Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse

Bubenhofer, Noah

DOI: https://doi.org/10.1515/9783110209372.6.407

Noah Bubenhofer

Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse

Discourse analysis has always relied on text corpora as its empirical basis. However, in the last decade corpus linguistics developed a rich variety of methods and tools which can be used to improve discourse analysis. It is now possible to process large corpora and have all the data at one's fingertips. Even more important for the analysis of discourse are, in my opinion, the possibilities of the ‚corpus driven' approach in corpus linguistics. This approach aims at unfolding patterns of language use and variation in the data, and deducing hypotheses from these observations – inductively. This ‚corpus driven' approach matches with important assumptions of Foucault's idea of discourse. At the centre of the approach lies the calculation of multiword units which are typical for diachronically or synchronically defined parts of language use. These multiword units can be automatically extracted from the corpus data and provide a basis for further investigations. They are indicators for metaphors, topoi, reasoning patterns and terms which play an important role in a discourse. But more general, these multiword units are also the result of patterns in language use and are thus indicators for the forces of discourses. These forces define what can be said in (and about) a specific discursive field. The framework proposed in this paper – i. e., a corpus linguistic way of analysing discourse – combines the ‚corpus driven' approach with classical ‚corpus based' analysis on the one hand, and quantitative calculations with qualitative reviewing of the deduced hypotheses on the other hand. In order to demonstrate the framework in practice, an analysis of a newspaper corpus is outlined. These examples prove the framework to be an important addition to discourse analysis.

Die linguistische Diskursanalyse nimmt sich Texte zum Untersuchungsgegenstand, um in einem hermeneutischen Prozess die Spuren von Diskursen darin zu finden. Damit ist eine solche Diskursanalyse schon immer korpuslinguistisch: Es wird ein Korpus von Texten definiert, das die Grundlage für die Analyse bildet.

Diese weitest mögliche Definition von ‚korpuslinguistischer Diskursanalyse' kann aber eingeengt werden, indem die Methoden, wie sie in den letzten Jahrzehnten im Rahmen der ‚elektronischen Korpuslinguistik' entwickelt wurden, fruchtbar gemacht werden. Diese Methoden wirken sich vornehmlich auf zwei Elemente aus: Auf den Umfang und die Gestalt des Korpus und die Art und Weise, wie dieses ausgewertet wird. Technisch ist es weitgehend unproblematisch geworden, mit Korpora zu arbeiten, deren

Anzahl Wörter im Millionenbereich liegt, und diese auch mit quantitativen Methoden auszuwerten, die über das Zählen von Wörtern weit hinaus gehen. Der vorliegende Beitrag möchte zweierlei aufzeigen:

1. Wie können korpuslinguistische Methoden für eine linguistische Diskursanalyse fruchtbar gemacht werden? Oder anders ausgedrückt: Wie können Fragestellungen der Diskursanalyse für die Möglichkeiten der Korpuslinguistik operationalisiert werden?
2. Umgekehrt erweisen sich korpuslinguistische Denkweisen aber auch als Möglichkeit, in der Diskursanalyse neue Wege zu gehen. Das bedeutet, dass ‚Korpuslinguistik' nicht bloß ein Werkzeug im Dienste eines Forschungsvorhabens ist, sondern auch Anreize für theoretische Überlegungen über Diskurse gibt.

1. Die Muster auf der sprachlichen Oberfläche

Folgt man den Foucault'schen Prinzipien der ‚Äusserlichkeit' und der ‚Spezifität' (Warnke 2007:17), bilden die sprachliche Oberfläche und deren diskursive Gebundenheit den Ausgangspunkt für eine linguistische Diskursanalyse. Doch die Palette der Phänomene, die auf der Textoberfläche analysiert werden können, ist breit. So rückt beispielsweise die Lexik in den Fokus, wenn Begriffsgeschichte geschrieben wird (Busse u. a. 1994). Oder das Untersuchungsinteresse gilt den Argumentationsfiguren und Topoi (Wengeler 2003) oder den Metaphern (Böke 1996), um nur einige Möglichkeiten zu nennen.[1]

Die Gemeinsamkeit dieser Phänomene ist ihr rekurrentes Auftreten im Sprachgebrauch. So treten z. B. Lexeme gehäuft zu bestimmten Zeitphasen und/oder in Verbindung mit bestimmten Themen, kommunikativen Situationen oder Textsorten auf. Gleiches gilt für Metaphern oder Argumentationsfiguren und Topoi. Diese Phänomene sind damit wichtige Bestandteile dessen, was den Sprachgebrauch einer bestimmten historischen Epoche und in bestimmten sozialen Zusammenhängen ausmacht. Die Analyse dieses Sprachgebrauchs in Abhängigkeit von spezifischen (synchron oder diachron) definierten Kontexten – also *typischen* Sprachgebrauchs – ist der Ausgangspunkt für diskursanalytische Fragestellungen.[2]

Typischer Sprachgebrauch kann operationalisiert werden als rekurrentes Auftreten von textuellen Einheiten in bestimmten Sprachausschnitten. Solche textuelle Einheiten können Morpheme sein, Lexeme oder aber

1 Vgl. auch Jung (1996:465) für eine knappe Übersicht.
2 Vgl. zum Vorschlag einer Sprachgebrauchsgeschichte Linke (2003).

komplexe Kombinationen von solchen Einheiten. Zudem können sie in Kombinationen mit ‚Slots' erscheinen, wie z. B. in *in der Nacht auf [Wochentag]*. ‚Typisch' ist dabei nicht einfach mit ‚häufig' gleichzusetzen, sondern meint immer ‚häufig in Bezug auf bestimmte Sprachausschnitte'. Ein Sprachgebrauch ist dann typisch, wenn er in einer bestimmten Art von Kontexten, z. B. in Artikeln der Auslandberichterstattung einer Zeitung X im Jahre 2008, überzufällig häufiger auftritt, als ein Jahr zuvor. Ich schlage vor, dieses Phänomen typischen Sprachgebrauchs ‚Muster' zu nennen.

Der Musterbegriff ist ein schillernder Begriff. Etymologisch gesehen stammt ‚Muster' vom Italienischen ‚mostra' ab, was mit ‚Probestück', bzw. ‚Ausstellung', ‚Auslage' übersetzt werden kann und seinerseits auf lat. ‚monstrare', ‚zeigen', zurückgeht (Kluge 1995, ‚Muster'). Ich verstehe im Kontext von Sprachgebrauch ‚Muster' in diesem Sinne als Vorbild oder Vorlage, und nicht in der zweiten Bedeutung, die z. B. im *gemusterten Tischtuch* steckt. Interessant beim Muster im Sinn von ‚Vorlage' ist die Tatsache, dass das Muster zwar ein Objekt ist, das als Vorbild für weitere Objekte diente, selber jedoch grundsätzlich gleicher Klasse ist wie die in der Folge sich daran orientierenden Objekte. Deshalb ist ein Objekt nie per se ein Muster, sondern kann nur in einer bestimmten Situation diese Funktion einnehmen. Außerhalb dieser Situation kann aber nicht mehr darüber entschieden werden, ob das Objekt diese Funktion je inne hatte – genau wie das (zumindest bei den meisten) sprachlichen Mustern ebenfalls der Fall ist: Wenn vom *Kampf gegen den Terrorismus* die Rede ist, kann ein ursprüngliches Muster, das Vorbild für diesen Ausdruck war, nicht mehr ausgemacht werden. Trotzdem werden in der Sprachproduktion immer wieder Instanzen des Ausdrucks *Kampf gegen den Terrorismus* musterbildend wirken und so zu typischem bzw. musterhaftem Sprachgebrauch führen. Mit ‚musterhaftem Sprachgebrauch' wird betont, dass anscheinend im untersuchten Sprachausschnitt immer wieder Instanzen einer bestimmten Phrase als Muster (als Vorbilder) für die Produktion weiterer Instanzen dienten. Im Nachhinein ist aber nicht mehr erkennbar, welche Instanzen je diese Musterfunktion übernahmen. Aber der Effekt dieser unzähligen Instanzen, die einerseits einem Muster folgten und andererseits Musterfunktion übernahmen, ist das Phänomen eines typischen, oder eben: musterhaften Sprachgebrauchs.

Hier kann eingewandt werden, dass für einen bestimmten Ausdruck sehr wohl oft ein Erstbeleg nachweisbar ist, dem wiederum Vorbildfunktion für weitere Verwendungen zugeschrieben werden kann. Das ist jedoch eine Ex-post-Interpretation, bei der dieser Verwendung eine Musterfunktion zugeschrieben wird. Im direkt der Erstverwendung anschließen-

den Sprachgebrauch wäre diese Musterfunktion wahrscheinlich nicht erkannt worden.

Ich möchte den Musterbegriff nicht in Bezug auf den Grad der Abstraktion einschränken. Die Instanzen, die Musterfunktion aufweisen, können konkrete Phrasen wie das oben Erwähnte *Kampf gegen den Terrorismus* sein, aber auch abstraktere Phrasen wie KAMPF GEGEN X.

Ich habe bereits erwähnt, dass diese Sprachgebrauchsmuster teilweise Phänomene sind, die in der linguistischen Diskursanalyse als Basis für begriffsgeschichtliche Beschreibungen verwendet werden, oder aber als (Teile von) Metaphern, Argumentationsfiguren oder Topoi gefasst werden können. Wo liegt denn der Gewinn, wenn auf der Ebene der Analyse zunächst mit einem umfassenderen (und dadurch auch ungenaueren) Konzept wie Sprachgebrauchsmustern gearbeitet wird?

Zu Sprachgebrauchsmustern zählen, neben den oben erwähnten Phänomenen, auch solche Muster, die in keine dieser Kategorien passen, da sie kaum einen spezifischen inhaltlichen Bezug aufweisen und eher im Sinne von Feilke (1996) als „idiomatische Prägungen" und/oder als „Routineformeln" (Burger 1998:50/52) bezeichnet werden können:

- in der Nacht auf gestern
- zum ersten Mal
- auf Grund der
- nach Angaben von/der

Wenn nachgewiesen werden kann, dass Muster dieser Art typisch für einen bestimmten Sprachausschnitt sind, zu einem bestimmten Zeitpunkt signifikant zu- oder abnehmen oder plötzlich in ganz bestimmten Kontexten auftauchen, können solche Muster ebenfalls hilfreich sein, um einen Diskurs zu beschreiben. Sie sind zwar weniger Indikatoren für die Wirkkräfte des Diskurses auf das *inhaltlich* Sagbare, dafür zeigen sie an, *wie* etwas gesagt werden muss. Solche Muster sind also Indikatoren für die „Sprechweise" oder die „diskursive Praxis" (Foucault 1981:275), die in einem Diskurs gilt.

Mit dem Auffinden und der Analyse von Sprachgebrauchsmustern sind also zwei Vorteile verbunden:

1. Es wird möglich, die eher inhaltliche Beschreibung von Diskursen zu ergänzen durch die Analyse der diskursiven Praxis, also die Frage „nach Gebrauchsweisen von Begriffen, den Ausformungen von Texten und den Routinen sprachlichen Handelns" (Linke 2003:40) zu stellen.
2. Das Auffinden von formal bestimmten Sprachgebrauchsmustern ist die algorithmisch einfacher operationalisierbare Aufgabe als das Auffinden von Konzepten wie Topos

> oder Metapher, die auf der Textoberfläche sehr unterschiedlich formalisiert werden können. Damit ebnet sich ein Weg, um mit korpuslinguistischer Hilfe und großen Datenmengen diskursanalytisch zu arbeiten.

Bevor ich im Detail zeigen möchte, wie mit korpuslinguistischen Methoden diskursanalytisch gearbeitet werden kann, gehe ich im Folgenden auf zwei wichtige Perspektiven der Korpuslinguistik ein, die auf das Verständnis von Diskursanalyse zurück wirken.

2. Korpuslinguistik: Corpus-based und corpus-driven

Die (elektronische) Korpuslinguistik zeichnet sich primär dadurch aus, dass sie Methoden entwickelt, um in großen Datenmengen rasch und gezielt Belege für sprachliche Phänomene zu finden. Für diskursanalytische Arbeiten ist damit auf den ersten Blick nicht viel gewonnen: Das zu untersuchende Korpus kann zwar besser verwaltet werden, und es ist komfortabler, Belege für eine bestimmte Wortform oder ein Lexem zu finden. Damit kann zwar korpusbasiert, ‚corpus-based', gearbeitet werden, doch das alleine ist kein großer Gewinn, überwiegt doch das Problem, bei der Analyse mit noch größeren Datenmengen umgehen zu müssen.

Auf den zweiten Blick jedoch wird das Potenzial korpuslinguistischer Forschung sichtbar:

> Sind Korpora nur Belegsammlungen oder Zettelkästen in elektronischer Form? Mitnichten! In entsprechender Größe […] und mit den entsprechenden Analysemethoden eröffnen sie eine eigene Perspektive in der linguistischen Forschung – die korpuslinguistische Perspektive. (Perkuhn/Belica 2006:2)

Das Zitat verweist auf ein in der Linguistik noch immer häufiges Missverständnis von Korpuslinguistik. Natürlich können Korpora auch als Nachschlagewerke benutzt werden, um zu überprüfen, ob sich darin ein bestimmtes sprachliches Phänomen findet. Doch gerade mit Hilfe der immer schneller werdenden Computer und der Verfügbarkeit von elektronischen Korpora ist für viele Fragestellungen eine andere Perspektive der Korpuslinguistik interessanter: Die ‚corpus-driven'-Perspektive:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories

> and classifications that are needed to answer a given research question. This is the corpus-driven approach. (Teubert 2005:4)

Dieser Zugang zeichnet sich also im Versuch aus, das Korpus als Datenbestand aufzufassen, in dem mit geeigneten Methoden Strukturen sichtbar gemacht werden, die erst im Nachhinein klassifiziert werden.[3] Für eine Analyse von Sprachgebrauchsmustern muss also ein Weg gefunden werden, in den Daten musterhaften Sprachgebrauch zu finden, ohne bereits vorher definieren zu müssen, aus welchen konkreten sprachlichen Elementen, also meist Lexemen oder auch morphosyntaktischen Strukturen, das Muster besteht. Weiter unten werde ich zeigen, wie das methodisch gelöst werden kann.

3. Repräsentativität und thematische Eingrenzung

Wenn mit automatischen Methoden große Textmengen korpuslinguistisch analysiert werden können, stellt sich die Frage nach der Korpusdefinition neu. Nach welchen Kriterien sollen die Texte ausgewählt werden?

Die Diskursanalyse möchte „Spiele von Beziehungen" beschreiben, und dabei alle

> Beziehungen der Aussagen untereinander (selbst wenn diese Beziehungen dem Bewusstsein des Autors entgehen; selbst wenn es sich um Aussagen handelt, die nicht den gleichen Autor haben; selbst wenn diese Autoren einander nicht kennen); Beziehungen zwischen so aufgestellten Gruppen von Aussagen (selbst wenn diese Gruppen nicht die gleichen Gebiete oder benachbarte Gebiete treffen; selbst wenn sie nicht das gleiche formale Niveau haben; selbst wenn sie nicht der Ort bestimmbaren Austausches sind); Beziehungen zwischen Aussagen oder Gruppen von Aussagen oder Ereignissen einer ganz anderen (technischen, ökonomischen, sozialen, politischen) Ordnung (Foucault 1981:45)

analysieren. Dabei sollen die diskursiven Einheiten kontrolliert in „andere Einheiten" gruppiert werden, Mengen, die „nicht arbiträr […], indessen aber unsichtbar geblieben wären" (Foucault 1981:45).
Dies bedeutet:

1. Es werden Aussagen analysiert, von denen angenommen werden muss, dass sie nicht nur in Form von abgeschlossenen Texten, sondern auch in anderen textuellen Größen auffindbar sind.

---

3 Das corpus-driven-Paradigma ist nicht neu. Bei Sinclair (1991) bereits angedacht, wird es bei Tognini-Bonelli (2001) mit dem Terminus ‚corpus-driven Linguistics' (CDL) explizit gemacht.

2. Die ‚Spiele von Beziehungen' machen nicht an den Grenzen von eindimensional (z. B. thematisch) definierten textuellen Einheiten Halt.

Auf den ersten Punkt möchte ich an dieser Stelle nicht weiter eingehen und verweise stattdessen auf Jung (1996:460) und Spitzmüller (2005:47), die auch von einer Untersuchungseinheit ausgehen, die unterhalb des Textes liegt. Grundsätzlich kann dieser Aspekt auch nach der Zusammenstellung eines Korpus, das aus Texten besteht, in der Analyse berücksichtigt werden.

Für die Korpuszusammenstellung bedeutender ist jedoch die Frage nach dem Auswahlkriterium von Texten für das Korpus, die mit dem zweiten Punkt oben aufgeworfen wird. Nach Busse/Teubert (1994:14) sollen die ausgewählten Texte repräsentativ „hinsichtlich eines jeweils als Untersuchungsleitfaden gewählten Inhaltsaspekts sein". Repräsentativität in der Diskursanalyse sei „vor allem ein inhaltliches (semantisches) Problem" und im Vordergrund stünden „inhaltlich begründbare Relevanzkriterien" (Busse/Teubert 1994:14). Diese heuristische Definition von Repräsentativität ist nicht nur unnötig, sondern auch kritisch, wenn das Ziel einer Diskursanalyse ist, auch verborgene und unauffällige Strukturen aufzuzeigen. Relevanzempfinden ist ein Ergebnis diskursiver Formationen. Dieses als Kriterium zu wählen, bedeutet, die offensichtlichen Strukturen des Diskurses zu replizieren.

Dass es anders geht, zeigt Wengeler (2003:294), der für sein Korpus aus Zeitungstexten zum Einwanderungsdiskurs zwar nur Texte berücksichtigt, die das Thema in irgendeiner Form behandeln, jedoch alle Texte des Untersuchungszeitraums aufnimmt, auch wenn das Thema nur gestreift wird. Er nimmt keine Unterscheidung in Leittexte oder besonders relevante Texte vor. Es wurden „ausführliche Hintergrundreportagen zum Thema Einwanderung ebenso berücksichtigt wie kurze Berichte über punktuelle Ereignisse oder einzelne Stellungnahmen von Politikern" (Wengeler 2003:295). Auch Gardt (2007:43) betont, dass eine Diskursanalyse

> sich ja gerade nicht auf die klassischen, großen Texte beschränken [kann], sondern […] die ganze Vielfalt der Äußerungen, die ein Thema konstituieren, zu erfassen versuchen [muss] mit entsprechenden Konsequenzen für die Textmenge. (Gardt 2007:43)

Eine induktive Analyse, bei der Sprachgebrauchsmuster im Zentrum stehen, bietet aber zusätzlich die Chance, Diskurse nicht nur thematisch zu definieren, sondern als Cluster von ähnlichem Sprachgebrauch. Dazu wird in einem großzügig gewählten Ausgangskorpus (z. B. alle Zeitungsartikel

aus allen Ressorts eines Jahrgangs) nach typischen Sprachgebrauchsmustern gefahndet. Mit der weiteren Analyse können die für die Untersuchung relevanten Muster als Teilmenge aller typischen Sprachgebrauchsmuster definiert werden. Dabei können die Muster nicht nur nach inhaltlichen, sondern nach beliebigen anderen Kriterien aus den bereits berechneten Mustern ausgewählt werden. Beispielsweise können Muster gewählt werden, denen man eine argumentative Funktion zuschreibt, wie die folgenden:[4]

- ist zu hoffen, dass …
- ein Dorn im Auge
- an der Zeit, dass …
- je länger je mehr …
- auf Teufel komm raus …

In einem nächsten Schritt wird das Korpus für die detaillierte Analyse auf jene Textpassagen reduziert, die diese Muster enthalten. Damit wird das Korpus für qualitative Analysen besser handhabbar, beruht aber trotzdem auf einer breiten empirischen Basis.

Um statistischen Gütekriterien zu genügen, muss ein repräsentatives Korpus als zufällige Stichprobe aus einer definierten Grundgesamtheit gezogen werden. Bisher sprachen oft forschungspraktische Gründe gegen ein solches Verfahren. Für immer mehr Datenquellen gibt es aber diesbezüglich kaum mehr Probleme, so ist es beispielsweise relativ einfach möglich, aus zehn Jahrgängen einer Tageszeitung eine Zufallsstichprobe von mehreren Tausend Artikeln zu ziehen und diese korpuslinguistisch zu verarbeiten.[5] Wird dann in einem ersten Schritt zunächst corpus-driven musterhafter Sprachgebrauch darin berechnet, kann die Datenmenge für nachgelagerte qualitative Analysen merklich verringert werden, ohne die empirische Basis zu verlieren, um statistisch gesicherte Aussagen machen zu können.

4 Diese Sprachgebrauchsmuster stammen aus der corpus-driven-Analyse von Leserbriefen einer Tageszeitung (vgl. Bubenhofer 2008).

5 Natürlich gibt es noch immer genügend sprachliche Daten, die nicht elektronisch in genügend umfangreichen und gut zugänglichen Mengen vorliegen. Man denke an historische Texte, mündliche Sprache und visuelle Daten (vgl. die Beiträge von Kersten Roth und Stefan Meier in diesem Band).

4. Methoden der Korpuslinguistik

4.1 Berechnung von Kollokationen

Welche Techniken stehen nun zur Verfügung, um in Korpusdaten nach Sprachgebrauchsmustern zu suchen? Der Kernbegriff in diesem Zusammenhang ist ‚Kollokation'. Er geht auf Firth (1957:194) zurück und wird in der Korpuslinguistik heute als statistisch signifikantes gemeinsames Auftreten von Wörtern (je nach Definition von Lexemen oder Wortformen) definiert.[6] Mit unterschiedlichen statistischen Maßen[7] wird berechnet, welche Wörter häufiger als erwartet in einer definierten Distanz zu einem gegebenen Wort auftreten.

Der Mangel solcher Berechnungen von Kollokationen ist die Beschränkung auf Zweiwort-Verbindungen und die nötige Festlegung von Wörtern als Ausgangspunkte für die Berechnung der Kollokationen. Letzteres würde gerade nicht einer corpus-driven-Perspektive entsprechen. Um diese Limitierungen zu vermeiden, können in einem Korpus systematisch für jedes auftretende Wort die Kollokationen berechnet werden, und statt bloß Zweiwort-Verbindungen können die Gruppen um beliebig viele Wörter zu Mehrworteinheiten (auch: ‚Multi Word Units') erweitert werden.[8] Die Berechnung solcher Mehrworteinheiten ist eine simple Auszählarbeit, die spezialisierte Programme automatisch erledigen.[9]

Das Resultat solcher Berechnungen sind Listen, wie in Tabelle 1 in einem Auszug dargestellt. Sie basieren auf einem Korpus (im Folgenden ‚NZZ-Korpus') aus knapp 45.000 zufällig ausgewählten Artikeln der Neuen Zürcher Zeitung der Periode von 1995 bis 2005.[10] In dieser Liste wurden Mehrworteinheiten mit einer Länge von drei Wörtern berechnet, wobei die Distanz zwischen dem jeweils ersten und letzten Wort im

---

6 Vgl. für die mitunter schwierige Unterscheidung von ‚Kookkurrenzen' und ‚Kollokationen' auch Manning/Schütze (2000:151), Lemnitzer/Zinsmeister (2006:147) und Steyer (2004: 96). Ich werde in der vorliegenden Arbeit diese Unterscheidung ignorieren und nur den Begriff ‚Kollokationen' verwenden.

7 Vgl. für eine Diskussion dieser Maße Evert (2005) oder Manning/Schütze (2000).

8 Für Wortgruppen, die mehr als drei Wörter enthalten, wird die Anwendung von statistischen Maßen zur Berechnung der Kollokationsstärke sehr kompliziert. Deshalb wird in solchen Fällen oft auf die Verwendung solcher Maße verzichtet und es werden stattdessen nur die Frequenzen (in Relation zur Korpusgröße) berücksichtigt. Das reicht meist aus, da die Menge der rekurrenten Mehrworteinheiten mit steigender Länge markant abnimmt.

9 Für Hinweise zu solchen Programmen und weiteren Details der Berechnung vgl. Bubenhofer (2006).

10 Im Detail: Das Korpus besteht aus einer Zufallsstichprobe von 44.843 Artikeln (27.946.381 Tokens) aus der ‚Neuen Zürcher Zeitung' aus dem Zeitraum vom 3. Januar 1995 bis 31. Dezember 2005. Die Stichprobe macht damit etwa 6% der Artikel des gesamten Publikationsumfangs und damit der Grundgesamtheit aus. Es wurden alle redaktionellen Artikel aus allen Ressorts berücksichtigt.

Tab. 1: Ausschnitt aus einer Liste von berechneten Dreiwort-Einheiten in der Auslandberichterstattung der ‚Neuen Zürcher Zeitung' geordnet nach Signifikanz (ausgedrückt in Rängen).

Mehrworteinheit	Rang	Mehrworteinheit	Rang
in der Provinz	137	Zeitung date 29	172
hat am Montag	138	war in der	173
gegen p Neue	139	gegen in der	174
in der Vergangenheit	140	um in der	175
in der Woche	141	in der einer	176
Zeitung date 24	142	werden in der	177
in der nur	143	Regierung in der	178
in der Ukraine	144	Ausland Press in	179
Zeitung date 09	145	S Ausland Gaupp	180
in der haben	146	in der Nacht	181
aber in der	147	in der Armee	182
in der noch	148	sind in der	183
hat am Donnerstag	149	in der sei	184
in der sind	150	in der durch	185
S Ausland Kocher	151	in der amerikanischen	186
des p Zuercher	152	in der Regierung	187
in der über	153	Zeitung date 13	188
in der vor	154	in der zur	189
hatte in der	155	in der einen	190
in der einem	156	in der Hand	191
in der Öffentlichkeit	157	Ausland dpa Deutsche	192
2 Ausland der	158	S Ausland ChM	193
Ausland Associated in	159	in der Grenze	194
in der ersten	160	in der deutschen	195
in der war	161	zum in der	196
Das in der	162	die vor allem	197
Ausland reu in	163	zur in der	198
in der aus	164	in der sein	199
in der Lage	165	nur in der	200
in der hat	166	der fuer die	201
einen in der	167	in der seit	202
in der Geschichte	168	in der Zeit	203
in der werden	169	in der Nato	204
in der Schweiz	170	noch in der	205
in der um	171	in der eines	206

Korpus mit maximal zehn Wörtern festgelegt wurde.[11] Das heißt, in den Belegen zu diesen Wortgruppen könnten zwischen den Wörtern der Gruppe jeweils noch weitere Wörter stehen.[12]

11 Die Liste ist ungefiltert und enthält deshalb auch Einzelbuchstaben, die im Text Abkürzungen darstellen wie *S* für Seite, Zahlen, sowie Metainformationen des Datenbanksystems wie *p* für page.

12 Für alle in diesem Beitrag dargestellten Mehrworteinheiten galten bei der Berechnung die folgenden Parameter: Anzahl der Wörter = 3; maximale Distanz zwischen erstem und

4.2 Typische Kollokationen eruieren

Listen dieser Art sind noch nicht von großem Wert, da sie sehr lang sind und erst Sprachgebrauchsmuster zeigen, die zwar eine häufige Wortkombination darstellen, jedoch nicht unbedingt typisch für das Korpus sind. Um die Typik zu berechnen, muss diese Liste mit einer auf gleiche Weise generierten Liste eines Referenzkorpus verglichen werden. Die Wahl von Korpus und Referenzkorpus richtet sich nach den Untersuchungsinteressen, wobei die Grundidee darin liegt, durch die Kontrastierung von ähnlichen Korpora, die für das eigentlich zu untersuchende Korpus typischen Sprachgebrauchsmuster zu finden. Beispielsweise könnte ein Korpus aus Artikeln des Ausland-Ressorts einer Zeitung mit den Artikeln aus dem Inland-Ressort der gleichen Zeitung verglichen werden. Oder in diachroner Perspektive könnten die Artikel aus dem Ausland-Ressort eines Jahrgangs mit den Artikeln eines älteren Jahrgangs kontrastiert werden. Aber auch thematische Eingrenzungen sind denkbar: So könnten Artikel zum Thema ‚Terrorismus' aus der Zeit 2000, 2002 und 2004 miteinander verglichen werden.

Beim Vergleich dieser Listen können im einfachsten Fall jene Mehrworteinheiten als typisch für das Korpus angesehen werden, die im Referenzkorpus nicht vorkommen. Differenzierter ist jedoch ein Vergleich, bei dem gemessen wird, wie signifikant die Differenz der unterschiedlichen Frequenzen einer Mehrworteinheit in den beiden Korpora ist. Tabellen 2 und 3 zeigen jeweils den Beginn von Listen von Mehrworteinheiten, die für die Auslandberichterstattung in Artikeln der Neuen Zürcher Zeitung im Zeitraum 1995–1999 bzw. 2000–2005 im Vergleich zur jeweils anderen Zeitperiode typisch sind. Die Mehrworteinheiten sind nach ihrer Signifikanz geordnet.

Für die Berechnung der Signifikanz können unterschiedliche Maße verwendet werden. Basis der Berechnung ist eine ‚Kontingenztafel', in der die Verteilung der Frequenzen in den beiden Korpora aufgeführt werden (vgl. Tabelle 4). Rayson/Garside (2000) und Kilgarriff (2001:105) schlagen den ‚Log-likelihood'-Test vor, um die Signifikanz der Verteilung in zwei Korpora zu messen. Der Log-likelihood-Koeffizient G^2 berechnet sich wie folgt:[13]

---

letztem Wort = 10 Wörter; minimale Frequenz im Korpus = 3. Die Listen wurden mit der Software ‚Ngram Statistic Package' (NSP) berechnet (vgl. Banerjee/Pedersen 2003).

13 Selbstverständlich gibt es eine Reihe von Programmen, die das komfortabel automatisch erledigen, so z. B. Paul Raysons ‚log-likelihood calculator', der online verwendet werden kann: *http://ucrel.lancs.ac.uk/llwizard.html*. Vgl. für weitergehende Hinweise auch Bubenhofer (2006).

$$G^2 = 2\,(A \log A + B \log B + C \log C + D \log D - (A + B) \log (A + B) - (A + C) \log (A + C) - (B + D) \log (B + D) - (C + D) \log (C + D) + (A + B + C + D) \log (A + B + C + D))$$

Tab. 2: Dreiwort-Einheiten in der Auslandberichterstattung der ‚Neuen Zürcher Zeitung', die typisch sind für die Periode 1995–1999 im Vergleich zu 2000–2005.

G^2	**Mehrworteinheit**	**95–99** #/Mio.	**00–05** #/Mio.
106,58	der bosnischen Serben	4,13	0
60,24	in der Slowakei	2,33	0
57,15	der der bosnischen	2,21	0
57,15	die bosnischen Serben	2,21	0
55,61	am abend in	2,15	0
55,61	Human Rights Watch	2,15	0
55,61	den und nicht	2,15	0
54,06	der der Menschenrechte	2,09	0
46,34	der der Ukraine	1,8	0
46,34	in einem Interview	1,8	0
43,25	die der Menschenrechte	1,68	0
41,71	Aung San Suu	1,62	0
39,54	der Serbischen Republik	3,77	0,77
37,07	am Montag abend	1,44	0
37,07	fuer Sicherheit und	1,44	0
37,07	die Rechte der	1,44	0
35,53	Aung Suu Kyi	1,38	0
33,98	der achtziger Jahre	1,32	0
32,44	Indien und Pakistan	1,26	0
30,89	Scud B 1000	1,2	0
29,35	nach dem Ende	1,14	0
27,8	der besetzten Gebiete	1,08	0
27,8	in Serbischen Republik	1,08	0
27,8	der Kurdischen Arbeiterpartei	1,08	0
27,8	der in Sarajewo	1,08	0
27,76	der in Kosovo	4,25	1,39
27,48	der in Bosnien	2,57	0,51
26,26	in Phnom Penh	1,02	0
26,26	die der bosnischen	1,02	0
26,26	die der ETA	1,02	0
26,26	auf der Basis	1,02	0
24,72	Aufhebung der Sanktionen	0,96	0
24,72	Karadzic und Mladic	0,96	0
23,17	russische Praesident Jelzin	0,9	0
21,63	gegen Praesident Clinton	0,84	0
20,08	Kim Dae Jung	0,78	0
8,9	Beziehungen zwischen und	1,98	0,82

Tab. 3: Dreiwort-Einheiten in der Auslandberichterstattung der ‚Neuen Zürcher Zeitung', die typisch sind für die Periode 2000–2005 im Vergleich zu 1995–1999.

G^2	Mehrworteinheit	95–99 #/Mio.	00–05 #/Mio.
-128,91	die im Irak	0	5,34
-109,08	gegen den Terrorismus	0	4,52
-79,33	in die USA	0	3,29
-69,41	hiess es in	0	2,88
-68,17	mit den USA	0	2,83
-67,79	der im Irak	0,9	5,65
-65,69	in den Irak	0	2,72
-59,5	im Irak die	0	2,47
-58,26	vor Jahren der	0	2,41
-58,26	und die USA	0	2,41
-58,26	der der politischen	0	2,41
-55,78	Anschlag auf in	0	2,31
-50,82	die in Afghanistan	0	2,11
-50,82	Usama bin Ladin	0	2,11
-47,1	und im Irak	0	1,95
-44,62	Kampf gegen Terrorismus	0	1,85
-44,62	die EU und	0	1,85
-44,62	die den Irak	0	1,85
-44,62	in besetzten Gebieten	0	1,85
-40,9	Abzug aus dem	0	1,7
-39,66	der israelischen Armee	0	1,64
-39,66	getoetet und verletzt	0	1,64
-39,66	eine Loesung des	0	1,64
-37,96	im Kampf gegen	1,26	4,73
-37,19	der im Gazastreifen	0	1,54
-37,19	den gegen den	0	1,54
-35,95	Krieg gegen den	0	1,49
-35,95	fuer den Irak	0	1,49
-35,95	die USA in	0	1,49
-34,71	Tote bei in	0	1,44
-34,71	Kampf gegen und	0	1,44
-34,71	der Kampf gegen	0	1,44
-34,71	im Irak und	0	1,44
-34,71	Personen ums Leben	0	1,44
-32,23	dem Ende der	0	1,34
-32,23	die Zukunft der	0	1,34
-32,23	den neunziger Jahren	0	1,34
-30,99	der im Osten	0	1,28
-29,75	von Praesident Bush	0	1,23
-28,51	eine Mehrheit der	0	1,18
-28,51	die im Gazastreifen	0	1,18
-27,57	in den Gebieten	1,62	4,68
-27,27	zu Jahren verurteilt	0	1,13
-27,27	die Verantwortung fuer	0	1,13
-27,27	Personen verletzt worden	0	1,13
-23,55	Truppen aus dem	0	0,98
-23,55	zum ersten Mal	0	0,98

Tab. 4: Kontingenztafel für die Frequenzen einer Mehrworteinheit (MWE) X in den Korpora A und B.

	Korpus A	*Korpus B*	Total
Frequenz MWE X	A	B	A+B
Freq. aller anderen MWE	C	D	C+D
Total	A+C	B+D	N

Die Variablen A, B, C und D stehen für die entsprechenden Felder der Kontingenztafel (vgl. Tabelle 4). Beim Signifikanztest werden nun die beobachteten Werte mit den bei zufälliger Verteilung erwarteten Frequenzen verglichen. Je höher G^2 ist, desto signifikanter ist die Verteilung der Frequenzen. Um mit dem Log-likelihood-Koeffizienten gleichzeitig anzuzeigen, dass die Mehrworteinheit in Korpus B relativ häufiger vorkommt als in Korpus A, kann ein Minuszeichen vor G^2 gesetzt werden. Im gegenteiligen Fall, der höheren relativen Frequenz in Korpus A, bleibt der Wert positiv. Für jede Mehrworteinheit kann nun G^2 berechnet und die Liste danach geordnet werden (vgl. Tabellen 2 und 3).[14]

Diese Listen von Mehrworteinheiten sind nun Ausgangspunkt für die weitere Analyse, bei der die einzelnen Mehrworteinheiten interpretiert und kategorisiert werden. Mit diesen Schritten muss entschieden werden, ob es sich bei diesen Mehrworteinheiten, die zunächst bloß statistisch auffällige Wortkombinationen sind, um Sprachgebrauchsmuster handelt. Methodisch gesehen findet mit der Analyse der berechneten Mehrworteinheiten ein Wechsel der Perspektive von corpus-driven nach corpus-based statt.

14 Der Wert von *G2* ließe sich anschließend darauf hin prüfen, ob die Verteilung, gemessen an einem bestimmten Signifikanzniveau, signifikant ist oder nicht. Es würde also getestet, ob die Nullhypothese H0, dass die Verteilung zufällig zustande gekommen ist, mit genügend großer Sicherheit verworfen werden kann. Allerdings sind solche Signifikanztests bei Korpusdaten kritisch, da nicht ohne weiteres von einer zufälligen Verteilung der Wörter im Korpus ausgegangen werden kann. Deshalb schlägt z. B. Kilgarriff (2001:114) vor, solche Signifikanztests nur im oben dargestellten Sinn zu verwenden, um die Grade der Signifikanz der Wortvorkommen untereinander zu vergleichen.

5. Korpuslinguistische Diskursanalyse: Ein Verfahren zwischen corpus-driven und corpus-based

Eine korpuslinguistische Diskursanalyse, die corpus-driven begonnen wurde, wird erst dann hilfreich, wenn sie corpus-based ergänzt wird. Die vorliegenden Listen von Mehrworteinheiten enthalten spezifisches Wortmaterial, dessen Verwendung im Korpus genauer überprüft werden muss. Dafür stehen die ‚klassischen' Verfahren korpuslinguistischer Analysemethoden zur Verfügung: Sichtung der Belege, Überprüfung der Distribution, Kategorisierungen.[15] Es versteht sich von selbst, dass bei diesen Korpusrecherchen die ursprünglich generierten Mehrworteinheiten modifiziert und ergänzt werden müssen. Sie dienen sozusagen der Generierung von Hypothesen, die corpus-based überprüft werden müssen. So führt beispielsweise die für eine bestimmte Zeitperiode in einer Zeitung typische Mehrworteinheit *zum ersten Mal* zu Recherchen nach *zum/das erste(n/s) Mal*, *erstmals* etc., um abschätzen zu können, wie stark die Mehrworteinheit zu einem Sprachgebrauchsmuster abstrahiert werden kann.[16]

Im Kontext einer Diskursanalyse wird deutlich, dass der Wechsel zwischen corpus-driven und corpus-based nicht bloß einmal vorgenommen wird, sondern dass ein zirkulärer Prozess entsteht, der immer wieder diese Perspektiven durchläuft. Abbildung 1 zeigt diesen Prozess im Überblick: In einem ersten Schritt muss ein Korpus in Kombination mit einem oder mehreren Referenzkorpora definiert werden, wobei auch Teile des Untersuchungskorpus als Referenzkorpus dienen können. Aus dem Korpus und den Referenzkorpora werden corpus-driven Listen von Mehrworteinheiten berechnet. Durch Kontrastierungen der Listen untereinander können die für bestimmte Teilkorpora typischen Mehrworteinheiten berechnet werden.

---

15 Ich verzichte an dieser Stelle auf detailliertere Beschreibungen dieser Techniken und verweise stattdessen auf korpuslinguistische Literatur zur Einführung: Lemnitzer/Zinsmeister (2006), Scherer (2006), Sinclair (1991), Tognini-Bonelli (2001) und Bubenhofer (2006).

16 Ein nicht zu vernachlässigender Aspekt ist die Frage nach der Lemmatisierung der Daten. Es ist möglich, mit großer Zuverlässigkeit ein Korpus automatisch mit Wortarten und Lemmaformen zu annotieren und die Berechnung von Mehrworteinheiten auf Basis der Lemmata statt der laufenden Textwörter vorzunehmen. Allerdings liegen in unterschiedlichen Flexionsformen oft semantische Differenzierungen, die mit der Lemmatisierung verloren gingen, so wenn z. B. *im letzten Jahr* und *in den letzten Jahren* zu IN [DEF. ART.] LETZTES JAHR überführt und damit bei der Zählung nicht mehr differenziert würde. Während corpus-driven diese Differenzierung aber erwünscht ist, muss später corpus-based überprüft werden, ob das Muster zu dieser lemmatisierten Form generalisiert werden kann, oder ob tatsächlich semantische Differenzen vorliegen. Vgl. zu diesem Problem auch Tognini-Bonelli (2001:92), die explizit von einem Informations*verlust* spricht, zu dem es bei der Lemmatisierung komme.

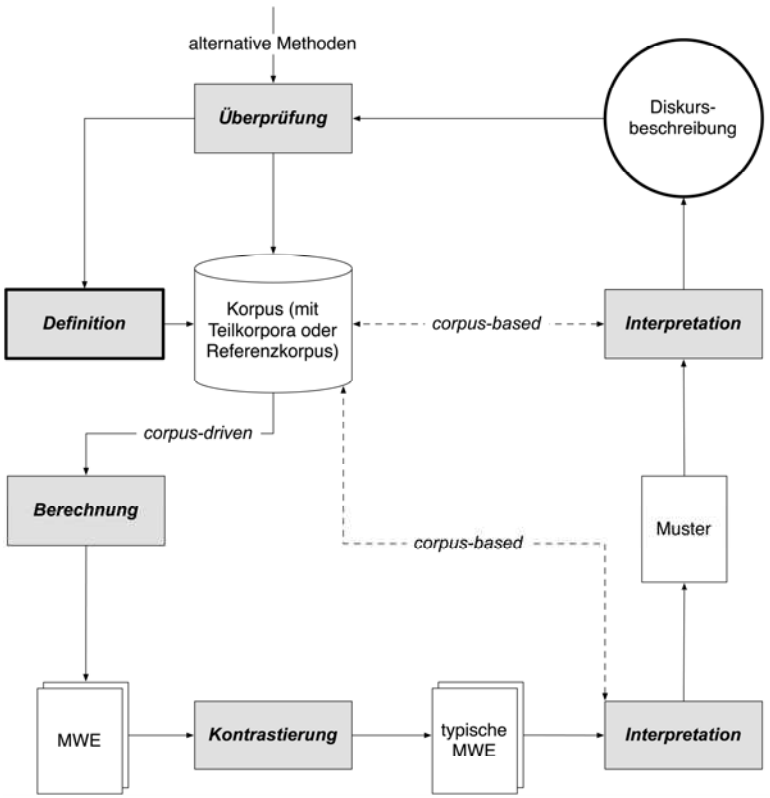

Abb. 1: Die korpuslinguistische Diskursanalyse im Überblick

Nun erfolgt unter corpus-based-Rückgriffen in das Korpus die Interpretation der Mehrworteinheiten, um aus ihnen abstraktere Sprachgebrauchsmuster abzuleiten. Die weitere Analyse der Verwendung dieser Sprachgebrauchsmuster – ebenfalls corpus-based – führt zu einer Diskursbeschreibung. Doch auch jetzt ist der Prozess nicht abgeschlossen: Die Diskursbeschreibung muss aufgrund der Korpusdaten auf ihre Plausibilität hin geprüft werden. Dabei ist es sinnvoll, die Daten auch mit alternativen Methoden auszuwerten, um die Diskursbeschreibung einer erweiterten Prüfung zu unterziehen. Je nach Resultat dieser Überprüfungen muss der Prozess mit veränderten Parametern erneut durchlaufen werden, um zu einer korrigierten Diskursbeschreibung zu gelangen.

Neben dem Wechsel zwischen corpus-driven und corpus-based bewegt sich dieser Prozess auch zwischen quantitativen und qualitativen Me-

thoden. Die Entscheidung darüber, welche der berechneten Mehrworteinheiten weiter verfolgt werden, die Gewichtung von Belegen im Korpus, um die Sprachgebrauchsmuster abzuleiten und letztlich die weitere Abstrahierung der Befunde zu einer Diskursbeschreibung, sind qualitativ-interpretative Akte. Allerdings beruhen sie auf einer empirischen Basis.

6. Vom Sprachgebrauchsmuster zum Diskurs: Beispiele

Man kann sich leicht vorstellen, wie bei einem spezifischen inhaltlichen diskursanalytischen Interesse die Sprachgebrauchsmuster thematisch gefiltert und auf ihre diskursive Funktion hin analysiert werden können. So wäre ein größerer Teil der in Tabelle 3 aufgeführten Mehrworteinheiten, die *Terror*, *Irak*, *USA* etc. enthalten, sowie Einheiten wie *im Kampf gegen*, *Anschlag auf … in …* oder *… Personen ums Leben* eine Basis, um den Terrorismus-Diskurs genauer unter die Lupe zu nehmen. Es wäre nach alternativen Ausdrücken und deren Verwendung zu fragen, nach den Kontexten dieser Muster oder danach, ob sie Indikatoren für Metaphern oder Topoi sind.[17]

Einen etwas anderen Fokus setzt die Frage nach der Verwendung von *Kampf gegen*, motiviert durch die berechneten Mehrworteinheiten *Kampf gegen Terrorismus*, *im Kampf gegen* etc. aus Tabelle 3. Es scheint zwar plausibel, dass *Kampf gegen Terrorismus* ein Ausdruck ist, der typisch für die Zeit nach dem 11. September 2001 ist, wie Tabelle 5 zeigt.[18] Doch zu fragen ist allgemeiner nach dem Gebrauch des Musters KAMPF GEGEN X und den Füllungen des Slots X.

Tabelle 6 zeigt die Frequenzen von *Kampf gegen* im NZZ-Korpus.[19] Die Verteilung ist nicht signifikant, die Verwendung also über die Jahre ungefähr ähnlich. Abbildung 2 zeigt die Füllungen von X im Muster KAMPF GEGEN X für die Perioden vor und nach dem 11. September 2001. Dafür wurden alle Belege (842) von *Kampf gegen* im Korpus untersucht, indem automatisch die rechten Kollokatoren des Suchausdrucks extrahiert wurden. Dabei gilt als Kollokator alles Wortmaterial bis und mit dem ersten

17 Vgl. für eine methodisch ähnlich gelagerte Untersuchung zu Argumentationsfiguren in Leserbriefen Bubenhofer (2008).

18 Um die Signifikanz der Verteilung der Frequenzen über die Jahre zu testen, wurde der $\chi 2$-Test verwendet. Vgl. Belica (1996) und Manning/Schütze (2000:169).

19 Vgl. für die Korpusdefinition Seite 423.

Tab. 5: χ^2-Statistik (Einheit: Artikel) für die Mehrworteinheit *Kampf(s/es) gegen (den) Terror** im NZZ-Korpus. In Klammern: Frequenzen in Prozent zum Total

Jahr	***Kampf(s/es) gegen (den) Terror****	***alle anderen MWE***	**Total**
1995	2 (0,05%)	3664	3666
1996	2 (0,05%)	3709	3711
1997	0 (0%)	3827	3827
1998	3 (0,08%)	3613	3616
1999	2 (0,05%)	4246	4248
2000	0 (0%)	4496	4496
2001	9 (0,21%)	4364	4373
2002	17 (0,39%)	4368	4385
2003	11 (0,26%)	4224	4235
2004	15 (0,36%)	4140	4155
2005	13 (0,31%)	4118	4131
Total	74	44769	44843
χ^2 =55,568, df=10, p<0,001 (signifikant)			

Tab. 6: χ^2-Statistik (Einheit: Artikel) für die Mehrworteinheit *Kampf gegen* im NZZ-Korpus. In Klammern: Frequenzen in Prozent zum Total

Jahr	*Kampf(s/es) gegen*	*alle anderen MWE*	Total
1995	60 (1,64%)	3606	3666
1996	58 (1,56%)	3653	3711
1997	55 (1,44%)	3772	3827
1998	58 (1,6%)	3558	3616
1999	43 (1,01%)	4205	4248
2000	60 (1,33%)	4436	4496
2001	80 (1,83%)	4293	4373
2002	67 (1,53%)	4318	4385
2003	70 (1,65%)	4165	4235
2004	71 (1,71%)	4084	4155
2005	64 (1,55%)	4067	4131
Total	686	44157	44843
χ^2 =13,274, df=10, p>0,10 (nicht signifikant)			

groß geschriebenen Wort.[20] Damit wird sowohl *(Kampf gegen das) organisierte Verbrechen* als auch *(Kampf gegen den) Terror* gefunden. Es wurden nur die Kollokatoren berücksichtigt, die in der Periode vor oder nach dem 11. September 2001 mindestens dreimal erschienen. Teilweise wurden Kollokatoren in unterschiedlichen Schreibvarianten, Flexionsformen oder mit Ergänzungen zu Gruppen zusammengefasst. Solche Zusammenfassungen sind durch die Verwendung des Sterns (*) ersichtlich, der für beliebig viele Zeichen steht.

20 Es wurde mittels regulären Ausdrücken gesucht. Das entsprechende Suchmuster lautete: (.+? [[:upper:]].+? \s).

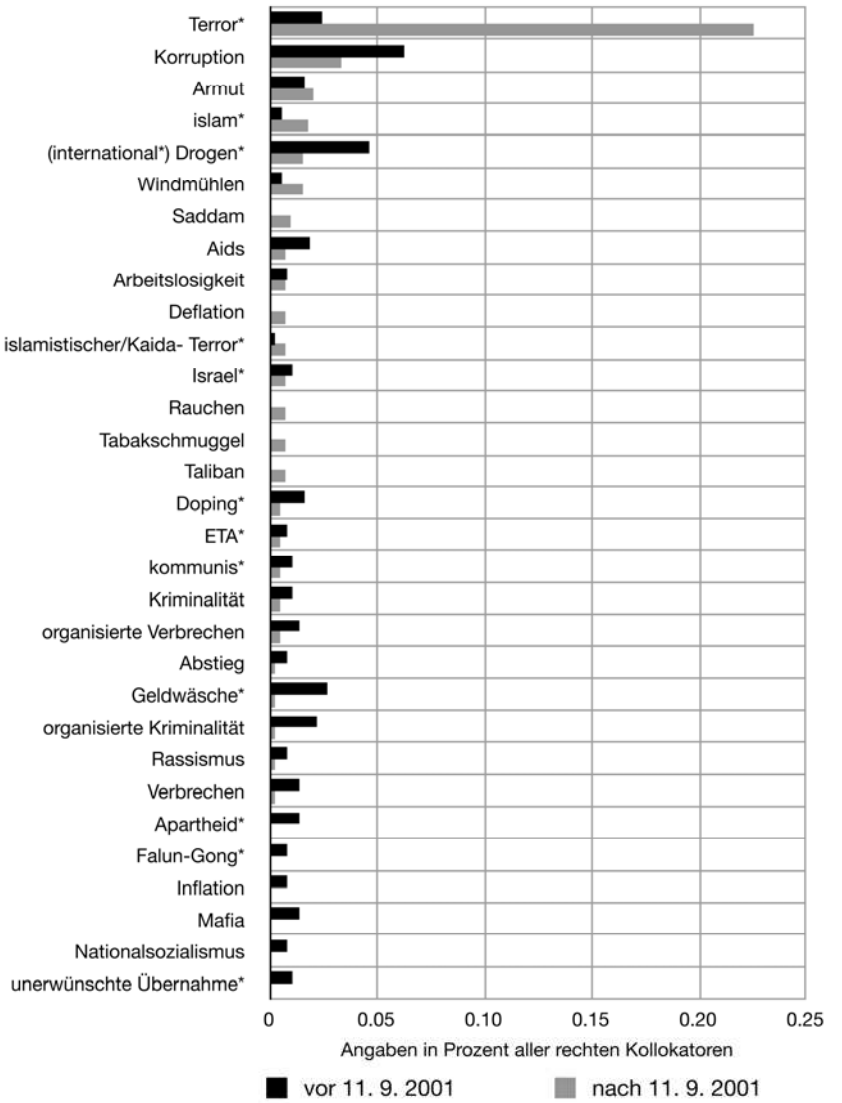

Abb. 2: Die Füllungen von X im Muster KAMPF GEGEN [ARTIKEL] X mit Mindestfrequenz 3. Vor 11.9.2001 n = 368, nach 11.9.2001 n = 394. Details zur Berechnung können dem Text entnommen werden

Die Übersicht der rechten Kollokatoren in Abbildung 2 zeigt die inhaltlichen Gewichtungen an, die für die beiden Zeitperioden typisch sind. Während in der ersten Zeitperiode der *Kampf* der *Korruption*, den *Drogen*, der *Geldwäsche* und der *Kriminalität* galt, wird das Feld in der zweiten Periode vom *Terror* dominiert, gefolgt von *Korruption*, *Armut* und Wörter mit dem Morphem *islam-*.

Aus diskursanalytischer Sicht würde nun besonders interessieren, ob und wie Diskurse die Sprechweise verändern. Im Beispiel von KAMPF GEGEN X wäre also zu fragen:

1. Gibt es für bestimmte X zu bestimmten Zeiten oder Kontexten Beschränkungen, sie im Muster KAMPF GEGEN X zu verwenden?
2. Welche alternativen Ausdrucksmöglichkeiten zum Muster KAMPF GEGEN X existieren? Und wann werden sie verwendet?

Es ist in diesem Beitrag nicht der Ort, um diese Fragen zu beantworten. Die notwendigen Analysen, um die erste Fragestellung zu verfolgen, sind einsichtig: In erster Linie geht es darum, die Füllung von X im Muster KAMPF GEGEN X in Abhängigkeit von Zeit, Thema, Textsorte etc. zu untersuchen. Zum Schluss möchte ich zur zweiten Fragestellung aber noch einige Hinweise geben, in welche Richtung die weiteren Analysen gehen könnten.

Als engere alternative Ausdrucksmöglichkeiten von KAMPF GEGEN X liegen folgende Varianten auf der Hand: KAMPF MIT X und KAMPF DEM/DER [DATIV-OBJEKT]. Es ist zu fragen, ob der propositionale Gehalt der drei Varianten gleich ist:

1. Kampf gegen den Terrorismus
2. Kampf mit dem Terrorismus
3. Kampf dem Terrorismus

Das NZZ-Korpus enthält keine Belege für die Varianten 2 und 3, und nur wenige Belege für die von Varianten 2 und 3 abstrahierten Muster KAMPF MIT X und KAMPF DEM/DER [DATIV-OBJEKT]. Die Füllungen für X in den beiden Mustern sind in Tabelle 7 aufgeführt. Dabei wird sofort klar, dass sich in den untersuchten Daten deutlich unterschiedliche Sprachgebräuche zeigen, wobei es aber schwierig ist, die Unterschiede zu benennen. Vielleicht lässt es sich so fassen: KAMPF DEM/DER [DATIV-OBJEKT] wird überhaupt nicht, KAMPF MIT X [ARTIKEL] nur vereinzelt mit kriegerischen Themen in Verbindung gebracht. Etwas anders das Bild bei KAMPF GEGEN X (vgl. Abbildung 2), wo vor allem die hochfrequenten

Kollokatoren wie *Terror**, *Saddam*, *islamistischer/Kaida-Terror**, *Taliban*, *ETA* etc. in kriegerischen Kontexten zu finden sind.

Tab. 7: Die Füllungen von X in den Mustern KAMPF MIT X und KAMPF DEM/DER [DATIV-OBJEKT] im NZZ-Korpus[21]

KAMPF MIT X [ARTIKEL]	KAMPF DEM/DER [DATIV-OBJEKT]
allen Mitteln gegen Israel; Aussicht auf Erfolg; baskischen Separatisten und Terroristen; Bayern München; Behörden; Berg; Depression; Drachen; drohenden Überschuldung; eigenen Schwächen; Elementen; erbarmungslosen Feind; Gegnern; gleich langen Spiessen; gleicher Waffe; Grammatik; Great-West Lifeco; harten Bandagen; HIV; ihnen; ihren Texten; Immunsystem; Klavierschülergeschlecht; Konkurrenz; Leuten; Mächten einer niederdrückenden Ideologie; nassen Element; Neat-Gegnern; Schönheit; sehr jungen Gegnern; sich selber; sich selbst; Trieb; ungewissem Ausgang; vorpreschenden Spunden; Wettbewerbskommission; Wind, Wasser und Wellen; x Toren; Zeit	„Bussen- und Gebührenterror"; Dengue-Fieber; Diskriminierung von HIV-Positiven; Feinstaubbelastung; Guineawurm; Korruption; Kriminalität; Lobbying; Minenplage; Neoliberalismus; Schuldenmisswirtschaft; Schwefel; Spam; Stau; Straflosigkeit

Dieser Befund steht auf Grund der Datenlage auf schwachen Füßen. Es gibt im NZZ-Korpus zu wenig Belege für die beiden Muster KAMPF MIT X und KAMPF DEM/DER [DATIV-OBJEKT]. Es bietet sich deshalb an, ein weiteres Referenzkorpus, beispielsweise das DeReKo IDS (o.J.)-Korpus des Instituts für Deutsche Sprache, heranzuziehen. Grundlage für die Recherche waren alle im IDS-Korpus öffentlich verfügbaren Tageszeitungen aus Deutschland, Österreich und der Schweiz im Zeitraum von 1991 bis 2006.[22] Die Tabellen 8 und 9 zeigen die häufigsten Füllungen für X in den drei Mustern.[23]

21 Bis auf *Kampf mit dem Berg* (vier Vorkommen) gibt es für jede Mehrworteinheit jeweils nur einen Beleg im Korpus

22 Das so zusammengestellte Korpus enthält damit 4.129.847 Texte mit insg. 960.395.973 Wörtern. Folgende Zeitungen sind in unterschiedlichen Zeiträumen vertreten: Berliner Morgenpost, Die Presse, Frankfurter Rundschau, Hamburger Morgenpost, Kleine Zeitung, Mannheimer Morgen, Neue Kronen-Zeitung, Oberösterreichische Nachrichten, Salzburger Nachrichten, St. Galler Tagblatt, Tiroler Tageszeitung, Vorarlberger Nachrichten, Zürcher Tagesanzeiger.

23 Für das Muster KAMPF DEM/DER [DATIV-OBJEKT] mussten im IDS-Korpus Belege mit Dativ-Objekten mit femininem Genus (*Kampf der Arbeitslosigkeit*) ignoriert werden, da das Korpus nicht syntaktisch annotiert ist und deshalb Dativ-Objekte nicht von Genitiv-Objekten (*der Kampf der Regierung*) unterschieden werden konnten.

Tab. 8: Die Füllungen von X in den Mustern KAMPF GEGEN X und KAMPF MIT X im IDS-Korpus, deren Mindestfrequenzen mindestens 0,5% aller Kollokatoren ausmachen.

KAMPF(S/ES) GEGEN X	#	%	KAMPF(S/ES) MIT X	#	%
Abstieg	453	4,53	[Sportresultat X:Y]	344	11,21
Terror/Terrorismus	329	3,29	Behörde(n)	32	1,04
organisiert*	225	2,25	sich selbst/sich selber	31	1,01
Doping	183	1,83	Drachen	26	0,85
Korruption	161	1,61	Natur	21	0,68
Armut	129	1,29	Bürokratie(n)	20	0,65
Drogen	119	1,19	Tücke(n) des/der	17	0,55
Geldwäsche/Geldwäscherei	105	1,05	allen Mitteln	16	0,52
Kriminalität	96	0,96	Elementen	16	0,52
Uhr	94	0,94	Konkurrenten/Konkurrenz	16	0,52
Mafia	90	0,90	Waffe(n)	16	0,52
AIDS	79	0,79			
illegale(n)	77	0,77			
Windmühle(n)	77	0,77			
internationale(n)	76	0,76			
Krebs	62	0,62			
seine(n)	61	0,61			
Rassismus	58	0,58			
Israel	51	0,51			
Verbrechen	51	0,51			
Total	10.000	100,00	Total	3070	100,00

Tab. 9: Die Füllungen von X im Muster KAMPF DEM X im IDS-Korpus, deren Mindestfrequenzen mindestens 0,5% aller Kollokatoren ausmachen

KAMPF DEM X	#	%	*Fortsetzung…*	#	%
Krebs	33	5,45	Faschismus	4	0,66
Stau	32	5,28	Hochwasser	4	0,66
Hunger	21	3,47	Müll	4	0,66
Dickdarmkrebs	13	2,15	Schimmel*	4	0,66
Brustkrebs	10	1,65	Tod	4	0,66
Elektrosmog	9	1,49	Übergewicht	4	0,66
Fahrraddiebstahl	9	1,49	Atomtod	3	0,50
Herztod	8	1,32	blauen Dunst	3	0,50
Schlaganfall	8	1,32	Chaos	3	0,50
Terror*	8	1,32	Darmkrebs	3	0,50
Feuerbrand	7	1,16	Feinstaub	3	0,50
Rassismus	7	1,16	Grünen Star	3	0,50
Sterilen und Leblosen	7	1,16	Hanf	3	0,50
Alkohol	6	0,99	Kommunismus	3	0,50
Kinderkrebs	6	0,99	Mißbrauch	3	0,50
Pusch	6	0,99	Orgasmus	3	0,50
Doping	5	0,83	Proporz	3	0,50
Kindsmissbrauch	5	0,83	Schilderwald	3	0,50
Verbrechen	5	0,83	Sozialmissbrauch	3	0,50
Winterspeck	5	0,83	Verkehrschaos	3	0,50
Drogen*	4	0,66			
Total				606	100,00

Mit Hilfe dieser Analysen lässt sich der oben formulierte Befund präzisieren:

1. KAMPF GEGEN X: Dominant sind Füllungen mit Themen, die eine internationale, politische Dimension haben und wohl als ‚gesellschaftlich bedeutend' umrissen würden. Ausnahme davon: Der *Kampf gegen den Abstieg*, der im sportlichen Kontext verwendet wird.
2. KAMPF MIT X: Dominant sind Füllungen, die eher Anekdotisches umschreiben (*Kampf mit den Behörden*, der *Bürokratie*, den *Tücken (der Technik*), oder die eigene Persönlichkeit meinen (*Kampf mit sich selbst*). Die Verwendung ist aber sehr heterogen. An der Spitze stehen auch hier Verwendungen im Zusammenhang mit Sportresultaten: *Den Kampf mit 2:0 (gewonnen).*
3. KAMPF DEM X: Auffallend sind hier die dominanten Kontexte Krebs- und andere Krankheiten. Mehrheitlich bewegen sich die Themen eher auf der individuellen, aber durchaus tragischen Dimension, daneben sind aber auch innenpolitische Themen oder Probleme im Bereich Naturgewalten und Verkehr vertreten.

Neben diesen Präferenzen für Füllungen von X gibt es aber auch Belege, die zeigen, dass die Muster teilweise austauschbar sind. So finden sich sowohl Belege für *Kampf gegen das Verbrechen* als auch *Kampf dem Verbrechen* oder *Kampf gegen Rassismus* und *Kampf dem Rassismus* und andere mehr.

Interessant im Zusammenhang mit *Terror* ist natürlich die seltene Verwendung von *Kampf dem Terror**: Gibt es einen semantischen oder pragmatisch-funktionalen Unterschied zwischen *Kampf dem Terror** und *Kampf gegen den Terror**? Fordert der politische Terror-Diskurs eine bestimmte Sprachgebrauchsvariante?

Alle acht Belege für *Kampf dem Terror** im IDS-Korpus sind Verwendungen in Titeln oder als Schlagzeilen von Zeitungsartikeln und haben Parolen-Charakter. Bei dreien können direkte oder indirekte Zitate dahinter vermutet werden:

- Vranitzky, Schüssel: Kampf dem Terror, ob von links oder rechts! (Neue Kronen-Zeitung, 27.4.1995:2)
- US-Gegenoffensive mit Atomwaffen? „Kampf dem Terrorismus mit allen verfügbaren Mitteln" (Salzburger Nachrichten, 24.8.1998, Ressort: Weltpolitik; US-Gegenoffensive mit Atomwaffen?)
- Kampf dem Terrorismus. Der neue FBI-Chef Louis Freeh bezeichnete es als eine seiner wichtigsten Aufgaben, scharf gegen den Terrorismus vorzugehen. (Die Presse, 4.9.1993; IN KÜRZE)

Bei übersetzten Zitaten können sprachliche Feinheiten wie die Unterscheidung von *Kampf gegen* und *Kampf dem* natürlich verloren gehen. Doch aus diskursanalytischer Sicht ist trotzdem interessant zu sehen, welche Sprachgebrauchsvariante auch in der Übersetzung für einen bestimmten

Zweck verwendet wird. Auf Grund der gemachten Beobachtungen könnte z. B. die These aufgestellt werden, dass mit *Kampf dem Terror** eine Sprachgebrauchsvariante verwendet wird, die politische Rhetorik, und vielleicht sogar Distanzierung des Verfassers/der Verfasserin gegenüber dieser Rhetorik, signalisieren soll. Weitere Evidenz für diese These ergibt eine erweiterte, unsystematische Suche in deutschsprachigen Tageszeitungen, wie folgende Belege zeigen:

- Die konservative Zeitung fordert mehr Kampf dem Terrorismus: „Deutschland wird in den kommenden Jahren weiter massiv in den Ausbau der Sicherheitsarchitektur investieren müssen. Es wäre gut, wenn die staatstragenden Parteien, zu denen man neben Union und SPD auch Grüne und FDP zählen möchte, sich darüber gemeinsam Gedanken machten." (SonntagsZeitung, 9.9.2007:23)
- Geiger ist sich sicher, dass für den Finanzplatz Schweiz die Bedeutung des Bankgeheimnisses steigen wird. Denn mit der Begründung „Kampf dem Terrorismus" nähmen die Zugriffsbegehrlichkeiten auf die Privatsphäre in vielen Ländern zu. (Süddeutsche Zeitung, 29.11.2006:V2/2)
- Ein weiteres Beispiel für Abzocke in großem Stil: die Gebührenerhöhung für den neuen Reisepass um 127 Prozent, natürlich unter dem Deckmantel „Kampf dem Terrorismus". (Stuttgarter Zeitung, 31.10.2005, Leserbriefe:40)
- Das Schlagwort „Kampf dem Terrorismus" muss herhalten, die Folgen dieses Kriegs um Einfluss und Öl zu rechtfertigen. Sind nicht auch die Deutschen mitschuldig, wenn die amerikanischen Stützpunkte im Land für Militärtransporte etc. benützt werden? (Stuttgarter Zeitung 16.11.2004, Leserbriefe:8)

Die Liste ließe sich fortsetzen, und dabei zeigt sich, dass die Verwendung von *Kampf dem Terror** auch hier (z. B. in Leserbriefen) in argumentativer Funktion verwendet wird, um diesen ‚Kampf' zu kritisieren.

Die bisherigen Analysen führen zu folgenden Hypothesen:

1. Die Muster KAMPF GEGEN X, KAMPF MIT X und KAMPF DEM X sind grundsätzlich syntaktisch austauschbar. Trotzdem finden sich klare Präferenzen, in welchen Kontexten (mit welchen X) die Muster verwendet werden. Diese Präferenzen verändern sich diachron und unterscheiden sich synchron (z. B. bezüglich Textsorte, kommunikativer Funktion etc.), sind also abhängig von Diskursen.
2. Für *Kampf gegen (den) Terror* und *Kampf dem Terror* gelten in den untersuchten Daten relativ klare Präferenzen für bestimmte kommunikative Zwecke.
3. In den untersuchten Daten wird der Ausdruck *Kampf gegen (den) Terror* weit häufiger verwendet, als die syntaktisch ebenfalls möglichen Alternativen *Kampf dem Terror* und *Kampf mit dem Terror*. Die Präferenz für die eine Variante ist diskursiv begründet. Die vermehrte Verwendung von *Kampf dem Terror* würde z. B. auf einen veränderten Status von *Terror* schließen lassen, einen ähnlichen Status, wie er heute z. B. für *Krebs* oder *Stau* (*Kampf dem Krebs/Stau*) gilt.

Ich breche die Analyse an dieser Stelle ab – sie ist weit davon entfernt, vollständig zu sein. Trotzdem soll sie demonstrieren, in welche Richtung die corpus-driven gestartete und corpus-based vertiefte Sprachgebrauchsanalyse geht und wie die Befunde beschaffen sind, die die Basis für die Beantwortung diskursanalytischer Fragestellungen bilden.

7. Möglichkeiten und Grenzen: Fazit

Die Beispiele im vorherigen Kapitel haben deutlich gemacht: Korpuslinguistische Methoden ersetzen nicht bestehende diskurslinguistische Methoden, sondern ergänzen sie. Allerdings setzen sie an einem grundlegenden Punkt von Diskurslinguistik ein: Eine corpus-driven operierende Korpuslinguistik geht induktiv, und damit hypothesenbildend, vor. Statt nur als Hilfsmittel zur Hypothesenüberprüfung zu dienen, verhilft sie der Diskursanalyse zu einem anderen Startpunkt, in dem zunächst ein Korpus auf seinen musterhaften Sprachgebrauch untersucht wird. Die quantitative und qualitative Analyse der Sprachgebrauchsmuster führt alsdann zu bekannten Kategorien wie Metapher, Topos, Argumentationsmuster, Begriff etc., aber auch zu neu zu entwickelnden Kategorien, und ermöglicht die Beschreibung von deren Funktionen im Diskurs.

Die Sprachgebrauchsmuster bieten darüber hinaus jedoch das Potenzial, neben einem eher thematischen Fokus auch einen Fokus auf die Sprechweise, die diskursive Praxis, einzunehmen. Auch da bietet der induktive Zugang einen Vorteil: Es wäre bedeutend schwieriger, deduktiv vorzugehen, indem vorab mögliche Sprachgebrauchsmuster lexikalisch exakt definiert werden müssten, die als typisch für einen Diskurs stehen könnten. Der induktive Zugang führt jedoch genau zu den Mustern, die auf einer synchronen oder diachronen Achse für einen bestimmten Sprachgebrauchsausschnitt signifikant sind. Und die algorithmische Operationalisierung dieser Muster ermöglicht es, auch umfangreiche Daten auszuwerten.

Aber führt ein Mehr an Daten auch zu einer besseren Diskursanalyse? Ein Mehr an Daten ermöglicht statistisch sicherere Aussagen. Aber der bedeutendere Vorteil von einem größeren Datenumfang zeigt sich in der corpus-driven-Perspektive: Sind nicht jene sprachlichen Veränderungen besonders interessant, die so gering sind, dass sie unserer Aufmerksamkeit als Leserinnen und Leser entgehen, statistisch jedoch signifikant sind, also mit hoher Wahrscheinlichkeit nicht dem Zufall angerechnet werden können? Diskursive Kräfte können ihre Wirkung auf das Sagbare und die

Sprechweise auf subtile Art entfalten – die Wirkung ist deshalb nicht weniger stark.

Auch wenn eine korpuslinguistische Diskursanalyse induktiv vorgeht, vorurteilslos tut sie es nicht. Wenn der Ausgangspunkt rekurrente lexikalische Elemente sind, die in die Beschreibung von Sprachgebrauchsmuster münden, wird mit diesem Ausgangspunkt bereits eine starke Hypothese vorausgesetzt, die solche Sprachgebrauchsmuster als grundlegende Indikatoren für Diskurse annimmt. Die Grenzen einer solchen Diskursanalyse liegen dort, wo eine Gegenhypothese aufgestellt wird. Oder anders ausgedrückt: Auch hier handelt es sich um eine getönte Brille, durch die auf Diskurse geblickt wird. Sie ist m. E. jedoch weniger stark getönt als eher deduktiv und primär hermeneutisch vorgehende Methoden und bietet damit die Chance, Diskurse zu beschreiben, ohne vorschnell selbst Opfer des Diskurses zu werden.

8. Literatur

Banerjee, Satanjeev/Pedersen, Ted (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.

Belica, Cyril (1996): Analysis of Temporal Changes in Corpora. In: International Journal of Corpus Linguistics 1/1, 61–73.

Böke, Karin (1996): Überlegungen zu einer Metaphernanalyse im Dienste einer ‚parzellierten' Sprachgeschichtsschreibung. In: *Karin Böke/Matthias Jung/Martin Wengeler* (Hgg.): Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Georg Stötzel zum 60. Geburtstag gewidmet. Opladen, 431–452.

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

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.

Burger, Harald (1998): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin.

Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hgg.) (1994): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen.

Busse, Dietrich/Teubert, Wolfgang (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: *Busse/Hermanns/Teubert* (1994), 10–28.

DeReKo IDS (o.J.): Das Deutsche Referenzkorpus DeReKo. Online unter *http://www.ids-mannheim.de/kl/projekte/korpora/*.

Evert, Stefan (2005): The Statistics of Word Cooccurrences. Word Pairs and Collocations. Phil. Diss. Institut für maschinelle Sprachverarbeitung. Stuttgart.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt am Main.

Firth, John Rupert (1957): Modes of Meaning. In: Papers in Linguistics 1934–1951, London, 190–215.

Foucault, Michel (1981): Archäologie des Wissens. 10. Aufl. Frankfurt am Main.

Gardt, Andreas (2007): Diskursanalyse – Aktueller theoretischer Ort und methodische Möglichkeiten. In: *Ingo H. Warnke* (Hg.): Diskursanalyse nach Foucault. Theorie und Gegenstände. Berlin/New York, 27–52.

Jung, Matthias (1996): Linguistische Diskursgeschichte. In: *Karin Böke/Matthias Jung/Martin Wengeler* (Hgg.): Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Georg Stötzel zum 60. Geburtstag gewidmet. Opladen, 453–472.

Kilgarriff, Adam (2001): Comparing Corpora. In: International Journal of Corpus Linguistics 6/1, 1–37.

Kluge (1995): Kluge. Etymologisches Wörterbuch der deutschen Sprache. 23., erw. Aufl. Berlin/New York.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen.

Linke, Angelika (2003): Begriffsgeschichte – Diskursgeschichte – Sprachgebrauchsgeschichte. In: *Carsten Dutt* (Hg.): Herausforderungen der Begriffsgeschichte. Heidelberg, 39–49.

Manning, Christopher D./Schütze, Hinrich (2000): Foundations of Statistical Natural Language Processing. 5th printing Cambridge, Massachusetts.

Perkuhn, Rainer/Belica, Cyril (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. In: Sprachreport 22/1, 2–8.

Rayson, Paul/Garside, Roger (2000): Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing corpora. Morristown, N.J., 1–6.

Scherer, Carmen (2006): Korpuslinguistik. Heidelberg.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Spitzmüller, Jürgen (2005): Metasprachdiskurse. Einstellungen zu Anglizismen und ihre wissenschaftliche Rezeption. Berlin/New York.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: *Kathrin Steyer* (Hg.): Wortverbindungen – mehr oder weniger fest. Berlin/New York, 87–116.

Teubert, Wolfgang (2005): My version of corpus linguistics. In: International Journal of Corpus Linguistics 10/1, 1–13.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam.

Warnke, Ingo H. (2007): Diskurslinguistik nach Foucault – Dimensionen einer Sprachwissenschaft jenseits textueller Grenzen. In: *Ders.* (Hg.): Diskursanalyse nach Foucault. Theorie und Gegenstände. Berlin/New York, 3–24.

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