wie gezeigt wurde, nicht allein durch die Bundles selbst bestimmbar sind. Mit dieser Schlussfolgerung ist nicht gesagt, dass die Bundles in den verschiedenen Funktionskategorien von Topik und Fokus in den Studien von Biber und Kollegen grundsätzlich falsch zugeordnet sind. Nur die Terminologie und die Richtlinien zur Klassifikation sind in Bezug auf die deutschen Daten nicht klar und daher nicht anwendbar. Im Hinblick auf Funktionen, die die Informationsstruktur einer Äußerung näher analysieren, ist es schwierig, ohne Kontext und genaue Definitionen hinreichende Erkenntnisse zu gewinnen.

Bereits in Kapitel 3.4.2 wurde darauf hingewiesen, dass Tracy-Ventura, Cortes und Biber

³¹Für eine weitere Diskussion zu Topik und Fokus vgl. auch Molnár (2002).

(2007) eine funktionale Kategorie *Structural Only* als reine Artefakte der Datenerhebung einordnen und somit nicht weiter betrachten. Diese Artefakte hingegen weisen zum Teil ableitbare funktionale Eigenschaften auf:

Beispiel	Struktur	möglicher Kontext
<i>a mi no me</i>	<i>a mi</i> im Vokativ + NEG + PRF	<i>a mi no me interesa en ...</i>
<i>que a mi me</i>	PPERREL + <i>a mi</i> im Vokativ + PRF	<i>que a mi me interesa ...</i>

Tabelle 6.9: Fokuskonstruktion im Spanischen. Beispiel aus Tracy-Ventura, Cortes und Biber (2007, 225).

Mittels *a mi* in diesen Bundles kann, entweder im Hauptsatz *a mi no me* oder Relativnebensatz mit *que a mi me*, Fokus ausgedrückt werden: *MICH, mich interessiert [...] nicht* oder *der MICH, mich interessiert*. Hier zeigt sich, dass *Fokus* nicht in allen Kategorien als eine Möglichkeit der funktionalen Eigenschaften diskutiert worden ist. Im Gegensatz zu den diskutierten Beispielen enthalten die *Functional Only Bundles* Pronomina statt direkter NPn oder PPn.

Grundsätzlich werden Pronomen, wie die Personalpronomen *I/You* in die Auswertung eingeschlossen. Reflexiva wie im Spanischen *me* oder auch anaphorische Konstruktionen, die nicht auf einen situativen, prominenten Diskursreferenten wie *Ich/Du* referieren, sind hingegen kaum in den Beispielen enthalten. Solche Konstruktionen werden per Definition nicht vorab ausgeschlossen. Eine genauere Betrachtung dieser erscheint aber notwendig.

6.3.3 Redewendungen

In beiden Korpora lassen sich Bundles identifizieren, die nach Quasthoff, Schmidt und Hallsteinsdóttir (2010) als deutsche Phraseologismen identifizierbar sind:

- | | |
|---|---|
| (68) <i>Tür und Tor geöffnet</i> (5 F) | (72) <i>ohne Wenn und Aber</i> (20 P) |
| (69) <i>darüber im Klaren sein</i> (13 P) | (73) <i>nicht die Rede sein</i> (11 P) |
| (70) <i>auf den Punkt gebracht</i> (10 P) | (74) <i>in der Tasche haben</i> (11 P) |
| (71) <i>an einem Strang ziehen</i> (10 P) | (75) <i>nichts aber auch gar</i> (17 P) |

Damit erscheint eine theoretische Auseinandersetzung, wann eine Wortsequenz ein Lexical Bundle und wann ein Phraseologismus ist, notwendig. Der allgemeine Status als 'building block of discourse' reicht hier nicht aus, um die Phänomene hinreichend zu diskutieren. Phraseologismen scheinen eine spezifische Teilmenge der Lexical Bundles zu sein, da nicht bei allen hier erhobenen Daten diese Zuordnung möglich ist. Conrad (2004) argumentieren, dass Idiome ziemlich rar im allgemeinen Sprachgebrauch vorkommen und deren Bedeutung wenig transparent ist (ebd., 57). Nun sind die hier angegebenen Phrasen durchaus frequent, in bestimmter Hinsicht auch bedeutungstransparent - *an einem Strang ziehen* = *zusammenarbeiten* - aber dennoch nicht vergleichbar mit den anderen vorab diskutierten Bundles. Trotz der enger gefassten Richtlinien zur Identifikation von Lexical Bundles im Vergleich zu Phraseologismen wie bei Altenberg (1998) oder zu spezifischeren Datensets wie bei Quasthoff, Schmidt und Hallsteinsdóttir (2010) sind dennoch

solche Phänomene vorhanden. Hier könnte der Ansatz von Wulff (2008), mit einem graduellen Begriff von Idiomatizität mögliche Erklärungen dafür geben.

7 Schlussfolgerung und Ausblick

Aus den vorangegangenen Darstellungen wird ersichtlich, dass die Replizierung der Lexical Bundle-Studien nach der Methode von Biber und Kollegen nicht gelingt. Darum wird hier keine funktionale Klassifikation aller Bundles gegeben. Allein die Datenerhebung konnte für die deutschen Korpora übernommen werden. Eine Diskussion der so gewonnenen Daten hinsichtlich ihrer Struktur und Funktion konnte nicht in Hinblick auf die Register unterscheidenden Funktionen von Lexical Bundles durchgeführt werden. Die strukturelle Heterogenität der deutschen Sequenzen im Vergleich zu den Englischen lässt keine klare Auswertung und keinen direkten Vergleich mit den Vorgängerstudien zu. Die umfangreiche Flexionsmorphologie der deutschen Sprache ist ein Grund dafür. Diese Daten spiegeln das Type-Token-Verhältnis nicht so wieder, dass die hier vorgestellten Zahlenverhältnisse aussagekräftig sind. Weiterhin spielt die variable Wortstellung im Deutschen eine wichtige Rolle. Es hat sich gezeigt, dass hier ein neuer Ansatz zur strukturellen Klassifikation für das Deutsche diskutiert werden muss. Die Möglichkeit, die Datenerhebung auf Grundlage der Lemma-Ebene des jeweiligen Korpus zu realisieren und deren Vorteile sind diskutiert worden. In wie weit dieses Vorgehen Einfluss auf das Konzept *Lexical Bundles* hat, muss noch weiter betrachtet werden. Die Textgrundlage verlässt damit in jedem Fall die Ebene der rein quantitativen Methode, da es sich dabei um interpretierte Daten - Lemmata - handelt. Offen bleibt, ob die so errechneten Frequenzlisten noch *Lexical Bundles* enthalten oder von einem anderen Konzept, möglicherweise von *Lemma-Bundles*, ausgegangen werden muss.

Die Verteilung nach eigener Systematik der Bundles zeigt keine konkreten Unterschiede zwischen den beiden Registern *Plenarprotokoll* und *Politikforum*. Es hat sich ein ähnliches Verhältnis von Bundles, die als NP-/PP-basiert beziehungsweise als MF klassifiziert worden sind, und der Bundles, die verbale Bestandteile beinhalten, herausgestellt. Dies ist kohärent mit den ähnlichen Verteilungen der verschiedenen Wortarten in beiden Registern. Darüber hinaus können keine weiteren Schlussfolgerungen bezüglich der Charakterisierung der Register durch die Lexical Bundles gezogen werden, da die funktionale Klassifikation der Vorgängerstudien nicht klar definiert und damit wenig transparent ist. Dies liegt auch unter anderem daran, dass keine kompletten Datensets einsehbar sind und gerade die Ausschlüsse aus der Untersuchungsmenge unter anderem durch die Richtlinie für die lokalen Wiederholungen nicht ausreichend diskutiert beziehungsweise präsentiert sind.

Die Diskussion der funktionalen Klassifikation an den deutschen Daten zeigt, dass Kategorien wie die Multifunktionalität in dieser Studie weitergefasst werden müsste, als es bei den Vorgängerstudien der Fall ist. Die strukturelle Vielfalt der deutschen Sequenzen bereitet auch bei der funktionalen Zuordnung Schwierigkeiten. *Referential Bundles*, die einen anaphorischen Ausdruck beinhalten, sind in den Vorgängerstudien, soweit die

Daten präsentiert sind, nicht vorhanden. Koreferenzen innerhalb der Bundles sind nur durch *you* oder *I* realisiert, in den deutschen Daten finden sich hingegen vielfältigere Koreferenzen. Fraglich ist, ob diese gerade im Gegensatz zu lexikalischen Einheiten wie Nomen separat diskutiert werden müssen. Die Definition der **Lexical Bundles** schließt diese Bundles nicht aus.

Weiterhin ist die Definition und der Anwendungsbereich von linguistischen Begriffen wie *Topik* und *Fokus* nicht eindeutig genug für die deutschen Daten diskutiert, sodass eine Zuordnung nach diesen Kriterien nicht erfolgen konnte. Die Kategorien setzen zum Teil an unterschiedlichen Text- und Strukturebenen an, die nicht genau motiviert werden. Bestimmte Kategorien wie die *Framing Attributes* der *Referential Bundles* umfassen die phrasale Umgebung der Bundles, Funktionen der *Discourse Organizers* hingegen eine weitere, textuelle. Beide Kategorien wiederum beinhalten informationsstrukturelle Charakteristika, die in den deutschen Daten nicht immer ohne Kontext identifizierbar sind. Dies steht im Widerspruch zueinander. Eine hierarchisierte Klassifikation nach diesen Ebenen könnte diese Problematik auffangen. Der Überblick der Vorgängerstudien hinsichtlich der Klassifikationen hat ganz allgemein gezeigt, dass in vielen Studien das Konzept abgeändert werden musste, um die Daten erklären zu können. Es wird auch keine mögliche Beziehung der Lexical Bundles untereinander beziehungsweise eine Beziehung zwischen den einzelnen Funktionen betrachtet. Die hier geführte Diskussion gibt Anlass zu der Annahme, dass es möglicherweise solche Beziehung gibt. Diese müssten genauer betrachtet und der linguistische Status der Lexical Bundles daraufhin diskutiert werden. Der theoretische Status der Lexical Bundles als 'building block of discourse' umfasst diese Aspekte nicht.

Lexical Bundles werden in Abgrenzung zu Idiomen und Phraseologismen definiert. Dennoch enthalten die hier untersuchten Daten Sequenzen, die als solche eingeordnet werden können. Somit muss geklärt werden, in wie weit Lexical Bundles - oder zumindest die Definition - grundsätzlich solche Phänomene erfassen kann und ob es nur eine Frage der nachträglichen Diskussion ist, in welchem Umfang diese Sequenzen relevant für die Forschungsfragen sind. Generell scheint der Begriff *Lexical Bundle* in Verbindung mit den Klassifikationen nicht das zu umschreiben, was tatsächlich empirisch erhoben wird. Die Rohdaten, Frequenzlisten, sind zu einem kleinen Teil angesprochen und diskutiert, die überwiegende Mehrheit der rekurrenten Wortketten wird nach der Datenerhebung ausgeschlossen.

Da aufgrund der Probleme bei der Klassifizierung der Bundles keine Auswertung und kein Vergleich zwischen den Registern erfolgen konnte, bleibt auch eine Analyse von gemeinsamen Lexical Bundles noch offen. Diese Stichprobe zeigt, dass hier ganz unterschiedliche Bundles in beiden Korpora gebraucht werden:

- | | |
|------------------------------|---------------------------------|
| (76) im Zusammenhang mit der | (82) Sie in diesem Zusammenhang |
| (77) mit Blick auf die | (83) im Bezug auf die |
| (78) aber auch dass wir | (84) aber auch in den |
| (79) auch in Bezug auf | (85) auch im Hinblick auf |
| (80) auf der Grundlage des | (86) danke Ihnen für Ihre |
| (81) die Frage nach der | (87) in den kommenden Jahren |

Möglicherweise sagen Gemeinsamkeiten hier auch etwas über die Register aus. Ein weitere Aspekt, der hier nicht diskutiert werden konnte, ist die Kookkurenz von Lexical Bundles selbst, welche sich in Beispiel (39) zeigt:

- (39)
*Außerdem ist es für mich in diesem Zusammenhang und **vor allen Dingen auch vor dem Hintergrund der Debatte am heutigen Vormittag völlig unverständlich [...]** (Plenarprotokoll)*

Wenn Bundles bestimmte Funktionen im Diskurs übernehmen, so müsste auch untersucht werden, in wie fern diese Funktionen auch kombiniert Verwendung finden und welche Beziehung, zum Beispiel welche hierarchische Beziehung, zwischen diesen besteht. Oder ob von einer einzelnen N-Wortketten ausgegangen werden muss. So stellt sich hier wieder die Frage nach dem nötigen oder nur möglichen Umfang eines Lexical Bundles.

Dieser Arbeit zeigt, dass das Konzept und die Methode, die hier repliziert werden sollten, nicht genügend erklärt und motiviert sind und dadurch keine Transparenz für eine Übernahme besitzen. Weiterhin wurde auf einige Besonderheiten der deutschen Sprache hingewiesen, die eine Analyse zusätzlich erschweren. Durch die hier geführte Diskussion wird ebenfalls ersichtlich, dass das Konzept der Lexical Bundles nicht ausreichend umrissen ist, und dass eine Abgrenzung zu anderen frequenzbasierten Phänomenen nicht ohne weiteres möglich ist. Die Frage, ob Sprecher eher auf holistische Phrasen zurückgreifen, kann hier wie auch in den Vorgängerstudien nicht beantwortet werden. Bezüglich des linguistischen Status der Lexical Bundles könnten mit Hilfe der Konstruktionsgrammatik (u.a. Goldberg 1995 und 2006) Ansätze für eine theoretische Einordnung und Erklärung des Phänomens gegeben werden. Die bereits diskutierte Rahmenstruktur der Bundles und die Annahme, dass auf nicht analysierte Chunks innerhalb der Sprachproduktion zurückgegriffen wird (vgl. Biber, Conrad und Cortes 2004, 399-400), lassen sich innerhalb dieser Grammatiktheorie diskutieren. Die Beziehung zwischen Form und Bedeutung sprachlicher Äußerungen könnten auf die von Biber und Kollegen entwickelte Taxonomie angewandt werden. Die Bundles geben einen bestimmten strukturellen Rahmen vor, der bestimmte lexikalische Füllung erlaubt. Die lexikalischen Füllungen könnten enge funktionale Eigenschaften ähnlich denen der *Stance Bundles* oder der *Referential Bundles* besitzen. Im Kontext erfüllen sie dann möglicherweise weitere Diskurs organisierende Funktionen, die auch informationsstrukturelle Charakteristika wie Topik und Fokus einschließen. Auf diese Weise könnten die Zuordnungsprobleme gerade in Hinblick auf Topik und Fokus besser ausgewiesen und von anderen Funktionen abgegrenzt werden.

So könnten Lexical Bundles als eine bestimmte, hochfrequente Konstruktionsklasse mit Vererbungsbeziehungen eingeordnet werden. Eine genauere Diskussion und Überprüfung dieser Überlegungen ist dazu notwendig. Darüber hinaus wäre es interessant, ob mithilfe von Produktivitätsmodellen auch die kontextspezifische Verwendung der Bundles näher untersucht werden kann (vgl. Zeldes 2011). Gerade die lexikalischen Lücken vieler Sequenzen bieten hier eine Grundlage für diesen Ansatz.

Abschließend kann nicht geklärt werden, ob Lexical Bundles auch im Deutschen durch ihre strukturellen und funktionalen Eigenschaften Register näher beschreiben und unterscheiden können. Grundsätzlich erscheint es möglich, wenn die Definition und die Klassifikationen ausführlich motiviert und angepasst werden. In jedem Fall muss ein solches Konzept auf andere, in diesem Fall auch verwandte, Sprache anwendbar sein. In einem größeren theoretischen Rahmen ist es dennoch fraglich, ob eine frequenzbasierte Methode eine so enge strukturelle wie funktionale Interpretation zulassen kann.

8 Quellen- und Literaturverzeichnis

8.1 Quellen

Originaldaten Korpus Plenarprotokolle

<http://www.bundestag.de/dokumente/protokolle/plenarprotokolle/index.html>.

Zuletzt angeschaut am 22.09.2011.

Originaldaten Korpus Politikforum

<http://www.abgeordnetenwatch.de/>.

Zuletzt angeschaut am 22.09.2011.

Mediathek des Deutschen Bundestags

<http://www.bundestag.de/Mediathek/>.

Zuletzt angeschaut am 06.09.2011.

Such- und Visualisierungstool ANNIS2

<http://www.sfb632.uni-potsdam.de/d1/annis/>.

Zuletzt angeschaut am 22.09.2011.

8.2 Software

TreeTagger

<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>.

Zuletzt angeschaut am 13.09.2011.

Perl

<http://strawberryperl.com/>

Zuletzt angeschaut am 25.09.2011.

8.3 Literatur

Altenberg, Bengt (1998). On the Phraseology of Spoken English. The Evidence of Recurrent Word-Combinations. In Cowie, Anthony P. (Hg.). *Phraseology. Theory, Analysis, and Application*. Clarendon Press, Oxford. 101-122.

- Altmann, Hans (1993). Fokus-Hintergrund-Gliederung und Satzmodus. In Reis, Marga (Hg.). *Wortstellung und Informationsstruktur*. Max Niemeyer, Tübingen. 1-37.
- Barbour, Stephen und Stevenson, Patrick (1998). *Variation im Deutschen. Soziolinguistische Perspektiven*. Übersetzt aus dem Englischen von Konstanze Gebel. De Gruyter, Berlin [u.a.].
- Baroni, Marco (2009). Distributions in text. In Lüdeling, Anke und Kytö, Merja (Hg.). *Corpus Linguistics. An International Handbook*. Band 2. Reihe Handbücher zur Sprach- und Kommunikationswissenschaft. Mouton de Gruyter, Berlin. 803-822.
- Biber, Douglas (2006). *University Language*. John Benjamins, Amsterdam [u.a.].
- Biber, Douglas (2004). Lexical bundles in academic speech and writing. In Lewandowska-Tomaszczyk, Barbara (Hg.). *Practical applications in language and computers*. PALC 2003 Peter Lang, Frankfurt am Main. 165-177.
- Biber, Douglas (1994). An Analytical Framework for Register Studies. In Biber, Douglas und Finegan, Edward (Hg.). *Sociolinguistic Perspectives on Register*. Oxford University Press, Oxford [u.a.]. 29-56.
- Biber, Douglas und Barbieri, Federica (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263-286.
- Biber, Douglas und Conrad, Susan (2006). *Register, Genre, and Style*. Cambridge University Press, Cambridge [u.a.].
- Biber, Douglas und Conrad, Susan (1999). Lexical Bundles in Conversation and Academic Prose. In Hasselgard, Hilde und Oksefjell, Signe (Hg.). *Out of Corpora. Studies in Honor of Stig Johansson*. Radopi, Amsterdam [u.a.]. 181-190.
- Biber, Douglas/ Conrad, Susan und Cortes, Viviana (2004). If you look at ... Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25/3: 371-405.
- Biber, Douglas/ Conrad, Susan und Cortes, Viviana (2003). Lexical bundles in speech and writing. An initial taxonomy. In Wilson, Andrew et al. (Hg.). *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech*. Peter Lang, Frankfurt am Main. 71-92.
- Biber, Douglas/ Conrad, Susan/ Reppen, Randi (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge University Press, Cambridge.
- Biber, Douglas/ Connor, Ulla und Upton, Thomas A. (2007). *Discourse on the move. Using corpus analysis to describe discourse structure*. John Benjamins, Amsterdam [u.a.].

- Biber, Douglas/ Johansson, Stig/ Leech Geoffrey/ Conrad, Susan und Finegan, Edward (1999). *The Longman Student Grammar of Spoken and Written English*. Longman, Harlow.
- Borsch, Robert (2008). *Politiker und ihr Umgang mit Anglizismen*. Kovac, Hamburg.
- Conrad, Susan und Biber, Douglas (2004). The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica* 20: 56-71.
- Cortes, Viviana (2004). Lexical Bundles in published and student disciplinary writing. Examples from history and biology. *English for Specific Purposes* 23: 397-423.
- Deutscher Bundestag (Hg.). *Geschäftsordnung*. In der Fassung der Bekanntmachung vom 2. Juli 1980 (BGBl. I S.1237), zuletzt geändert laut Bekanntmachung vom 17. Dezember 2010 (BGBl. I S. 2199) http://www.bundestag.de/dokumente/rechtsgrundlagen/go_btg/go06.html.
Zuletzt angeschaut am 22.09.2011.
- Evert, Stefan (2009). Corpora and collocations. In Lüdeling, Anke und Kytö, Merja (Hg.). *Corpus Linguistics. An International Handbook*. Band 2. Reihe Handbücher zur Sprach- und Kommunikationswissenschaft. Mouton de Gruyter, Berlin. 1212-1248.
- Firth, John R. (1957). *Papers in linguistics*. Oxford University Press, London.
- Friesdorf, Marita (2011). *Bundestagsreden. Analyse und Bewertung der redaktionellen Bearbeitung*. Verlag für Gesprächsforschung, Mannheim. http://opus.kobv.de/tuberlin/volltexte/2011/3149/pdf/friesdorf_marita.pdf.
Zuletzt angeschaut am 22.09.2011.
- Goldberg, Adele (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. The University of Chicago Press, Chicago [u.a.].
- Goldberg, Adele (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford University Press, Oxford [u.a.].
- Gries, Stefan Th. (2008). *Quantitative Corpus Linguistics with R*. Routledge, New York [u.a.].
- Hyland, Kan (2008). As can be seen. Lexical Bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.
- Jacobs, Joachim (2001). The dimensions of topic-comment. *Linguistics* 39: 641-681.

Jalali, Hassan und Ghayoomi, Sumayeh (2010). *A Comparative Qualitative Study of Lexical Bundles in Three Academic Genres of Applied Linguistics*. MJAL 2:4. <http://www.mjal.org/Journal/A%20Comparative%20Qualitative%20Study%20of%20Lexical%20Bundles%20in%20Three%20Academic%20Genres%20of%20Applied%20Linguistics4.pdf>.
Zuletzt angeschaut am 20.09.2011.

Lemnitzer, Lothar und Zinsmeister, Heike (2006). *Korpuslinguistik*. Narr, Tübingen.

Molnár, Valéria (2002). Contrast. From a contrastive perspective. In Hasselgard, Hilde et al. (Hg.). *Information Structure in a Cross-Linguistic Perspective*. Rodopi, Amsterdam [u.a.]. 147-161.

Nerius, Dieter (2005). Spoken and Written Language. Gesprochene und geschriebene Sprache. In Ammon, Ulrich et al. (Hg.). *Sociolinguistics. An International Handbook of the Science of Language and Society*. Band 2.2. 2. Ausgabe. Mouton de Gruyter, Berlin. 1628-1638.

Musan, Renate (2010). *Informationsstruktur*. Universitätsverlag Winter, Heidelberg.

Pang, Winnie (2010). *Lexical Bundles and the Construction of an Academic Voice. A Pedagogical Perspective*. Asian EFL. Journal. Professional Teaching Articles. Band 47. http://www.asian-efl-journal.com/PTA/October_2010-WPang.pdf.
Zuletzt angeschaut am 20.09.2011.

Partington, Alan (1998). *Pattern and Meanings. Using Corpora for English Language Research and Teaching*. John Benjamins, Amsterdam [u.a.].

Partington, Alan und Morley, John (2004). At the heart of ideology. Word and cluster/bundle frequency in political debate. In Lewandowska-Tomaszczyk, Barbara (Hg.). *Practical applications in language and computers*. PALC 2003. Peter Lang, Frankfurt am Main. 179-192.

Quasthoff, Uwe/ Schmidt, Fabian und Hallsteinsdóttir, Erla (2010). Häufigkeit und Struktur von Phraseologismen am Beispiel verschiedener Web-Korpora. In Ptashnyk, Stefaniya (Hg.). *Korpora, Web und Datenbanken*. Schneider Verlag Hohengehren, Baltmannsweiler. 37-53.

Strunkyte, Gintarė und Jurkunaite, Eglė (2008). *Written Academic Discourse. Lexical Bundles in Humanities and Natural Sciences*. Unpublished BA dissertation, Vilnius University. http://www.lexically.net/wordsmith/corpus_linguistics_links/Lexical%20Bundles%20in%20Humanities%20and%20Natural%20Sciences.pdf.
Zuletzt angeschaut am 22.09.2011.

Teliya, Veronika/ Bragina, Natalya/ Oparina, Elena und Sandomirskaya, Irina (1998). Phraseology as a Language of Culture. Its Role in the Representation of a Collective Mentality. In Cowie, Anthony P. (Hg.). *Phraseology. Theory, analysis, and application*. Clarendon Press, Oxford. 55-75.

Tognini-Bonelli, Elena (2007). The corpus-driven approach. In Teubert, Wolfgang und Krishnamurthy, Ramesh (Hg.). *Corpus Linguistics. Critical Concepts in Linguistics*. Band 1. Routledge, London [u.a.]. 74-92.

Tracy-Ventura, Nicole/ Cortes, Viviana und Biber, Douglas (2007). Lexical bundles in speech and writing. In Parodi, Giovanni (Hg.). *Working with Spanish Corpora*. Continuum, London [u.a.]. 217-231.

Wulff, Stefanie (2008). *Rethinking Idiomaticity. A Usage-based Approach*. Continuum, London [u.a.].

Zeldes, Amir (2011), On the Productivity and Variability of the Slots in German Comparative Correlative Constructions. In Konopka, Marek et al. (Hg.). *Grammar and Corpora. Grammatik und Korpora 2009. Third International Conference. Mannheim, 22.-24.09.2009*. Tübingen, Narr. 429-449. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/pdf/je_desto_vorversion.pdf.

Zuletzt angeschaut am 22.09.2011.

A Funktionale Klassifikation nach Biber, Conrad und Cortes (2004)

Kategorie	Subkategorie	Subsubkategorie	Beispiel
<i>Referential Expressions</i>	identification/ focus		<i>is one of the</i>
			<i>or something like that</i>
	imprecision specification attributes	quantify specification	<i>have a lot of</i>
		tangible framing attributes	<i>the size of the</i>
		intangible framing attributes	<i>the nature of the</i>
	time/place/text reference	place reference	<i>in the United States</i>
	time reference	<i>at the time of</i>	
	text deixis	<i>shown in figure N</i>	
	multi-functional reference	<i>the end of the</i>	
<i>Discourse Organizers</i>	topic introduction/ focus		<i>if you look at</i>
	topic elaboration/ clarification		<i>has to do with</i>
<i>Stance Expressions</i>	epistemic stance		pers. <i>I don't know if</i>
			impers. <i>the fact that the</i>
	attitudinal/ modality stance	desire	pers. <i>if you want to</i>
		obligation/ directive	pers. <i>I want you to</i>
		intention	impers. <i>it is important to</i>
		prediction	pers. <i>I'm not going to</i>
		ability	impers. <i>is going to be</i>
		pers. <i>to be able to</i>	
		impers. <i>can be used to</i>	
<i>Special Conversation Function</i>	politeness		<i>thank you very much</i>
	simple inquiry		<i>what are you doing</i>
	reporting		<i>I said to her</i>

Tabelle A.1: Übersicht der funktionalen Klassifikation nach Biber, Conrad und Cortes (2004, 384-388).

Danksagung

Für die Unterstützung bei der technischen Umsetzung dieser Arbeit und die sehr hilfreiche Expertise möchte ich mich ganz besonders beim Sonderforschungsbereich 632: 'Informationsstruktur' B4 und bei Thomas Krause, Julia Richling, Amir Zeldes und Florian Zipser bedanken.