4. Kollokationen, n-Gramme, Mehrworteinheiten

Bubenhofer, Noah

DOI: https://doi.org/10.1515/9783110296310-004

Noah Bubenhofer

4. Kollokationen, n-Gramme, Mehrworteinheiten

Abstract: Der Beitrag beschreibt, wie die Berechnung von Kollokationen und Mehrworteinheiten in Textkorpora typische Sprachgebrauchsmuster freilegen kann. Zuerst werden in einer Gegenstandsbestimmung die verschiedenen Termini zur Bezeichnung von mehrgliedrigen Ausdrücken auf der sprachlichen Oberfläche diskutiert. Ein kurzer Forschungsüberblick nennt korpuslinguistische Arbeiten mit politolinguistischen Zielen, die Gebrauch von der Analyse solcher Mehrworteinheiten machen. Anschließend zeigt eine exemplarische Analyse von unterschiedlichen Typen von Mehrworteinheiten in parteispezifischen Teilkorpora der Wortprotokolle des Deutschen Bundestags die Einsatzmöglichkeiten.

1 Gegenstandsbestimmung

Die Terminologie zur Benennung von Einheiten, die aus mehreren Wörtern bestehen, ist unübersichtlich. Mehrworteinheiten, Kollokationen, Phraseologismen, usuelle Wortverbindungen oder n-Gramme sind die wichtigsten Termini, wobei die Differenzierung unterschiedlich gehandhabt wird.

Die Unterschiede rühren von den Provenienzen der Termini: Einerseits wurde mit dem britischen Kontextualismus der Begriff der *Kollokation* geprägt. Nach Firth (Firth 1957, 194) sind Kollokationen häufig auftretende Wortverbindungen, die für einen bestimmten Ausschnitt von Sprache typisch sind. Während bei Firth noch nicht genauer spezifiziert ist, was *häufig* und *typisch* bedeutet, wurde dieses Konzept in der Folge genauer spezifiziert und widerspiegelt heute die in der Korpuslinguistik dominierende Auffassung der „empirischen Kollokation" (Evert 2009, 1213). Demnach sind Kollokationen Paare von Worteinheiten (auf der Basis von Wortformen, Lemmata oder anderen sprachlichen Einheiten), die innerhalb einer bestimmten Distanz zueinander kookkurrieren und eine statistisch feststellbare Bindung zueinander aufweisen. Typischerweise wird diese Bindung, *Assoziation*, als statistische Signifikanz ausgedrückt, nach der die beiden Einheiten in einem Korpus häufiger miteinander vorkommen, als es bei einer zufälligen Verteilung im Korpus erwartbar wäre.

DOI 10.1515/9783110296310-004

Neben dieser empirischen Auffassung von Kollokationen gibt es eine theoretische Auffassung, die Kollokationen enger fasst: Danach handelt es sich um lexikalische Einheiten, die z. B. in einer syntaktischen Beziehung zueinander stehen (etwa Adjektiv und Nomen in einer Nominalphrase) oder deren gemeinsame Bedeutung nicht dem Kompositionalitätsprinzip folgt, also opak ist (z. B. blinder Passagier). Besonders die Phraseologie lenkte die Aufmerksamkeit auf solche Konstruktionen und deswegen herrscht dort auch heute meist eine engere Auffassung von „Kollokation" vor (Bartsch 2004; Evert 2009).

Evert (2009, 1213) verwendet den Begriff „Kollokation" für die empirische und „Mehrworteinheit" („Multiword Expression") für die theoretische Auffassung von Kollokationen. Die Begriffsverwendung ist in der Literatur jedoch nicht einheitlich.

Prinzipiell sehen beide Konzepte Ausdrücke vor, die aus mehr als zwei Einheiten bestehen können. Die (empirischen) Kollokationen werden jedoch normalerweise als Paare aufgefasst, was auch daran liegt, dass es kaum Ansätze gibt, die die statistische Assoziation zwischen mehr als zwei Elementen messen können (Evert 2009, 1244). Um Kollokationen der Länge *n* zu bezeichnen, bietet sich darüber hinaus der Begriff *n-Gramm* (Bigramm, Trigramm etc.) an. Unter Mehrworteinheiten – oder auch „usuellen Wortverbindungen" (Steyer 2013) – werden häufiger Einheiten verstanden, die aus mehr als zwei Wörtern bestehen können, da sich hier oft das Problem nicht stellt, mit statistischen Assoziationsmaßen arbeiten zu müssen.

Ebenso offen ist die Frage, welches die Bestandteile von Kollokationen und Mehrworteinheiten sind. Neben den Wortformen können auch die Grundformen oder andere Kategorien wie Wortart- oder semantische Klassen verwendet werden, wie weiter unten gezeigt wird.

Im Folgenden liegt der Fokus auf empirischen Kollokationen oder n-Grammen. Bei quantitativen Korpusanalysen ist es unabdingbar, eine algorithmisierbare Operationalisierung von mehrgliedrigen lexikalischen Ausdrücken nutzen zu können, um die Daten maschinell verarbeiten zu können. Im Nachgang können die berechneten Kollokationen manuell oder halbautomatisch kategorisiert werden, um daraus Mehrworteinheiten zu extrahieren, deren Assoziation nicht nur statistisch definiert ist.

2 Forschungsüberblick

Arbeiten, die Kollokationen und Mehrworteinheiten in den Fokus politolinguistischer Analysen nehmen, finden sich über verschiedene Teildisziplinen verstreut. Ein wichtiger theoretischer Hintergrund solcher Arbeiten ist die korpuslinguistische Diskursanalyse in den verschiedenen Ausprägungen. Eine Einführung in die modernere Diskurslinguistik, wie diejenige von Spitzmüller und Warnke (2011), zeigt die Relevanz korpuslinguistischer Analysemethoden. Diskurslinguistische Arbeiten verfolgen meistens auch politolinguistische Fragestellungen. Die Vielfalt solcher Ansätze ist

in einer Reihe von Sammelbänden ersichtlich (z. B. Bluhm u. a. 2000; Busse/Teubert 2013; Felder u. a. 2011; Jung 2001) und soll an dieser Stelle nicht erschöpfend dargestellt werden.

Die in der Politolinguistik erarbeiteten Analysekategorien (etwa bei Burkhardt 2003; Girnth/Spieß 2006; Girnth 2002; Schröter/Carius 2009) wurden teilweise korpuslinguistisch operationalisiert. Hauptsächlich die angelsächsische Korpuslinguistik hat bereits früh auf die soziokulturelle Dimension von Korpusanalysen aufmerksam gemacht und vor allem Kollokationen als wichtige Analysedimension etabliert (etwa bei Sinclair 1991; Olsen/Harvey 1988; Teubert 2006; Tognini-Bonelli 2001) und dafür plädiert, die sprachliche Oberfläche und statistische Analyseergebnisse ernst zu nehmen. Die frankophone Tradition der Lexikometrie verfolgt dabei ähnliche Ziele und zog schon früh auch komplexere multivariate Analysemethoden hinzu (Dzudzek u. a. 2009; Glasze 2007; Lebart/Salem 1994; Mattissek 2005; Scholz 2010, 145 ff.).

Die folgenden Arbeiten verfolgen im engeren Sinn politolinguistische Interessen und arbeiten mit statistischen Korpusanalysen: Im Kontext der Forschergruppe *semtracks* (www.semtracks.org) sind eine Reihe von Analysen entstanden, die vornehmlich mit komplexen n-Grammen und Kollokationsanalysen arbeiten, um Sprachgebrauchsmuster datengeleitet zu berechnen. Dies, um die partei- oder personenspezifischen Sprachcharakteristika (Bubenhofer u. a. 2009; Ebling 2010), extremistische Positionen (Ebling u. a. 2014) oder Phänomene der Skandalisierung (Bubenhofer 2013) zu beschreiben. Einen weiteren, stärker diskurslinguistischen Fokus, nehmen die diachronen Analysen von Veränderungen in Diskursen in der Wochenzeitung Die Zeit und dem Magazin Spiegel ein (Bubenhofer u. a. 2014; Scharloth u. a. 2013). Eine Einschränkung der Kollokationen auf Verbindungen mit Toponymen, an denen sich (auch) politische Diskurse über Orte und Regionen ablesen lassen, bietet das Konzept der Geokollokationen (Bubenhofer 2014).

In der lexikometrischen Tradition arbeitet Scholz (2010) mit der Berechnung von Kollokationen und Mehrworteinheiten („wiederholte Segmente"), um Diskurse um die Europäische Union zu untersuchen. Eher argumentationstheoretisch, jedoch korpuslinguistisch unterstützt, untersuchen Ziem u. a. (2013) Krisendiskurse. Vogel (2010) schlägt einen korpuslinguistischen Ansatz für Imageanalysen vor und entwickelte auch eine entsprechende Software (Vogel 2012). Ebenso klar korpuslinguistisch gehen Storjohann und Schröter (2011) oder Koller und Farrelly (2010) vor, um den Diskurs der Finanz- und Wirtschaftskrise zu untersuchen (mit einem Fokus auf Metaphern bei Koller 2006 in Unternehmensdiskursen). Viele weitere Arbeiten wären zu zitieren, die nicht im engeren Sinn politolinguistisch, sondern eher diskurslinguistisch vorgehen, wobei eine Abgrenzung naturgemäß schwierig ist.

Neben der linguistischen Tradition gibt es auch von politologischer Seite Ansätze, mit statistischen Verfahren an die Vermessung der sprachlichen Oberfläche zu gehen. Am bekanntesten ist wohl das „Wordscore-Verfahren" (Laver u. a. 2003), bei dem anhand eines manuell z. B. nach politischen Positionen kategorisierten Trainingskorpus die typische Distribution von Lexemen dieser Positionen statistisch modelliert

werden, um neue Texte maschinell kategorisieren zu können (vgl. z. B. Blätte 2012). Ähnlich gehen andere statistische Textklassifikationsalgorithmen vor, die hauptsächlich für Data-Mining-Aufgaben eingesetzt werden; die verschiedenen Topic-Modelling-Verfahren (Anthes 2010) – z. B. LDA, „Latent Dirichlet Allocation" (Blei u. a. 2003) – gehören zu den Algorithmen, die in den Geistes- und Sozialwissenschaften hin und wieder eingesetzt werden – vgl. die Website www.programminghistorians.org für entsprechende Beispiele und Anleitungen und Rohrdantz u. a. (2012) als Ansatz, visuell datengeleitet den Wandel der Semantik von Lexemen zu analysieren.

3 Exemplarische Analyse

Für die exemplarische Analyse werden im Folgenden Protokolle des Deutschen Bundestags verwendet. Leitend für die Analyse soll die Frage sein, ob sich für die Parteien typische Muster des Sprachgebrauchs (Bubenhofer 2009) finden lassen. Die Hypothese hinter dieser Fragestellung zielt auf die Bedeutung von sprachlichen Mustern als Indikatoren für unterschiedliche, diskursiv geprägte Sprechweisen:

> Hypothese: Die für eine Partei im Vergleich zu allen anderen Parteien typischen Mehrworteinheiten sind Indikatoren für unterschiedliche Funktionen (Regierung vs. Opposition), Themensetzungen und Sprechweisen der Parteien zu diesen Themen.

An dieser Stelle kann keine vollständige Prüfung dieser Hypothese geleistet werden. Stattdessen sollen unterschiedliche methodische Zugänge zu den Daten diskutiert und deren Potenzial für politolinguistische Analysen aufgezeigt werden.

Die Untersuchungskorpora speisen sich aus dem PolMine-Plenardebattenkorpus (PDK), das sämtliche Plenardebatten auf Bundes- und Landesebene seit 2000 korpuslinguistisch aufbereitet umfasst (Blätte 2013). Die Protokolle liegen in einem XML-Format vor, bei dem reichhaltige Metadaten zu den Sitzungen und Sprecher/innen erfasst sind. Daraus wurden Teilkorpora erstellt, die alle Äußerungen der Parlamentarier/innen (ohne Äußerungen des Parlamentspräsidiums) der Wahlperiode 17 (2009–2013), nach Parteien gegliedert, umfassen. Die Auswahl beschränkte sich allerdings auf die Parteien Bündnis 90/Die Grünen, CDU/CSU, Die Linke, FDP und SPD. Die Wortformen (Tokens) wurden mit Hilfe des TreeTaggers (Schmid 1995) maschinell nach Wortarten klassifiziert und lemmatisiert unter Verwendung der Standardbibliothek fürs Deutsche, die nach dem Stuttgart-Tübingen-Tagset (STTS) klassifiziert ist (Schiller u. a. 1995). Tabelle 1 gibt einen Überblick über die Größe der Teilkorpora.

Tab. 1: Datengrundlage der exemplarischen Analyse

Korpus	Tokens	Sätze
B90/Die Grünen	1.843.824	116.563
CDU/CSU	6.276.625	390.341
Die Linke	2.343.904	155.532
FDP	2.939.585	188.728
SPD	4.042.815	258.485
Total	17.446.753	1.109.649

3.1 Methode

3.1.1 Berechnung der Mehrworteinheiten

Die Berechnung einer Frequenzliste von Mehrworteinheiten in einem Korpus ist unkompliziert: Zunächst werden im Korpus

Wort-1 Wort-2 Wort-3 Wort-4 Wort-5 Wort-6 Wort-7 ... Wort-n

alle kombinatorisch möglichen n-Gramme der Länge n (hier 3) aufgeführt:

Wort-1 Wort-2 Wort-3
Wort-2 Wort-3 Wort-4
Wort-3 Wort-4 Wort-5
...

Danach wird ausgezählt, wie oft jedes n-Gramm insgesamt im Korpus vorkommt.

Je nach Definition der Mehrworteinheit verändert sich die Erstellung aller kombinatorisch möglichen Mehrworteinheiten:

- Bestandteile der Mehrworteinheit: Mögliche Bestandteile sind die Wortformen, die Grundformen, Wortartklassen, andere linguistische Einheiten oder Annotationen (z. B. semantische oder syntaktische Klassen) oder eine Kombination davon.
- Kookkurrenz: Die Mehrworteinheit kann als Kette unmittelbar aufeinanderfolgender Worteinheiten oder aber als in einem weiteren Kontext kookkurrierende Worteinheiten verstanden werden. Der weitere Kontext kann textoberflächlich über eine in Anzahl Wörtern gemessene Distanz angegeben werden oder aber sich an anderen Einheiten im Text orientieren (z. B. Kookkurrenz im gleichen Satz). Weiter sind aber auch syntaktische Restriktionen denkbar (z. B. nur Mehrworteinheiten die aus Nominalphrasen bestehen). Vgl. für weitere Ausführungen dazu Evert (2009, 1221 ff.).

– Assoziation: Weiterhin kann die Bindungsstärke der Mehrworteinheit mittels eines Signifikanzmaßes berechnet werden. Während dies für binäre Kollokationen üblich ist, wird das bei Mehrworteinheiten, die aus mehr als zwei Einheiten bestehen, selten gemacht. Anpassungen für solche Mehrworteinheiten finden sich bei Zinsmeister/Heid (2003) und da Silva/Lopez (1999). Zudem gibt es unterschiedliche Implementierungen als Software-Tools, so z. B. das „Ngram Statistics Package" (Banerjee/Pedersen 2003), bei dem ebenfalls für n-Gramme angepasste Assoziationsmaße verfügbar sind.

Im Folgenden werden diese Typen von Mehrworteinheiten berechnet:
– n-Gramme mit den Längen n von 3, 4 und 5 auf der Basis von Grundformen (Lemmata), direkt aufeinander folgend,
– komplexe n-Gramme mit den Längen n von 3 bis 8 auf der Basis von Kombinationen von Wortformen und Wortartklassen, direkt aufeinander folgend; Wortformen mit den Wortartklassen Nomen, Eigenname, Artikel, Adjektiv, Adverb, Modalverben und Zahlausdrücke werden dabei durch Wortartklassen statt Wortformen ausgedrückt (mehr dazu weiter unten).

3.1.2 Berechnung der *Keyness*

Ziel der Analyse soll die Berechnung der Mehrworteinheiten sein, die typisch für eine Partei im Vergleich zu allen anderen Parteien sind. Deswegen reicht es nicht aus, pro Partei die häufigsten Mehrworteinheiten zu berechnen, zusätzlich werden aus diesen Listen der Mehrworteinheiten mit den jeweiligen Frequenzen in jeder Partei diejenigen ausgewählt, die bei einer Partei häufiger vorkommen als in den anderen Parteien.

Die übliche Methode, um für ein Korpus im Vergleich zu einem Referenzkorpus typische Einheiten (Lexeme, Mehrworteinheiten etc.) zu berechnen, ist der „Keyness"-Ansatz (Scott/Tribble 2006; Bondi/Scott 2010), bei dem die Keyness, also die Typizität jedes n-Gramms im Untersuchungskorpus im Vergleich zum Referenzkorpus berechnet wird. Mit Keyness ist also ein Assoziationsmaß gemeint, mit dem ausgedrückt wird, ob ein bestimmtes Wort signifikant häufiger im Untersuchungskorpus vorkommt als im Referenzkorpus. Dieses Maß ist sehr verbreitet, um Schlüsselwörter (Keywords) in einem Korpus zu finden und in viele Korpustools implementiert. Natürlich kann dieses Verfahren auch eingesetzt werden, um die Keyness von Mehrworteinheiten zu berechnen, wie Bubenhofer (2009) gezeigt hat.

Im Beispiel der vorliegenden Analyse stellt jeweils ein Parteienkorpus A das Untersuchungs- und das Gesamtkorpus Z (inkl. A) das Referenzkorpus dar. Für jede in Korpus A vorkommende Mehrworteinheit werden die Frequenzen in Korpus A und im Referenzkorpus Z berechnet und in einer Kontingenztabelle diese *beobachteten Werte* aufgeführt:

Tab. 2: beobachtete Werte

beobachtete Werte	Korpus A (SPD)	Korpus Z (alle)	Summen
MWE („die Frage sein“)	235	640	875
alle anderen Wörter	4914087	22203651	27117738
Summen	4914322	22204291	27118613

In der Kontingenztabelle werden nicht nur die Frequenzen der Mehrworteinheit in den beiden Korpora aufgeführt, sondern auch die Anzahl aller anderen Mehrworteinheiten im Korpus und damit die Anzahl aller kombinatorisch möglichen Mehrworteinheiten in den Korpora und insgesamt (Summen).

Daraus können nun die erwarteten Werte abgeleitet werden. Die erwarteten Werte orientieren sich an den Korpusgrößen: Im Beispiel oben würde man erwarten, dass die insgesamt in beiden Korpora vorkommenden 875 Mehrworteinheiten im Verhältnis der Korpusgrößen A und Z verteilt sind. Die Tabelle der erwarteten Werte sieht demnach folgendermaßen aus:

Tab. 3: erwartete Werte

erwartete Werte	Korpus A (SPD)	Korpus Z (alle)	Summen
MWE („die Frage sein“)	159 =4914322* 875/27118613	716	875
alle anderen Wörter	4914163	22203575	27117738
Summen	4914322	22204291	27118613

Im nächsten Schritt wird nun gemessen, ob die Differenz zwischen den beobachteten und erwarteten Werten genug groß ist, um als signifikant, also nicht zufällig, eingeschätzt werden zu können. Dazu eignen sich sog. Signifikanztests, wobei der einfachste Test der „Chi-Quadrat-Test“ ist (Kilgarriff 2001):

$$X^2 = \Sigma \frac{(O - E)^2}{E}$$

O steht für die beobachteten (*observed Values*), *E* für die erwarteten Werte (*expected Values*). Für jede Zelle in der Kontingenztabelle wird der erwartete vom beobachteten Wert abgezogen und quadriert und dieses Ergebnis durch den erwarteten Wert dividiert. Die für die vier Zellen so berechneten Werte werden summiert und ergeben X^2 (Chi Quadrat) – im Beispiel 45. Wenn $X^2 \geq 3.84$ beträgt, dann ist der Frequenzunterschied zwischen den beiden Korpora *signifikant*, bei $X^2 \geq 6.64$ *hoch* und bei $X^2 \geq$

10.83 *höchst signifikant* (vgl. dazu eine Tabelle der kritischen Werte der Chi-Quadrat-Verteilung).

Anstelle eines Chi-Quadrat-Signifikanztests kann auch ein „Log-Likelihood-Maß" verwendet werden, das bei kleinen erwarteten Frequenzen geeigneter ist (Kilgarriff 2001, 121; Rayson/Garside 2000).

Wenn für jede Mehrworteinheit in jedem Teilkorpus das Signifikanzmaß (Chi Quadrat, Log-Likelihood oder ein anderes Maß) berechnet worden ist, können pro Korpus nach Signifikanzmaß absteigend geordnete Ranglisten der Mehrworteinheiten erstellt werden. Je weiter oben in der Rangliste die Mehrworteinheit steht, desto spezifischer ist sie für das jeweilige Korpus.

Einige Korpusanalyseprogramme erlauben die bequeme Berechnung von Keywords, also einzelnen Lexemen, meist jedoch nicht die Berechnung der Keyness von Mehrworteinheiten.

3.1.3 Clustering der Mehrworteinheiten nach Ähnlichkeit

Bei der Berechnung der Mehrworteinheiten ergeben sich umfangreiche Listen: Im hier verwendeten Korpus sind je nach Partei 2.000 bis 11.000 unterschiedliche Mehrworteinheiten der Längen 3–5 signifikant ($p \leq 0{,}05$) für das jeweilige Korpus. Durch die verschiedenen Längen gibt es darüber hinaus eine Reihe von Mehrworteinheiten, die sich gegenseitig enthalten, wie z. B.:

sein zuversichtlich , dass
zuversichtlich , dass
zuversichtlich , dass wir

Zudem gibt es Mehrworteinheiten, die sehr ähnlich sind. Mit einem Clustering-Verfahren sollen die Listen so gruppiert werden, dass ähnliche Mehrworteinheiten aufeinander folgen. Dafür eignen sich Verfahren des hierarchischen Clusterings:

- Es wird eine Matrix erstellt, bei der als Spalten die in allen n-Grammen vorkommenden unterschiedlichen Einheiten (Wortformen, Lemmata, Wortartklassen) aufgeführt werden. Für jedes n-Gramm (als Zeilen in der Matrix) wird mit den Werten 0 und 1 angegeben, ob die entsprechende Einheit darin vorkommt. Daraus ergibt sich für jede Mehrworteinheit ein Zahlenvektor aus Nullen und Einsen.
- Nun wird zwischen allen Vektoren die euklidische Distanz gemessen, um daraus ein Dendrogramm zu erstellen, das die Mehrworteinheiten so gruppiert, dass die jeweiligen Nachbarn möglichst ähnliche Vektoren aufweisen.
- Optional können die Mehrworteinheiten dann in eine Anzahl *k* Gruppen ähnlicher Mehrworteinheiten aufgeteilt werden.

Abbildung 1 zeigt einen Ausschnitt aus einem Dendrogramm. Die Endpunkte stellen die n-Gramme dar; die Kanten gruppieren die n-Gramme nach Ähnlichkeit.

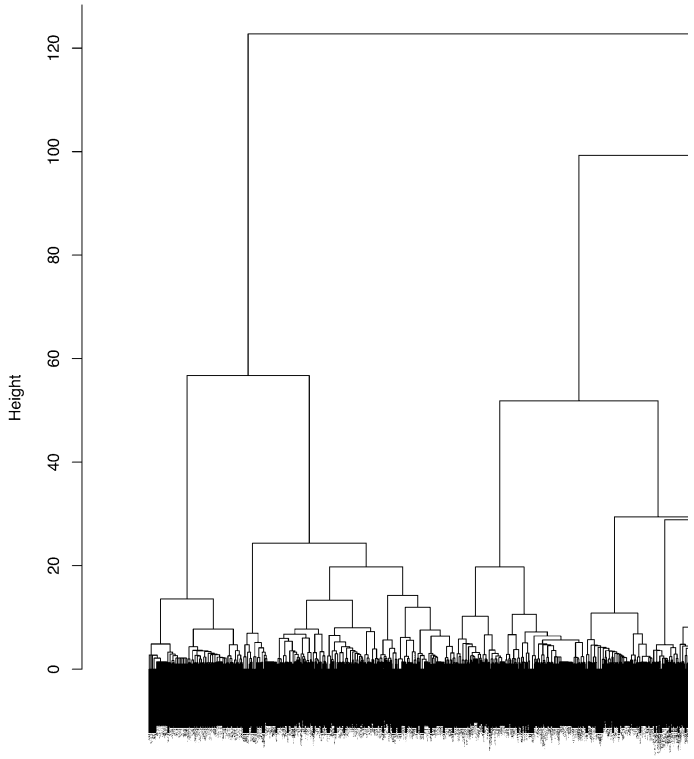

Abb. 1: Dendrogramm Clusteranalyse (Ausschnitt)

Damit werden nun die folgenden n-Gramme (alle typisch für die CDU) einer Gruppe zugeordnet, wobei sie untereinander ebenfalls nach Ähnlichkeit geordnet sind:

, dass es wir
, dass wir es
wissen , dass wir
wir wissen , dass
auch , dass wir
, dass wir auch
dass wir ,
, dass wir ,
, dass wir
, dass wir wir
darin , dass wir
haben , dass wir
sagen , dass wir
so , dass wir
aufpassen , dass wir
notwendig , dass wir
darum , dass wir
, dass wir hier
Meinung , dass wir
dass wir , wenn
darüber , dass wir
verweisen , dass wir
einig , dass wir
führen , dass wir

Wie das Beispiel zeigt, wird bei diesem Vorgehen die Reihenfolge der Einheiten im n-Gramm ignoriert. So werden die beiden Mehrworteinheiten „dass es wir“ und „dass wir es“ gleich nebeneinander aufgeführt, da sich deren Vektoren gar nicht unterscheiden. Diese Ordnung erleichtert die Analyse der Mehrworteinheiten deutlich.

3.2 Ergebnisse und Diskussion

Im Folgenden werden die Ergebnisse der Berechnungen in Teilen präsentiert und diskutiert. Die kompletten Ergebnisse können online eingesehen werden (www.bubenhofer.com/mwepol/).

3.2.1 Vorstudie: Keywords

Die Berechnung von Mehrworteinheiten ist nur eine der vielfältigen Möglichkeiten, um typische Sprachgebrauchsmuster in einem Korpus zu finden. Eine naheliegende Sonderform der n-Gramme ist die Berechnung von Unigrammen, die typisch für ein Teilkorpus sind, also der Berechnung von Keywords. Tabelle 4 zeigt die 20 typischsten nominalen Keywords (lemmatisiert).

Generell geben die Nomen politische Themen der Wahlperiode 17 wieder: *Atompolitik/Energiewende*, *Klimaschutz*, *Sozialwerke* (*Betreuungsgeld*, *Rente*, *Arbeitslosigkeit*), Außenpolitik etc. Interessanter sind jedoch parteitypische Schlüssel-, Schlag- und Fahnenwörter (Hermanns 1994): *„Atomkraftwerk“*, *„Klimaschutz“* (Grüne), *„Kernenergie“*, *„Wettbewerbsfähigkeit“*, *„Sicherheit“* (CDU/CSU), *„Krieg“*, *„Waffe“*, *„Armut“* (Linke). Dies setzt sich in anderen Wortartklassen fort, wie z. B. an *„ökologisch“* oder *„erneuerbar“* (Grüne) und *„anständig“* oder *„prekär“* (SPD) sichtbar ist.

Tab. 4: 20 typischste Nomen in vier Parteien im Bundestag (WP 17)

B90/Die Grünen	CDU/CSU	Die Linke	SPD
Bundesregierung	Land	Linke	Minister
Staatssekretär	Weg	Krieg	SPD-Bundestags-fraktion
Atomkraftwerk	Opposition	Bank	Sozialdemokrat
Schwarz-Gelb	Zukunft	Beschäftigte	Ministerin
Atomkraft	Bereich	Konzern	Herr
Klimaschutz	Rahmen	Rente	Staatssekretär
Leiharbeitskräfte	Erfolg	Lohn	Schwarz-Gelb
Einwanderer	Kernenergie	Waffe	Kanzlerin
Laufzeitverlängerung	Sicherheit	Bundesregierung	Sozialdemokratin
Subvention	Wettbewerbsfähigkeit	Skandal	SPD-Fraktion
Ministerin	Entwicklung	Armut	Regierung
Frage	Maßnahme	Euro	Staatssekretärin
Wirtschaftsminister	CSU-Bundestags-fraktion	Mensch	Betreuungsgeld
EU	Währung	Profit	Ankündigung
Regierungsfraktion	Dame	Erwerbslose	Kauder
Baustoff	Grundlage	Prozent	Vorschlag
Ankündigung	Ziel	Privatisierung	Brüderle
Kauch	Beitrag	Bevölkerung	Kürzung
Grüne	Bund	Leiharbeit	Steuersenkung
Antwort	Jahr	Osten	Frau

Keywords, basierend auf Wort- oder Grundformen, sind durchaus geeignet, um thematische Schwerpunkte, konkurrierende Wörter, Fahnenwörter etc. zu finden. Naturgemäß wird dabei aber der weitere Kontext ignoriert, weshalb Mehrworteinheiten eine wichtige Ergänzung dazu sind.

3.2.2 n-Gramme auf der Basis von Grundformen

Die manuelle Sichtung der nach einem hierarchischen Clustering geordneten Mehrworteinheiten zielt nun darauf, diese zu kategorisieren. Die Art der Kategorisierung ist dabei abhängig vom jeweiligen Forschungsinteresse. Bei den meisten Daten sind Phänomene auf semantischer, grammatischer und pragmatischer Ebene sichtbar.

Im Vergleich zu den Keyword-Analysen oben wird sofort deutlich, dass die Mehrworteinheiten Floskeln des Sprachgebrauchs wiedergeben, die weniger an Inhalten hängen, sondern pragmatische Funktionen erfüllen.

Tab. 5: Mehrworteinheiten (lemmatisiert), typisch für Bündnis 90/Grüne (Auswahl)

Mehrworteinheit	Freq. Korpus	Freq. Referenz	p
Kritisieren, fragen			
Sie|sie sicherstellen wollen ,	5	10	< 0.05
wie erklären Sie|sie	13	64	< 0.05
wie wollen Sie|sie eigentlich	6	14	< 0.05
glauben Sie|sie eigentlich ,	5	13	< 0.05
Sie|sie haben versuchen ,	10	45	< 0.05
Sie|sie sagen , dass	59	417	< 0.0001
Verschleierung/Zusammenhänge			
nichts andere als eine	31	228	< 0.01
schon bezeichnend , dass	5	13	< 0.05
es sein schon interessant	9	37	< 0.05
Anklage			
keine Wort dazu ,	5	9	< 0.05
einfach ignorieren ,	5	9	< 0.05
können Sie|sie ausschließen ,	8	25	< 0.05
Argument aus der Übereinstimmung aller			
wir alle wissen :	18	99	< 0.01
Metaphorizität/Topoi			
Schritt in die richtig	36	222	< 0.001
in die Praxis	85	759	< 0.05
in die Versenkung verschwinden	7	10	< 0.01
Regen stehen lassen .	9	34	< 0.05

Tabelle 5 zeigt eine Reihe von Mehrworteinheiten, die typisch für die Grünen im Vergleich zu allen anderen Mehrworteinheiten sind. Auffällig sind Mehrworteinheiten, die den typischen Sprachgebrauch einer Oppositionspartei widerspiegeln. So gibt es auffällig viele Muster, die mit dem Personalpronomen *Sie* vornehmlich die Regierung adressieren („wie wollen Sie eigentlich…“) und (rhetorische) Fragen stellen. Einige der Mehrworteinheiten tragen einen anklagenden Charakter („wie können Sie ausschließen…“, „kein Wort dazu…“) oder dienen dazu, eine angebliche Verschleierung anzuprangern („nichts anderes als eine…“). Auch die Referenz auf eine angebliche Mehrheit („wir alle wissen…“) dient der Behauptung, die wahren Zusammenhänge zu erkennen (Eggler 2006, 44).

Tabelle 6: Mehrworteinheiten (lemmatisiert), typisch für die CDU (Auswahl)

Mehrworteinheit	Freq. Korpus	Freq. Referenz	p
Überzeugen			
davon überzeugt , dass	190	349	< 0.0001
zuversichtlich , dass	100	183	< 0.01
froh , dass wir	106	213	< 0.05
Wir-Gefühl			
haben wir in Deutschland	54	92	< 0.05
dies Jahr haben wir	29	40	< 0.05
wir Deutsche haben , sondern wir müssen	78	162	< 0.05
Mensch in unser Land	253	493	< 0.0001
Bürger in unser Land	33	51	< 0.05
unser Soldat ,	26	36	< 0.05
Verteidigung			
sein schlichtweg falsch .	26	36	< 0.05
richtig , dass wir	302	544	< 1e-06
Sie|sie wissen ganz genau	58	111	< 0.05
Werte			
Hilfe zu Selbsthilfe	59	101	< 0.05
sozial Marktwirtschaft ,	49	84	< 0.05
Gott sein dank	132	239	< 0.01
ich ganz offen	30	42	< 0.05
unser freiheitlich-demokratisch Grundordnung	29	37	< 0.05

Bei den Mehrworteinheiten, die typisch für die CDU sind, zeigen sich andere kommunikative Funktionen (Tabelle 6). Als Regierungspartei müssen CDU-Angehörige das Parlament von Gesetzen überzeugen, indem sie sich selbstgewiss geben („davon überzeugt, dass…“, „froh, dass wir…“). Es gilt sich zu verteidigen („ist schlichtweg falsch“, „Sie wissen ganz genau…“) und natürlich versucht auch die CDU zu behaupten, für eine Mehrheit zu stehen. Dies realisiert sie auffällig oft über Referenzen auf das Land („haben wir in Deutschland“, „wir Deutsche haben…“) oder die Bürger/innen („Menschen/Bürger in unserem Land“). Häufig sind auch Floskeln, die Werte widerspiegeln („Hilfe zur Selbsthilfe“, „soziale Marktwirtschaft“, „unsere freiheitlich-demokratische Grundordnung“), die teilweise zu Fahnenwörtern der Partei gehören oder wie „ich ganz offen“ für kommunikative Ideale stehen (Schröter 2011).

Die Tabellen (die nur eine kleine Auswahl der Mehrworteinheiten zeigen) enthalten neben den Mehrworteinheiten auch Angaben zu den absoluten Frequenzen, mit denen die n-Gramme in den Korpora vorkommen, sowie das Signifikanzniveau (p),

also die statistische Sicherheit dafür, dass die Mehrworteinheit signifikant häufiger im jeweiligen Parteienkorpus vorkommt als im Referenzkorpus. Diese Werte unterstützen den Interpretationsprozess, indem etwa die Bedeutung der Mehrworteinheit eingeschätzt werden kann. Zusätzlich sind Distributionsanalysen der Mehrworteinheiten interessant, um zu sehen, ob diese breit über mehrere Sprecherinnen und Sprecher und/oder Debatten streuen oder idiosynkratisch sind.

3.2.3 Komplexe n-Gramme

Die Listen der parteitypischen komplexen n-Gramme sind schwieriger zu lesen. So ist beispielsweise für die Grünen das n-Gramm

> Wie/KOUS VMFIN Sie/PPER ADV

typisch, also die Konjunktion *wie* gefolgt von einem finiten Modalverb, dem Personalpronomen *Sie* und einem Adverb. Dieses Muster kommt in den Daten der Grünen 16 Mal vor, wobei es sprachlich folgendermaßen realisiert wird:

> Wie können Sie denn
> Wie wollen Sie denn
> Wie wollen Sie eigentlich
> Wie wollen Sie da

Das Muster streut zudem über 12 unterschiedliche Personen und kommt beispielsweise in folgenden Belegen vor:

> wir übrigens weit über 1 Billion Euro. **Wie können Sie denn** behaupten, dass diese Kosten eine Belastung für (28.02.13, Hans-Josef Fell)

> entziehen dem System Schiene Milliarden von Euro. **Wie wollen Sie da** etwas erreichen? (11.06.10, Dr. Anton Hofreiter)

> Das ist ein Ablasshandel zur Naturzerstörung . **Wie wollen Sie eigentlich** Ländern wie Brasilien und Indonesien erklären, dass (11.11.09, Bärbel Höhn)

> **Wie wollen Sie denn** Frauen in Unternehmen in Deutschland kriegen, wenn (28.03.12, Renate Künast)

> Wo ist denn das inhaltliche Konzept? **Wie wollen Sie denn** die gesellschaftliche Ausgrenzung benachteiligter Gruppen stoppen? (25.11.10, Stephan Kühn)

Erst mit diesen Informationen kann das Muster eingeschätzt und interpretiert werden. Daher ist es sinnvoll, bestimmte Kennzahlen zu berechnen und weiterfüh-

rende Angaben bereit zu stellen. Die Produktivität eines Musters, also die Anzahl unterschiedlichen Realisierungsvarianten, die auf ein Muster kommen, kann beispielsweise durch ein Type-Token-Verhältnis ausgedrückt werden: Im vorliegenden Fall liegt das bei ¼, also 0,25.

Im Vergleich zum komplexen n-Gramm oben ist das Muster

VMFIN Sie/PPER ADV ADV

weit produktiver; das Type-Token-Verhältnis liegt bei 0,01 (1/78). Realisierungen sind beispielsweise:

Können Sie bitte einmal
Können Sie hier noch
können Sie schon jetzt
könnten Sie hier sofort
müssen Sie endlich einmal
müssen Sie schon selbst
wollen Sie dann weiterhin
wollen Sie denn da
wollen Sie denn eigentlich
wollen Sie doch sicherlich
wollen Sie sogar noch

Damit ist dieses Muster auch weniger deutlich auf eine bestimmte pragmatische Funktion zu reduzieren – trotzdem lassen sich die Realisierungen auf das gemeinsame Muster zurückführen, das generell die Funktionen der Aufforderung aber auch des (rhetorischen) Fragens und Anklagens erfüllt und die typische Funktion einer Oppositionspartei beschreibt.

Tabelle 7 enthält eine kleine Auswahl von komplexen n-Grammen, die typisch für die Grünen sind, kategorisiert nach drei Bereichen. Auffallend sind Mehrworteinheiten, die kritisierende Funktionen aufweisen, wie z. B. die Konstruktion „ADV machen Sie ADV“:

Abgemacht? – Nein. **Jetzt machen Sie wieder** einen Rückzieher; das kennen wir schon. (Anton Hofreiter, 14. 4. 2011)

Da machen Sie überhaupt nichts. Da blockieren Sie nur. (Oliver Krischer, 8. 3.2012)

Damit wird dem politischen Gegner, meistens der Regierungskoalition, inkonsequentes Handeln vorgeworfen. Dies geschieht auch mit dem Muster „Sie haben ADV gesagt“:

[...] kann ich Ihnen immer noch nicht folgen. **Sie haben selber gesagt**, es gebe durchaus vorrangig zu bedienende Gläubiger [...]. (Hans-Christian Ströbele, 19. 5. 2010)

Oppositionelle Parteien kritisieren aber nicht nur, sondern machen auch Vorschläge für alternatives Handeln, wie das Muster „ADV wäre es ADJD“ zeigt:

> Wir wissen, dass die Aufstockung kommt. Aber meinen die Bundeskanzlerin und die Koalition nicht, dass man der Bevölkerung einmal reinen Wein einschenken sollte? Das führt mich zu Griechenland. **Hier wäre es notwendig**, deutlich zu sagen: Von Griechenland sind Anstrengungen notwendig. (Priska Hinz, 9. 2. 2012)

Allerdings könnte die pragmatische Funktion solcher Formulierungen komplexer sein, als nur eine alternative Handlung vorzuschlagen („es wäre besser, eine deutliche Aussage zu machen“). Zusätzlich scheint suggeriert zu werden, dass dieses Handeln vom Kritisierten sowieso nicht erwartet werden kann (vgl. für eine ausführliche Diskussion dazu Bubenhofer 2008; Hein/Bubenhofer 2015).

Tab. 7: Komplexe n-Gramme mit Realisierungen, typisch für die Grünen (Auswahl)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Inkonsequenz unterstellen				
ADV machen/VVFIN Sie/PPER ADV	17	47	< 0.01	0,06
Da machen Sie überhaupt				
dann machen Sie auch				
dann machen Sie bitte				
dann machen Sie doch				
Nun machen Sie aber				
Stattdessen machen Sie nur				
Stattdessen machen Sie wieder				
Sie/PPER ADV keine/PIAT NN	21	85	< 0.05	0,04
Sie auch keine Lösung				
Sie hier keine Antwort				
Sie hier keine Camouflage				
Sie hier keine Bereitschaft				
Sie also keine Unwahrheiten				
Sie/PPER haben/VAFIN ADV gesagt/VVPP	61	320	< 0.01	0,05
Sie haben eben/selber/nämlich/außerdem/ja gesagt				
Verbote/Ideale aussprechen				
NN VMFIN sich/PRF nicht/PTKNEG	28	120	< 0.05	0,03
Rassismus darf sich nicht				
Beschaffung darf sich nicht				
Staat darf sich nicht				
Behandlung dürfen sich nicht				
Kompromisses dürfen sich nicht				
Frauen müssen sich nicht				
Bundesregierung sollte sich nicht				

Tab. 7 (fortgesetzt)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Fragen und kommentieren				
ADV ADJD gefragt/VVPP	11	25	< 0.05	0,1
ganz konkret/deutlich/spezifisch/klar gefragt				
etwas salopp/präziser gefragt				
ADV wäre/VAFIN es/PPER ADJD	25	117	< 0.05	0,4
so wäre es besser				
Heute wäre es richtig				
Vielleicht wäre es sinnvoll				
Hier wäre es notwendig				
So wäre es dringend				
Insofern wäre es wirklich				

Weiterhin finden sich bei den Grünen typischerweise Aussagen, die über Modalverben Verbote aussprechen oder Ideale benennen („Rassismus/der Staat darf sich nicht"; „Frauen müssen sich nicht").

Bei der Partei „Die Linke", während der untersuchten Wahlperiode ebenfalls in der Opposition, finden sich ähnliche Muster, die auf diese Rolle zurückzuführen sind (vgl. Tabelle 8). Allerdings finden sich verstärkt Muster, die das Personalpronomen „Sie" enthalten und damit direkt die Regierung angreifen („Erklären Sie doch einmal", „wissen Sie eigentlich auch" etc.). Außerdem finden sich, wie bei allen Parteien, Mehrworteinheiten, die Zahlen beinhalten. Bei der Linken allerdings werden Zahlennennungen typischerweise in dass-Konstruktionen verwendet:

> Die Bundeskanzlerin hat in der gesamten heutigen Debatte keinen einzigen Satz zur Entwicklungspolitik gesagt. Keinen einzigen Satz! Angesichts der großen Herausforderungen, **dass fast 1 Milliarde** Menschen hungert und dass aufgrund der Wirtschaftskrise noch mehr Menschen in Armut gefallen sind, zeigt dies die Prioritätensetzung dieser Regierung. Es zeigt auch, dass diese Bundesregierung nicht nur in Deutschland, sondern auch in der internationalen Politik ein Totalausfall ist. (Heike Hänsel, 15. 9. 2010)

Bei der CDU/CSU scheinen die Zahlen nennenden Muster auf den ersten Blick eher deskriptiver Natur zu sein: „Ab dem Jahre CARD", Nennung von Paragraphen „NN nach NN CARD" („Telekommunikationsüberwachung nach § 100") oder Nennung von Beträgen und Anteilen: „von ADV CARD NN" – „von rund/etwa/fast/lediglich x Prozent/Milliarden".

Wiederum typisch für Oppositionsparteien sind „wir fordern"-Konstruktionen, wobei bei der Linken mit dem Muster „wir fordern ART ADJA" typischerweise Adjektive in verstärkender Funktion eingesetzt werden („wir fordern die bedingungslose/vollständige/sofortige [...]").

Tab. 8: Komplexe n-Gramme mit Realisierungen, typisch für die Linke (Auswahl)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Kritisieren, Inkonsequenzen aufzeigen				
Erklären/VVFIN Sie/PPER das/PDS ADV Erklären Sie das einmal/doch/mal	7	9	< 0.05	0,3
ADV sagen/VVFIN Sie/PPER jetzt/aber/hier/nun/dann sagen Sie	80	310	< 0.05	0,5
Wissen/VVFIN Sie/PPER ADV ,/$, Wissen Sie eigentlich/auch,	24	53	< 0.01	0,25
NN gehört/VVFIN ADV ,/$, dass/KOUS ART Wahrheit/Praxis/Realität/Würde gehört auch, dass die/der	7	17	< 0.05	0,2
ADV weitermachen/VVFIN wie/KOKOM ADV so/nur weitermachen wir bisher/zuvor	12	25	< 0.05	0,3
VMFIN ,/$, VMFIN ADV will, muss auch/endlich soll/kann, muss auch darf, kann heute will, soll weiterhin	45	193	< 0.05	0,03
Es/PPER VMFIN nicht/PTKNEG sein/VAINF ,/$, dass/KOUS Es kann/darf nicht sein, dass	53	372	< 0.05	0,3
ADJD ist/VAFIN ADV ,/$, dass/KOUS Sie/PPER Schade ist auch/nur, dass Sie Wahr ist vielmehr, dass Sie dramatischer ist aber, dass Sie unverständlicher ist aber, dass Sie Traurig ist allerdings, dass Sie	10	25	< 0.05	0,1
NN ADJD meint/VVFIN ,/$, Menschenrechtspolitik ernst meint, Kinderschutz ernst meint, Machtkontrolle ernst meint, UN-Behindertenrechtskonvention ernst meint, Bildungsdienst ernst meint, Beschäftigungssystem ernst meint,	13	01.02.04	< 0.05	0,08
Mit Zahlen argumentieren				
,/$, dass/KOUS ADV CARD , dass inzwischen [Zahl] , dass noch/nur [Zahl]	58	274	< 0.05	0,02
,/$, dass/KOUS ADV mehr/PIAT , dass zehnmal/immer/noch/wieder mehr	30	116	< 0.05	0,2

Tab. 8 (fortgesetzt)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Fordern				
Wir/PPER fordern/VVFIN ART ADJA Wir fordern ein umfassendes Wir fordern die sofortige/bedingungslose/vollständige Wir fordern den rechtssicheren	25	65	< 0.01	0,05

Die CDU/CSU-Vertreter/innen im Parlament argumentieren weniger angriffig (vgl. Tabelle 9). Viele Mehrworteinheiten haben das Potenzial, ein Wir-Gefühl auszulösen („unsere Bürgerinnen und Bürger", „unserer gemeinsamen Sache", „wirtschaftliche Erfolg unserer deutschen Unternehmen"). Zudem müssen die Angriffe der Oppositionsparteien abgewehrt werden, indem bestimmte politische Handlungen legitimiert werden. Dies geschieht z. B. durch Behauptung von Deutlichkeit und Offenheit:

> Abgewandert wird, weil in unserem Land gerade Hochschulabsolventen schlicht und ergreifend zu wenig gezahlt wird, weil sie mit Praktika abgespeist werden, während sie in anderen Ländern sofort unbefristete Anstellungen bekommen. Deswegen sage **ich in aller Deutlichkeit**: Es ist auch ein ganz entscheidender Beitrag der deutschen Wirtschaft gefragt, selbst etwas dafür zu tun, um attraktiver im Kampf um die klugen Köpfe in aller Welt zu werden. (Reinhard Grindel, 28. 10. 2010)

Um Gesetzesvorlagen zu legitimieren, beruft sich die CDU/CSU häufig auf „die Zukunft unseres Landes" – oder auch die „Zukunftsfähigkeit", „Sicherheit" oder „Fortentwicklung" des Landes. Die Vertreter/innen der Linken verwenden typischerweise weniger oft Referenzen auf die „Zukunft" und es findet sich kein signifikantes Muster, das von der Zukunft „unseres Landes" spricht.

Zur Thematisierung der Zukunft passend verwendet die CDU/CSU signifikant häufiger als die anderen Parteien die Phrase „wir sind auf einem guten Weg" (und Varianten), die meist eine beschwichtigende Funktion haben:

> Was Sie gesagt haben, ist nicht wahr! Das hat auf der Seite der Fachpolitiker natürlich keinen Jubel hervorgerufen; aber **wir sind auf dem richtigen Weg** und werden diesen Weg in den nächsten Beratungen fortsetzen. Wir haben eine Schuldenbremse in das Grundgesetz eingebaut. Diese Schuldenbremse wird erstmals 2011 Wirkung zeigen. (Jürgen Herrmann, 20. 5. 2010)

Eine ähnliche, beschwichtigende Funktion könnte der Mehrworteinheit „das eine oder andere Mal", „die eine oder andere Diskussion" etc. (mit Varianten) zugesprochen werden.

> Wir stehen am Anfang dieser neuen Regierungskoalition, die wir uns als Wahlziel gewünscht haben. Deswegen bin ich mir, auch wenn es **die eine oder andere Diskussion** gibt – wo gibt es sie nicht? –, sicher, dass uns dieser gemeinsame Wunsch, unserem Land zu helfen, aus der Krise herauszukommen, neue Perspektiven zu entwickeln und jungen Menschen Chancen zu geben, der getragen davon ist, Deutschland in eine gute Zukunft zu führen, die Kraft geben wird, nicht nur am Anfang stark zu sein, sondern über vier Jahre hinweg stark zu bleiben. (Volker Kauder, 10. 11. 2009)

Die großen Meinungsverschiedenheiten während der Koalitionsbildung relativiert Kauder im Zitat als „die eine oder andere Diskussion", um danach die gemeinsame Grundlage der Koalition zu beschwören.

Tab. 9: Komplexe n-Gramme mit Realisierungen, typisch für die CDU/CSU (Auswahl)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Wir-Gefühl				
ADJA NN unserer/PPOSAT ADJA	87	149	<0.05	0,01
identitätsstiftendes Merkmal unserer deutschen				
wesentlichen Punkte unserer Auswärtigen				
nachvollziehbarer Wunsch unserer skandinavischen				
wirtschaftliche Erfolg unserer deutschen				
übergreifendes Ziel unserer gesamten				
unverzichtbare Grundlage unserer deutschen				
wichtiger Teil unserer gegenseitigen				
unserer/PPOSAT ADJA NN	729	1340	<0.0001	0,002
unserer pluralistischen Gesellschaft				
unserer gemeinsamen Sache				
unserer gemeinsamen Verantwortung				
unserer gemeinsamen Aufgabe				
unserer ganzen Kraft				
unserer schnelllebigen Zeit				
unserer/PPOSAT NN und/KON NN	161	304	< 0.05	0,01
unserer Kolleginnen und Kollegen				
unserer Bürgerinnen und Bürger				
unserer Städte und Gemeinden				
unserer Soldatinnen und Soldaten				
Legitimierung				
in/APPR aller/PIAT NN sagen/VVINF	47	70	< 0.05	0,14
in aller Deutlichkeit/Klarheit/Offenheit sagen				
für/APPR ART NN unseres/PPOSAT	84	134	< 0.05	0,02
für die Fortentwicklung unseres				
für die Sicherheit unseres				
für die Zukunft unseres				
für die Zukunftsfähigkeit unseres				
für eine Weiterentwicklung unseres				

Tab. 9 (fortgesetzt)

Mehrworteinheit	Freq. K.	Freq. R.	p	TTR
Relativierung/Beschwichtigung				
ART eine/PIS oder/KON andere/PIS NN das eine oder andere Mal der eine oder andere Staat der eine oder andere Kollege	159	286	< 0.01	0,007
wir/PPER sind/VAFIN auf/APPR ART ADJA NN wir sind auf einem guten Weg wir sind auf dem richtigen Weg	41	50	< 0.01	0,2

Es sollte klar geworden sein, dass die berechneten Mehrworteinheiten hinsichtlich unterschiedlicher Forschungsinteressen gedeutet werden können. Die skizzenhaften Analysen müssen an dieser Stelle genügen – das Material steht jedoch für weitere Analysen als Datenbank für die Öffentlichkeit zur Verfügung: www.bubenhofer.com/mwepol/.

4 Fazit

Die Beispielanalyse sollte deutlich gemacht haben, dass die Berechnung von Kollokationen und Mehrworteinheiten alleine keine ausreichende Methode ist, um in politolinguistischer Perspektive den typischen Sprachgebrauch von Akteuren, Gruppen, Institutionen etc. zu untersuchen. Sie ist jedoch ein interessanter Ausgangspunkt, um datengeleitet auf typische Formulierungsmuster aufmerksam zu werden, die dann weiter qualitativ und quantitativ untersucht werden. So können politolinguistische Konzepte wie Schlagwörter, Topoi, Argumentationsfiguren etc. hinzugezogen werden, um die berechneten Sprachgebrauchsmuster zu klassifizieren und zu deuten. Zudem können weitere quantitative Korpusanalysen eingesetzt werden, um beispielsweise die zeitliche Streuung oder die Verteilung des Musters über verschiedene Themenbereiche oder Domänen zu untersuchen.

Die Chance des datengeleiteten Vorgehens liegt darin, auf Muster aufmerksam zu werden, die nicht a priori bekannt gewesen sind. Es handelt sich also um ein Verfahren der Generierung von Hypothesen, die in einem zweiten Schritt auch über andere Methoden getestet werden können.

Ebenso bedeutend ist aber der Fokus auf die sprachliche Oberfläche (vgl. zur Rehabilitierung der Oberfläche Feilke/Linke 2009), die vor dem Hintergrund einer sozial- und kulturwissenschaftlich interessierten Linguistik nicht nur als Ergebnis des Zusammenspiels von Lexik und Grammatik angesehen wird, sondern als Emergenzphänomen, das soziales Handeln widerspiegelt (Feilke 1996; Bubenhofer 2009).

In sprachlichen Mustern ist eine pragmatische Komponente eingeschrieben; sie zu entdecken hilft, politisches Sprachhandeln von der Beobachtung der sprachlichen Oberfläche aus zu beschreiben. Die Erkenntnis, dass sprachliche Muster Resultat von sozialem Handeln sind, stellt für die Politolinguistik natürlich keine Neuigkeit dar, wie der Fokus auf Analysekategorien wie Schlag- und Wertwörter (Hermanns 1994; Burkhardt 2003) oder Argumentationstopoi (Wengeler 2003a, 2003b) zeigt. Diese sprachlichen Muster aber oberflächennah zu denken, ist der innovative Beitrag einer korpuslinguistischen Sicht auf Sprachgebrauch. Da sprachliche Muster rekurrente Phänomene auf der sprachlichen Oberfläche sind, können sie maschinell als statistisch überzufällig auftretende Kombinationen aus Wortformen, Lexemen und/oder Wortartklassen (und weiteren Elementen) berechnet werden.

5 Literatur

Anthes, Gary (2010): Topic models vs. unstructured data. In: Communications of the ACM (53), 16–18.

Banerjee, Satanjeev/Ted Pedersen (2003): The design, implementation, and use of the ngram statistic package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City.

Bartsch, Sabine (2004): Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence. Tübingen.

Blätte, Andreas (2013): PolMine-Plenardebattenkorpus, Duisburg/Essen,www.polmine.sowi.uni-due.de/daten.html. [04.03.2014]

Blätte, Andreas (2012): Unscharfe Grenzen von Policy-Feldern im parlamentarischen Diskurs. Messungen und Erkundungen durch korpusunterstützte Politikforschung. In: Zeitschrift für Politikwissenschaft (22), 35–68.

Blei, David M./Andrew Y. Ng/Michael I. Jordan (2003): Latent dirichlet allocation. In: Journal of Machine Learning Research (3), 993–1022.

Bluhm, Claudia/Dirk Deissler/Joachim Scharloth u. a. (2000): Linguistische Diskursanalyse. Überblick, Probleme, Perspektiven. In: Sprache und Literatur in Wissenschaft und Unterricht (88) , 3–9.

Bondi, Marina/Mike Scott (2010): Keyness in texts. Amsterdam/Philadelphia.

Bubenhofer, Noah (2008): „Es liegt in der Natur der Sache…". Korpuslinguistische Untersuchungen zu Kollokationen in Argumentationsfiguren. In: Carmen Mellado Blanco (Hg.): Studien zur Phraseologie aus textueller Sicht. Hamburg, 53–72.

Bubenhofer, Noah (2014): Geokollokationen – Diskurse zu Orten. Visuelle Korpusanalyse. In: Sondernummer Mitteilungen des Deutschen Germanistenverbandes. Korpora in der Linguistik – Perspektiven und Positionen zu Daten und Datenerhebung (1), 45–59.

Bubenhofer, Noah (2013): Skandalisierung korpuslinguistisch. Ein empirisch-linguistischer Blick auf die Berichterstattung zur „Wulff-Affäre". In: Linguistik online 4 (61), www.linguistik-online.de/61_13/bubenhofer.html. [04.03.2014]

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

Bubenhofer, Noah/Tobias Dussa/Sarah Ebling u. a. (2009): „So etwas wie eine Botschaft." Korpuslinguistische Analysen der Bundestagswahl 2009. In: Sprachreport (4) , 2–10.

Bubenhofer, Noah/Joachim Scharloth/David Eugster (2014): Rhizome digital. Datengeleitete Methoden für alte und neue Fragestellungen in der Diskursanalyse. In: Zeitschrift für Diskursforschung, Sonderheft Diskurs, Interpretation, Hermeneutik.
Burkhardt, Armin (2003): Das Parlement und seine Sprache. Studien zu Theorie und Geschichte parlamentarischer Kommunikation. Tübingen (Reihe Germanistische Linguistik, 241).
Dzudzek, Iris/Georg Glasze/Annika Mattissek u. a. (2009): Verfahren der lexikometrischen Analyse von Textkorpora. In: Handbuch Diskurs und Raum. Theorien und Methoden für die Humangeographie sowie die sozial- und kulturwissenschaftliche Raumforschung. Bielefeld, 233–260.
Ebling, Sarah (2010): Korpusgeleitete Zugänge zur Rhetorik deutscher und schweizerischer Politiker am Beispiel von Peer Steinbrück und Hans-Rudolf Merz. In: Kersten Sven Roth/Christa Dürscheid (Hg.): Wahl der Wörter – Wahl der Waffen? Politische Sprache und Kommunikation in der Schweiz. Bremen (Sprache – Politik – Gesellschaft), 79–101.
Ebling, Sarah/Joachim Scharloth/Tobias Dussa u. a. (2014): Gibt es eine Sprache des politischen Extremismus? In: Frank Liedtke (Hg.): Die da oben. Texte, Medien, Partizipation. Bremen, 43–68.
Eggler, Marcel (2006): Argumentationsanalyse textlinguistisch. Argumentative Figuren für und wider den Golfkrieg von 1991. Tübingen (Reihe Germanistische Linguistik, 268).
Evert, Stefan (2009): 58. Corpora and collocations. In: Anke Lüdeling/Merja Kytö (Hg.): Corpus Linguistics. Berlin/New York (Handbücher zur Sprach- und Kommunikationswissenschaft, 29), 1212–1248.
Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt a. M.
Feilke, Helmuth/Angelika Linke (Hg.) (2009): Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt. Berlin/New York.
Felder, Ekkehard/Marcus Müller/Friedemann Vogel (2011): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin/Boston.
Firth, John Rupert (1957): Modes of meaning. In: Papers in Linguistics 1934–1951. London, 190–215.
Girnth, Heiko (2002): Sprache und Sprachverwendung in der Politik. Eine Einführung in die linguistische Analyse öffentlich-politischer Kommunikation. Tübingen (Germanistische Arbeitshefte).
Girnth, Heiko/Constanze Spieß (2006): Strategien politischer Kommunikation. Pragmatische Analysen. Berlin.
Glasze, Georg (2007): Operationalisierung der Diskurstheorie nach Laclau und Mouffe in einer Triangulation von lexikometrischen Verfahren und der Analyse narrativer Muster. In: Forum Qualitative Sozialforschung/Forum. Qualitative Social Research 15 (12).
Hein, Katrin/Noah Bubenhofer (2015): Korpuslinguistik konstruktionsgrammatisch. Diskursspezifische n-Gramme zwischen statistischer Signifikanz und semantisch-pragmatischem Mehrwert. In: Alexander Lasch/Alexander Ziem (Hg.): Konstruktionsgrammatik IV. Konstruktionen als soziale Konventionen und kognitive Routinen. Tübingen, 179–206.
Hermanns, Fritz (1994): Schlüssel-, Schlag- und Fahnenwörter. Zu Begrifflichkeit und Theorie der lexikalischen „politischen Semantik". Mannheim (Arbeiten aus dem Sonderforschungsbereich 245, 81).
Jung, Matthias (2001): Diskurshistorische Analyse – eine linguistische Perspektive. In: Andreas Hirseland/Reiner Keller/Werner Schneider u. a. (Hg.): Handbuch Sozialwissenschaftliche Diskursanalyse. Band 1. Theorien und Methoden. Opladen, 29–52.
Kilgarriff, Adam (2001): Comparing corpora. In: International Journal of Corpus Linguistics 6 (1), 1–37.
Koller, Veronika (2006): Of critical importance. Using electronic text corpora to study metaphor in business media discourse. In: Anatol Stefanowitsch/Stefan Thomas Gries (Hg.): Corpus-Based

Approaches to Metaphor and Metonymy. Berlin (Trends in linguistics. Studies and monographs), 237–266.
Koller, Veronika/Michael Farrelly (2010): Darstellungen der Finanzkrise 2007/2008 in den britischen Printmedien. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 6 (2), 170–192.
Laver, Michael/Kenneth Benoit/John Garry (2003): Extracting policy positions from political texts using words as data. In: American Political Science Review 97 (2), 311–331.
Lebart, Ludovic/André Salem (1994): Statistique textuelle. Paris.
Mattissek, Annika (2005): Diskursive Konstitution von Sicherheit im öffentlichen Raum am Beispiel Frankfurt a. M. In: Georg Glasze/Robert Pütz/Manfred Rolfes (Hg.): Diskurs – Stadt – Kriminalität. Städtische (Un-)Sicherheiten aus der Perspektive von Stadtforschung und Kritischer Kriminalgeographie. Bielefeld, 105–136.
Olsen, Mark/Louis-Georges Harvey (1988): Computers in intellectual history. Lexical statistics and the analysis of political discourse. In: Journal of Interdisciplinary History 18 (3), 449–464.
Rayson, Paul/Roger Garside (2000): Comparing corpora using frequency profiling. In: Proceedings of the Workshop on Comparing Corpora. Morristown, 1–6.
Rohrdantz, Christian/Annette Hautli/Thomas Mayer u. a. (2012): Towards Tracking Semantic Change by Visual Analytics. www.kops.ub.uni-konstanz.de/handle/urn:nbn:de:bsz:352-186381. [04.03.2014]
Scharloth, Joachim/David Eugster/Noah Bubenhofer (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Dietrich Busse/Wolfgang Teubert (Hg.): Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden, 345–380.
Schiller, Anne/Simone Teufel/Christine Thielen (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart.
Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop.
Scholz, Ronny (2010): Die diskursive Legitimation der Europäischen Union. Eine lexikometrische Analyse zur Verwendung des sprachlichen Zeichens Europa/Europe in deutschen, französischen und britischen Wahlprogrammen zu den Europawahlen zwischen 1979 und 2004. Magdeburg, www.nbn-resolving.de/urn:nbn:de:101:1-201108243629. [04.03.2014]
Schröter, Juliane (2011): Offenheit. Die Geschichte eines Kommunikationsideals seit dem 18. Jahrhundert. Berlin/Boston.
Schröter, Melani/Björn Carius (2009): Vom politischen Gebrauch der Sprache. Wort, Text, Diskurs. Eine Einführung. Frankfurt a. M. (Leipziger Skripten. Einführungs- und Übungsbücher, 5).
Scott, Mike/Chris Tribble (2006): Textual patterns. Key words and corpus analysis in language education. Amsterdam.
Silva, Joaquim Ferreira da/Gabriel Pereira Lopes (1999): A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In: Sixth Meeting on Mathematics of Language. Orlando, 369–381.
Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.
Spitzmüller, Jürgen/Ingo H. Warnke (2011): Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin/Boston.
Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen.
Storjohann, Petra/Melani Schröter (2011): Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise und die (Un-) Ordnung des Ausgeblendeten. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 7 (1), 32–53.
Teubert, Wolfgang (2006): Korpuslinguistik, Hermeneutik und die soziale Konstruktion von Wirklichkeit. In: Linguistik online 28 (3), 41–60.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam (Studies in Corpus linguistics, 6).

Vogel, Friedemann (2012): Das LDA-Toolkit Korpuslinguistisches Analyseinstrument für kontrastive Diskurs- und Imageanalysen in Forschung und Lehre. In: Zeitschrift für angewandte Linguistik 57 (1).

Vogel, Friedemann (2010): Linguistische Imageanalyse (LIma). Grundlegende Überlegungen und exemplifi zierende Studie zum öffentlichen Image von Türken und Türkei in deutschsprachigen Medien. In: Deutsche Sprache (4), 345–377.

Wengeler, Martin (2003a): Argumentationstopos als sprachwissenschaftlicher Gegenstand. Für eine Erweiterung linguistischer Methoden bei der Analyse öffentlicher Diskurse. In: Susan Geideck/Wolf-Andreas Liebert (Hg.): Sinnformeln. Linguistische und soziologische Analysen von Leitbildern, Metaphern und anderen kollektiven Orientierungsmustern. Berlin/New York (Linguistik – Impulse & Tendenzen), 59–82.

Wengeler, Martin (2003b): Topos und Diskurs. Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960–1985). Tübingen (Reihe Germanistische Linguistik, 244).

Ziem, Alexander/Ronny Scholz/David Römer (2013): Korpusgestützte Zugänge zum öffentlichen Sprachgebrauch. Spezifisches Vokabular, semantische Konstruktionen, syntaktische Muster in Diskursen über „Krisen“. In: Ekkehard Felder (Hg.): Faktizitätsherstellung in Diskursen. Die Macht des Deklarativen. Berlin/New York, 329–358.

Zinsmeister, Heike/Ulrich Heid (2003): Significant triples. Adjective+noun+verb combinations. In: Proceedings of the 7th Conference on Computational Lexicography and Text Research. Budapest, 92–102.