Preprint von: Bubenhofer, Noah (2015): Muster aus korpuslinguistischer Sicht. In: Dürscheid, C./Schneider, J. G. (Hrsg.): Handbuch Satz – Äußerung – Schema, Sprachwissen. Berlin / New York: De Gruyter. S. 485-502.

Noah Bubenhofer

Muster aus korpuslinguistischer Sicht

Abstract
Die Stärke korpuslinguistischer Methoden liegt darin, in großen Textmengen das serielle Auftreten eines bestimmten sprachlichen Phänomens zu entdecken. Es existiert deshalb vor dem Hintergrund des Firth'schen Kontextualismus eine lange Tradition, typische Kookkurrenzen von sprachlichen Einheiten statistisch zu berechnen, beispielsweise durch Kollokationsanalysen. Doch auch komplexere Phänomene, die über Wortpaare hinaus gehen und typische Kookkurrenzen von sprachlichen Einheiten unterschiedlicher Ebenen (Wortform, Grundform, morphosyntaktische Klasse etc.) umfassen, und die für bestimmte Textdaten typisch sind, können datengeleitet eruiert werden. Im Beitrag wird zunächst anhand von Beispielen die Palette der korpuslinguistischen Zugriffe auf Musterhaftigkeit in Texten dargestellt. Ausgehend von grundsätzlichen Gedanken zur korpuslinguistischen Perspektive auf Musterhaftigkeit werden die beiden wichtigen Paradigmen der Korpuslinguistik, korpusbasierte und datengeleitete Verfahren, herausgearbeitet. Auf der Grundlage dieser Überlegungen werden dann die wichtigsten Ansätze der Korpuslinguistik vorgestellt, die mit quantitativen Methoden Musterhaftigkeit in Textdaten testen oder entdecken.

Korpuslinguistik, Kollokationen, Mehrworteinheiten, Usuelle Wortverbindungen, n-Gramme, komplexe n-Gramme, korpusbasiert, datengeleitet, Kontextualismus, Musterhaftigkeit

1 Einleitung

Nachdem in den vorangegangenen Kapiteln hauptsächlich aus einer theoretischen Perspektive Phänomene zwischen Satz/Äußerung und Schema modelliert worden sind, geht es im Folgenden darum, einen quantitativ-empirischen Blick einzunehmen. Aus korpuslinguistischer Perspektive treten Phänomene, die man als

„Muster“ bezeichnen könnte, in vielen unterschiedlichen Kontexten zutage. Zunächst sollen deshalb Beispiele vorgestellt werden, die unterschiedliche Typen von Musterhaftigkeit zeigen, bevor im Anschluss daran auf die theoretischen und methodischen Aspekte eingegangen wird.

Ein erstes Beispiel von Mustern in der Korpuslinguistik sind Kollokationen, wie in Tabelle 1 auszugsweise dargestellt (vgl. Belica 2007). Es handelt sich um die ersten elf Kollokatoren zum Lemma „Maßnahme” (in allen Flexionsformen), also um Wortformen, die im Deutschen Referenzkorpus DeReKo (Kupietz et al. 2010) überzufällig häufig zusammen mit „Maßnahme” vorkommen. Im Unterschied zu vielen Kollokationstabellen unterscheidet sich diese dadurch, dass weitere Informationen verfügbar sind. So ist zu jedem Kollokator ein „syntagmatisches Muster” genannt, das die typische Verwendung der Kollokation wiedergibt. Die Häufigkeit des syntagmatischen Musters ist mit einer Prozentzahl angegeben. Lesebeispiel: In 84% aller Fälle, in denen das Lemma „Maßnahme” zusammen mit dem Lemma „ergreifen” vorkommt, geschieht dies in der Form des syntagmatischen Musters „Maßnahmen [zu] ergreifen um” (wobei „zu” nur manchmal vorkommt). In der originalen Ausgabe der Kollokationen werden darüber hinaus eine Reihe weiterer Informationen angegeben, so z.B. der statistische Wert, der die Signifikanz der Kollokation wiedergibt, die absoluten Frequenzen, mit denen die Kollokation im Korpus auftritt, der Bereich vor bzw. nach dem Suchwort „Maßnahme“, in dem der Kollokator auftritt, sowie die Belege für die Kollokation im Korpus. Weiter werden sog. „sekundäre“ Kollokatoren berechnet: Für jede Kollokation wird angegeben, welche weiteren Kollokatoren zur jeweiligen Kollokation ebenfalls überzufällig häufig in den Daten vorkommen.

Kollokatoren	*syntagmatisches Muster*	
ergreifen	84%	Maßnahmen [zu] ergreifen um
ergriffen	80%	Maßnahmen [...] ergriffen werden
vertrauensbildende	56%	vertrauensbildende [...] Maßnahmen
konkrete	78%	konkrete [...] Maßnahmen
solche	45%	solche [...] Maßnahmen
flankierenden	71%	die|den flankierenden [...] Massnahmen
bauliche	57%	durch bauliche […] Maßnahmen
weitere	65%	weitere […] Maßnahmen
vorbeugende	62%	vorbeugende […] Maßnahmen
geeignete	79%	durch geeignete […] Maßnahmen
notwendigen	83%	die|alle notwendigen […] Maßnahmen

Tabelle 1: Ausschnitt (und um Spalten gekürzte Version) aus der Kookkurrenzdatenbank CCDB (Belica 2007) zum Lemma „Maßnahme“

Die Kollokationsanalyse trägt also dazu bei, die musterhafte Verwendungsweise des Lemmas „Maßnahme“ in einem bestimmten Korpus aufzudecken.

Im zweiten Beispiel werden musterhafte Strukturen in einem Korpus visualisiert. Die Datenbasis sind Bergsteigerberichte aus den Periodika des Schweizer

Abbildung 1: Kollokationsnetz eines Korpus alpinistischer Literatur (Text+Berg-Korpus)

Alpenclubs von 1864 bis 2010 (Text+Berg-Korpus: Bubenhofer et al. 2013). Ähnlich wie beim Beispiel oben wurden Kollokationen berechnet, allerdings wurde dies nicht selektiv für bestimmte Lemmata gemacht, sondern für alle, die für das Korpus (im Vergleich zu einem Referenzkorpus) statistisch signifikant sind. Anschließend wurden alle Kollokationen in einem Netz visualisiert. Abbildung 1 zeigt einen Überblick über das Netz. Durch die Visualisierung als Graph werden Verdichtungsbereiche sichtbar, also Bereiche, in denen einzelne Lemmata (Knoten) besonders viele Verknüpfungen (Kanten) untereinander aufweisen. In Abbildung 2 ist ein Detail aus dem Kollokationsgraph abgebildet. Es zeigt einen Verdichtungsbereich um die Themen Trinken/Essen, Unwetter, Müdigkeit und Tiere. Das Netz gibt in seiner Gesamtheit die typische Bergsteigererzählung wieder, indem die typischen Kollokationen und deren Vernetzung untereinander aufgezeigt werden.

Während Kollokationen auf der Grundlage von Wort- oder Grundformen berechnet werden und (im Normalfall) nur aus Basis und Kollokator bestehen, sind sog. n-Gramme längere Einheiten: Sie bestehen aus *n* Einheiten (Bigramme, Tri-

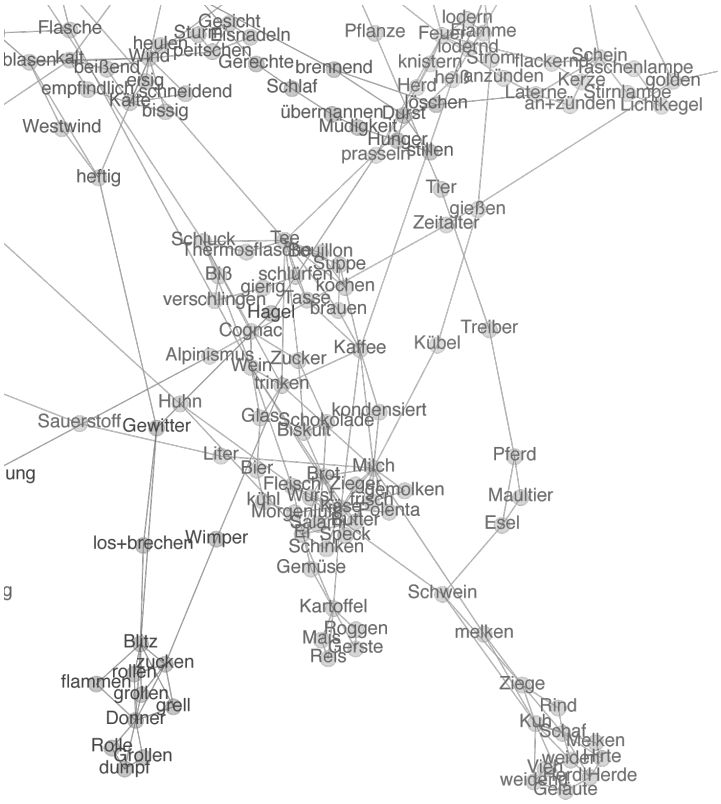

Abbildung 2: Ausschnitt aus dem Kollokationsgraphen des Text+Berg-Korpus

gramme etc.), z.B. Wortformen, die unmittelbar aufeinander folgen oder aber in einem bestimmten „Fenster“ der Länge *x* vorkommen. In einem Korpus lassen sich alle vorkommenden n-Gramme berechnen und nach Frequenz ordnen. Eine Spielart davon sind sog. „komplexe n-Gramme“, die aus einer Kombination von Wortformen und morphosyntaktischen Informationen (Part-of-Speech) oder beliebigen anderen Elementen bestehen können.

Berechnet man die komplexen n-Gramme im oben genannten Text+Berg-Korpus alpinistischer Berichte separat für bestimmte Zeitperioden, können die n-Gramme gefunden werden, die statistisch signifikant für eine Periode im Vergleich zum Rest sind. So sind z.B. die folgenden komplexen n-Gramme typisch für die Zeit von 1880 bis 1899 (vgl. für eine ausführliche Darstellung Bubenhofer/Scharloth 2011; Bubenhofer/Schröter 2012):

Tabelle 2: Beispiele für typische komplexe n-Gramme für die Zeit von 1880 bis 1899 im Text+Berg-Korpus (ADJA = Adjektiv, NN = Nomen, ADV = Adverb, APPR = Präposition, CARD = Kardinalzahl, ART = Artikel)

n-Gramm	Auswahl an Beispielen
ADJA Stunde ADJA (NN)	halben Stunde angenehmer Steigung
	halbe Stunde steilen Anstieges
	halbe Stunde langem Zeitaufwand
	halben Stunde weiteren Weges
	halbe Stunde langer
ADV APPR CARD NN CARD	schon um 9 Uhr 30
ADV APPR CARD Uhr	circa um 1 Uhr 30
	selbst um 12 Uhr 10
	bereits um vier Uhr
	Endlich gegen 9 Uhr
an der ADJA NN des	an der linken Seite des
	an der rechten Seite des
	an der anderen Seite des
	an der breiten Wand des
APPR ART Nähe ART NN	in der Nähe des Gipfels
	in der Nähe der Grenze
	in der Nähe des Muttensees
ADJA Weg APPR ART NN	alten Weg über den Feegletscher
	anderen Weg auf das Gabelhorn
	ausgetretenen Weg durch den Moränenschutt

Tabelle 3: Beispiele für typische komplexe n-Gramme für die Zeit von 1930 bis 1949 im Text+Berg-Korpus (VVFIN = finites Vollverb, PPER = irreflexives Personalpronomen, KOUS = unterordnende Konjunktion)

n-Gramm	Auswahl an Beispielen
dann VVFIN ART NN	dann kündete die Gipfelglocke
dann VVFIN ART ADJA	dann geschieht das Wunder
dann VVFIN PPER auf	dann folgt ein heikler
	dann standen wir auf (dem kühnen Gipfel)
KOUS PPER APPR ART NN	Wie ich in den Riss einstieg
VVFIN	als wir in der Gabel anlangten
KOUS PPER ART NN VVFIN	Bevor wir in das Couloir hinübersteigen
	während wir der Hütte zustrebten
	während wir die Steigeisen ablegten
	Als wir die Passhöhe erreichten
ADV VVFIN ART ADJA NN .	Draussen erwachte ein neuer Tag.
	Dann kam ein trüber Tag.
	Nun naht das schwierigste Stück.
ADV VVFIN PPER VVINF	So lasst uns eilen
	jetzt heisst es handeln

Das Muster „ADJA (=Adjektiv) Stunde ADJA“, ggf. gefolgt von einem Nomen (NN), ist demnach typisch für die älteren Bergsteigerberichte (vgl. für das Tagset Schiller et al. 1995). Das komplexe n-Gramm wird in den Texten beispielsweise

realisiert als „halben Stunde angenehmer Steigung“, „halben Stunde langem Zeitaufwand“ etc. Für die Zeit von 1930 bis 1949 sind dagegen andere komplexen n-Gramme typisch; eine Auswahl davon ist in Tabelle 3 dargestellt.

Die komplexen n-Gramme bewegen sich auf einer abstrakteren Ebene als n-Gramme oder Kollokationen auf der Ebene der Wortformen. Anhand der Beispiele wird sichtbar, dass diese alle dem gleichen syntaktischen Muster folgen, das als Folge von Wortformen und/oder morphosyntaktischen Angaben definiert ist.

Die Beispiele zeigen drei unterschiedliche Erscheinungsformen von rekurrenten sprachlichen Einheiten – musterhaftem Sprachgebrauch –, die allesamt über datengeleitete Verfahren der Korpusanalyse gewonnen wurden. Die terminologische Fassung dieser Phänomene ist uneinheitlich. Neben „Kollokationen“ (Evert 2005) und „(komplexen) n-Grammen“ (Manning/Schütze 2002, 192ff.; Bubenhofer/Scharloth 2013) werden in der Literatur die genannten Phänomene teilweise auch „Usuelle Wortverbindungen“ (Steyer 2013), „Kookkurrenzen“ (Lemnitzer 1997, 124), „Multi-Word Unit“ bzw. „Multi-Word Expression“ (Halliday et al. 2004, 121; Oakes 1998, 184; Sinclair 2004, 31) „Collostructions“ (Stefanowitsch/Gries 2003) oder „Concgrams“ (Cheng et al. 2006) genannt, um nur die wichtigsten Vertreter aus korpuslinguistischer Perspektive zu nennen.

2 Die korpuslinguistische Perspektive

Die korpuslinguistische Arbeit mit großen Textdatenmengen lässt Analysemethoden entstehen, die nicht den Einzelbeleg, sondern das musterhafte Auftreten bestimmter Phänomene in den Daten im Blick haben. Die Liste der Trefferstellen einer Suche in einem Korpus ist deswegen zunächst uninteressant. Bereits die einfachsten Darstellungsmodi von Korpustreffern verfolgen deshalb das Ziel, das Entdecken von Mustern in der Belegmenge zu ermöglichen. Die sog. „Key Word in Context“-Ansicht (KWiC) zentriert den Beleg auf das Suchwort und reduziert die Umgebung auf ein Minimum, um den Blick auf den unmittelbaren Kontext vor und nach dem Suchwort zu lenken. Die Sortierung der Belege ist die nächste Möglichkeit, einen Überblick über die Musterhaftigkeit des Kontextes zu gewinnen: Abbildung 3 zeigt einen Ausschnitt aus einer KWiC-Darstellung der Suche nach „Maßnahmen“ im Deutschen Referenzkorpus DeReKo über die Schnittstelle COSMAS II (Kupietz et al. 2010). Die Ergebnisse sind nach dem ersten und zweiten Vorgänger-Wort geordnet und die KWiC-Darstellung vermittelt so bereits einen ersten Überblick über häufige syntaktische Einbettungen des Suchwortes. Ähnlich ist die Darstellung von Treffermengen in anderen Korpora gestaltet, so z.B. auch in den DWDS-Korpora (vgl. Geyken 2007).

Besonders produktiv für die Zusammenfassung von Belegmengen hat sich das auf Firth zurückgehende Konzept der Kollokation erwiesen (Firth 1957). Firth

Abbildung 3: Beispiel einer KWiC-Darstellung im DeReKo über COSMAS II

definiert mehrere „Modi von Bedeutung“, darunter den Modus „meaning by ‚collocation’“:

> One of the meanings of *ass* is its habitual collocation with an immediatley preceding *you silly*, and with other phrases of address of personal reference (Firth 1957, 194).

Mit „habitual“ tritt das empirische Moment in die Diskussion ein: Das Bindungsverhalten von Wörtern ist nicht zufällig; empirisch zeigen sich offensichtlich bestimmte Bindungsmuster, also Bindungen die gebräuchlicher sind als andere, obwohl sie syntaktisch und semantisch ebenfalls möglich wären (vgl. „Zähne putzen“ statt „Zähne waschen“ – aber „laver les dents“ statt „nettoyer les dents“).

Firth selber formalisiert sein Konzept der Kollokationen nicht weiter, dies wurde erst später geleistet (Evert 2009, 1213). Aus korpuslinguistischer Perspektive ist es naheliegend, Phänomene, die über Gebrauchshäufigkeiten gefasst werden können, empirisch-quantitativ zu operationalisieren. Im Fall der Kollokationen wurde eine breite Palette von Formalisierungen für verschiedene Einsatzzwecke entwickelt und diskutiert (dazu mehr weiter unten). Die in Tabelle 1 als Ausschnitt gezeigten Kollokatoren zu „Maßnahme“ zeigen, wie die riesige Treffermenge durch die statistische Zusammenfassung über die signifikantesten Kollokatoren und syntaktischen Muster den Blick auf die typischen Verwendungsweisen lenkt. Jetzt ist es möglich, eine musterhafte Struktur zu entdecken und linguistisch zu interpretieren.

Musterentdeckende Verfahren sind bei empirisch-quantitativen Analysen also zentral, da bei großen Datenmengen die Belegmengen nicht mehr überblickt werden können. Es ist allerdings nicht nur die Not der praktischen Analysearbeit, die eine Fülle von korpuslinguistischen Methoden entstehen ließ, um Muster in großen Textmengen zu entdecken. Dies wird bei einem Blick auf Diskussionen zum Selbstverständnis der Korpuslinguistik deutlich: Ist die Korpuslinguistik eine Hilfswissenschaft, eine Methode oder eher ein Denkstil, der den linguistischen Zugang zu Sprache grundsätzlich verändert?

Perkuhn und Belica machen deutlich, dass digitale Korpora nicht nur „Belegsammlungen oder Zettelkästen in elektronischer Form" sind, sondern eine eigene „korpuslinguistische Perspektive" ermöglichen (Perkuhn/Belica 2006, 2). Sinclair formulierte diese neue Perspektive 1990 wie folgt:

> The study of language is moving into a new era in which the exploitation of modern computers will be at the centre of progress. The machines can be harnessed in order to test our hypotheses, they can show us things that we may not already know and even things which shake our faith quite a bit in established models, and which may cause us to revise our ideas very substantially. In all of this my plea is to trust the text. (Sinclair 2004, 23)

Korpora dienen also nicht nur der Überprüfung von Hypothesen entlang von bestehenden linguistischen Kategorien. Korpuslinguistik ermöglicht es, ausgehend von den Daten und diese ernst nehmend, neue Hypothesen zu generieren und damit auch neue linguistische Kategorien zu bilden. Dieses Paradigma wird als „corpus-driven" oder „datengeleitet" bezeichnet (Teubert 2005, 4; Tognini-Bonelli 2001; Belica/Steyer 2008; Steyer 2004; Bubenhofer 2009; Scharloth et al. 2013); es hebt sich ab von korpusbasierten („corpus based") Paradigmen und erhebt den Anspruch, dass eine solche korpuslinguistische Perspektive einen neuen Blick auf linguistische Daten ermöglicht.

Für eine datengeleitete Korpuslinguistik sind musterentdeckende Verfahren noch wichtiger als für klassische Ansätze. Sie ermöglichen überhaupt erst die Analyse der entsprechend großen Datenmengen und sind die Grundsteine, um durch eine linguistische Interpretation der Muster zu neuen Kategorien zu gelangen. Dabei wird die Methodenpalette gegenwärtig weiter angereichert, indem Verfahren des Data Minings (maschinelles Erkennen von Korrelationen in Daten) und der visuellen Analyse für linguistische Fragestellungen fruchtbar gemacht werden (Risch et al. 2008; Rohrdantz et al. 2010; Bubenhofer im Druck).

Muster jeglicher Art sind in der Korpuslinguistik also zentrale Analysekategorien, die vor allem in datengeleiteter Perspektive als Emergenzphänomen wahrgenommen werden:

> Die Strukturen in der Sprache kommen nicht erst dadurch zustande, dass die Gesetzmäßigkeiten durch unseren Geist *er*-funden werden. Das Systemhafte steckt vielmehr in der Sprache selbst, es tritt *emergent* aus ihr hervor, so dass es von unserem Geist quasi nur noch *ge*-funden werden muss. (Perkuhn et al. 2012, 13)

Diese Sichtweise im Hintergrund werden im Folgenden verschiedene methodische Zugriffe auf Musterhaftigkeit aus korpuslinguistischer Perspektive dargestellt.

3 Methodische Zugriffe auf sprachliche Muster

Grundsätzlich ist „Muster“ ein schillernder Begriff und wird, überlappend oder in Abgrenzung zu Konzepten wie „Schema“ und „Konstruktion“ in den linguistischen Teildisziplinen unterschiedlich verstanden. Eine ausführliche Begriffsbestimmung leistet Bücker (in diesem Band). Die bisherigen Ausführungen haben deutlich gemacht, dass aus korpuslinguistischer Perspektive musterhafte Sprache ein Phänomen des Sprach*gebrauchs* ist. Ob ein Sprachgebrauchsmuster eine mentale Realität widerspiegelt, also z.B. kognitiv als lexikalische Einheit vorgeprägt und abrufbar ist (vgl. Ziem, in diesem Band), oder sich aus kultureller Praxis oder sozialem Sprachhandeln ergibt (vgl. Linke, in diesem Band), bleibt dabei vorerst offen und soll an dieser Stelle auch nicht diskutiert werden. Es ist aber hilfreich, mit Steyer (2013, 41) zwei Aspekte von Muster auseinanderzuhalten:

1) Muster als „durch den Sprachgebrauch erreichte Vorgeprägtheit von Wortkombinationen als auch die Struktur von Wortverbindungen im Sinne einer Konstruktion“ (Häcki Buhofer 2011, 506).

2) Musterhafter Sprachgebrauch als post-hoc festgestellte rekurrente Verwendung beliebiger sprachlicher Einheiten (Bubenhofer 2009, 24).

Natürlich sind beide Aspekte die Kehrseiten derselben Medaille: Durch den Sprachgebrauch geprägte Wortkombinationen (z.B. „Guten Tag“) werden – weil sie Ergebnis von rekurrentem Sprachgebrauchs sind – musterhaft in bestimmten Situationen (z.B. Begrüßungen) verwendet. Daraus können sich Handlungsmuster verfestigen, die man als „kommunikative Gattungen“ (Günthner/Knoblauch 1994) fassen kann.

Beide Aspekte spielen in der Korpuslinguistik eine Rolle: Bei der Berechnung von Kollokationen geht es z.B. darum, die Assoziationsstärke zwischen „Guten“ und „Tag“ zu berechnen um Hinweise über die Vorgeprägtheit dieser Wendung zu gewinnen (1. Aspekt). Weiter dienen Distributionsanalysen über die Verteilung bestimmter Phänomene (z.B. von „Guten Tag“) über verschiedene Korpora oder die Berechnung der Korrelation eines Phänomens mit weiteren Phänomenen (z.B. „Guten Tag“ in Abhängigkeit von Textsorten, zeitlichen Epochen etc.) dazu, abzuschätzen, wie musterhaft ein Phänomen verteilt ist (2. Aspekt). Das Schwergewicht liegt im Folgenden jedoch auf dem ersten Aspekt, wobei, da die Aspekte eng miteinander verknüpft sind, die Abgrenzung nicht immer trennscharf sein kann.

3.1 Kollokationen

Kollokationen sind Kombinationen von zwei Wörtern, die in natürlicher Sprache eine Tendenz aufweisen, nahe beieinander aufzutreten (Evert 2009, 1214). Dabei bleibt zu klären, was mit „Wort", „Tendenz" und „nahe beieinander auftreten" gemeint ist. Zudem ist es wichtig darauf hinzuweisen, dass es unterschiedliche Auffassungen von Kollokationen gibt. Evert (2009, 1213) unterscheidet zwischen einem „empirischen Konzept" (wofür er die Bezeichnung „Kollokationen" verwendet) und einem „theoretischen Konzept" (von Evert „Mehrworteinheit" genannt), die mit den Namen Sinclair (im Nachgang zu Firth) und Hausmann (1985) verbunden sind. Hausmann nimmt die phraseologische Perspektive ein und definiert „Kollokation" enger als die Vertreter des Firth'schen und Sinclair'schen Kollokationenbegriffs. Auch Bartsch (2004, 76) definiert Kollokationen enger als „lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other".

Die Differenzen zwischen den Kollokationendefinitionen sind es aber nicht Wert, deswegen einen „Kollokationskrieg" (Hausmann 2004) zu führen. Aus der Perspektive einer datengeleiteten Korpuslinguistik sind Kollokationen ein Konzept, dessen genaue Operationalisierung je nach Forschungsinteresse über verschiedene Parameter gesteuert werden kann. Die wichtigen Parameter sind oben bereits kurz erwähnt und werden nun ausführlicher dargestellt:

1) Bestandteile der Kollokation: Klassischerweise werden als Bestandteile der Kollokation Wortformen oder Grundformen angenommen. Wie später gezeigt werden soll, sind aber auch andere sprachliche Elemente denkbar, wie z.B. (morpho-)syntaktische Kategorien oder semantische Klassen.

2) Kookkurrenz: Wann bei zwei sprachlichen Einheiten Kovorkommen (Kookkurrenz) vorliegt, ist ebenfalls unterschiedlich definierbar. Evert (2009, 1221ff.) unterscheidet drei Typen von Kookkurrenz:

a) *Kookkurrenz auf der sprachlichen Oberfläche:* Der einfachste Ansatz, Kookkurrenz zu messen, ist die Definition einer maximalen Spannweite vor und nach dem Suchwort, gemessen in Anzahl Wörtern. Die Definition der Spannweite bewegt sich oft zwischen drei und fünf Wörtern, muss aber je nach Erkenntnisinteresse festgelegt werden. Weiter kann definiert werden, ob die Spannweite die Satzgrenze überschreiten darf oder nicht. Zudem muss die Spannweite nicht symmetrisch sein, um beispielsweise nur den Kontext vor dem Suchwort mit einzubeziehen.

b) *Kookkurrenz in der gleichen Texteinheit:* Nach diesem Prinzip wird Kookkurrenz als Kovorkommen in der gleichen Texteinheit (z.B. Satz, Äußerung, Text etc.) definiert. Damit wird dem Problem begegnet, dass bei der Definition einer Spannweite auf der Textoberfläche eine willkürliche Entscheidung über die Grenze des Kontextes getroffen wird. So schiebt z.B. unter Umständen ein eingeschobener Nebensatz die Distanz zwischen dem Suchwort und dem Kollokator über

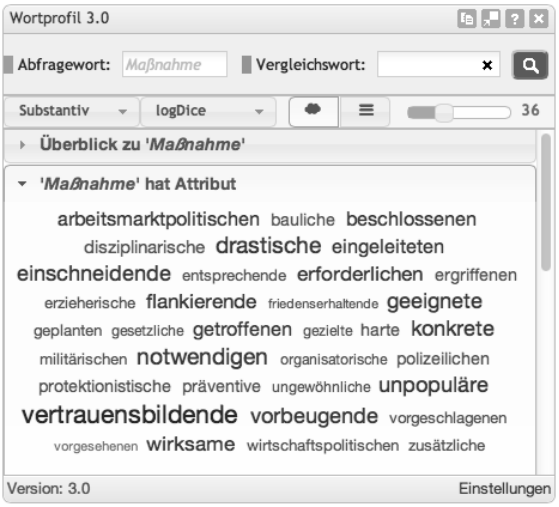

Abbildung 4: Ausschnitt aus einem Wortprofil zu „Maßnahme“ im DWDS-Korpus – Attribute zum Suchwort

die Spannweitengrenze, während der Kollokator in einem Satz ohne den eingeschobenen Nebensatz noch innerhalb der Spannweite wäre.

c) *Syntaktische Kookkurrenz:* Bei diesem Ansatz müssen zwei Wörter in einer bestimmten syntaktischen Relation zueinander stehen, um als potenzielle Kollokation gefunden zu werden. So könnten z.B. als Kollokator zu einem nominalen Suchwort nur Adjektive der gleichen Nominalphrase akzeptiert werden. Dies setzt ein entsprechend computerlinguistisch aufbereitetes Korpus voraus, bei dem etwa morphosyntaktische Klassen („Part-of-Speech“) annotiert sind. Die sog. „Wortprofile“ im DWDS-Korpus-Abfragesystem (vgl. Abbildung 4) sind ein Beispiel für solche syntaktisch definierten Kollokationen (Geyken et al. 2008).

3) Assoziation: Ausgehend von einem Suchwort können nun innerhalb der Spannweite die Häufigkeiten aller vorkommenden Wort-Types bestimmt werden. Um die Bindungsstärken zwischen Suchwort und den Kollokatoren zu berechnen, werden statistische Assoziationsmaße verwendet, die die beobachtete Häufigkeit der Kollokation mit der erwarteten Häufigkeit vergleichen. Die erwartete Häufigkeit ist abhängig von den Frequenzen von Suchwort und Kollokator im Korpus generell und der Korpusgröße. Dies kann man sich an folgendem Beispiel plastisch vorstellen:

- Wenn zwei sehr häufige Wörter auch hin und wieder zusammen vorkommen, ist das statistisch gesehen nicht überraschend.
- Wenn zwei Wörter hingegen je eher selten, jedoch in fast allen Fällen zusammen vorkommen, ist dies statistisch überraschend.

Mit der Berechnung der erwarteten Häufigkeit geht man also von der sog. Nullhypothese aus. Diese behauptet eine gleichmäßige Verteilung der Wörter im Korpus: Alle Wörter haben die gleiche Wahrscheinlichkeit, zusammen aufzutreten, die natürlich von ihrer allgemeinen Häufigkeit im Korpus abhängt. Je stärker

die beobachtete von der erwarteten Häufigkeit abweicht, desto statistisch signifikanter ist die Bindungsstärke der Kollokation (Evert 2009, 1224f.). Es gibt eine Reihe von unterschiedlichen Assoziationsmaßen, die die Abweichung von beobachteten und erwarteten Häufigkeiten bewerten (z.B. t-Test, Log-Likelihood-Test oder Mutual Information). Es führt an dieser Stelle zu weit, detaillierter auf die unterschiedlichen Assoziationsmaße einzugehen. Eine ausführliche Diskussion verschiedener Maße führt Evert (2005).

Diese Prämisse der Nullhypothese, die von einer gleichmäßigen Verteilung der Wörter im Korpus ausgeht, ist in natürlicher Sprache eigentlich nicht haltbar. Aufgrund grammatischer und semantischer Restriktionen kann nicht davon ausgegangen werden, dass Wörter in einem Korpus zufällig verteilt sind – die meisten Assoziationsmaße, die für die Berechnung von Kollokationen verwendet werden, gehen aber davon aus. Die Diskussionen darüber, welche Maße streng statistisch gesehen überhaupt für die Berechnung von Kollokationen verwendet werden dürfen, halten deswegen noch an (Evert 2009, 1244; Kilgarriff 2005; Gries 2005).

Über die genannten Parameter kann relativ genau beeinflusst werden, welche Art von Kollokationen aus den Daten extrahiert werden sollen. Gerade die Wahl des Assoziationsmaßes hat einen großen Einfluss auf die Berechnung der Bindungsstärke, wobei es sinnvoll ist, unterschiedliche Maße miteinander zu vergleichen. Evert (2009, 1236ff.) gibt zudem in Abhängigkeit von unterschiedlichen Forschungsinteressen Entscheidungshilfen.

3.2 Mehrworteinheiten

Der Einblick in eine Kollokationstabelle aus der Kookkurrenzdatenbank CCDB (Belica 2007) in Tabelle 1 hat bereits gezeigt, dass es gewinnbringend ist, die Beschränkung von Kollokationen auf zwei Elemente (z.B. „Maßnahmen ergreifen“) aufzuheben. Bei dem dort verwendeten Algorithmus werden zusätzlich zum (primären) Kollokator („ergreifen“) auch sekundäre Kollokatoren berechnet, also Wörter, die in der Umgebung der Kollokation („Maßnahmen ergreifen“) weiter auftreten (z.B. „um“ – „Maßnahmen ergreifen um“). Es gibt unterschiedliche Ansätze, Kollokationen weiter zu fassen. Im Beispiel der CCDB werden für jede berechnete Kollokation weitere Kollokatoren zur Kollokation berechnet.

Ein anderer Ansatz operiert mit sog. n-Grammen, also Mehrworteinheiten, die aus *n* aufeinander folgenden Wörtern bestehen (Manning/Schütze 2002, 192ff.). Dafür werden in einem Korpus alle kombinatorisch möglichen n-Gramme berechnet. Für n = 3, Trigramme, werden alle Kombinationsmöglichkeiten

Wort 1 – Wort 2 – Wort 3
Wort 2 – Wort 3 – Wort 4
Wort 3 – Wort 4 – Wort 5

etc.

aufgelistet und anschließend gezählt. Auch hier können über Parameter die Typen der zu erfassenden n-Gramme definiert werden:

1) Bestandteile des n-Gramms: Wortform oder Grundform.

2) Diskontinuität: Es kann definiert werden, ob die Wörter kontinuierlich aufeinander folgen müssen oder nicht. Falls nicht, wird ein „Fenster" („Spannweite") definiert, innerhalb dessen sich das n-Gramm bewegen kann.

3) Beachtung von Satz- und Textgrenzen: Ja oder Nein.

Eine nach Auftretensfrequenz geordnete Liste von n-Grammen enthält auf den ersten Plätzen eine Reihe von trivialen Fällen, die aus Wörtern bestehen, die auch für sich genommen sehr häufig sind. Deshalb können ähnlich wie bei den Kollokationen Assoziationsmaße verwendet werden, um die Bindungsstärke der n-Gramme zu berechnen. Da mehr als zwei Elemente vorhanden sind, müssen die statistischen Maße allerdings angepasst werden. Anpassungen für n-Gramme 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.

Anstelle von Assoziationsmaßen zur Berechnung der Bindungsstärke des n-Gramms, gibt es eine weitere Methode, um hochfrequente aber triviale n-Gramme aus den Daten zu entfernen. Die Idee besteht darin, mit einem Referenzkorpus zu arbeiten und die „Keyness" (Scott/Tribble 2006; Bondi/Scott 2010) jedes n-Gramms im Untersuchungskorpus im Vergleich zum Referenzkorpus zu berechnen. Mit Keyness ist ein Assoziationsmaß gemeint (ähnlich wie bei den oben diskutierten Assoziationsmaßen für die Berechnung von Kollokationen), 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 n-Grammen zu berechnen, wie Bubenhofer (2009) gezeigt hat. Dort wurden beispielsweise n-Gramme in einem Zeitungskorpus berechnet und die Zeiträume 1995-1997 und 2003-2005 einander gegenübergestellt. N-Gramme wie „die bosnischen Serben", „und der Opposition" oder „gegen die Korruption" sind (im Ressort „Ausland") typisch für die ältere, „gegen den Irak", „Abzug aus dem", „[Kampf/Krieg] gegen den Terrorismus" dagegen für die neuere Periode (Bubenhofer 2009, 210). Bei Verfahren dieser Art ist die Wahl des Referenzkorpus offensichtlich entscheidend um zu steuern, welche Vergleichsparameter von Interesse sind. Im genannte Fall liegt das Interesse darin, diachrone Veränderungen in der Verwendung von n-Grammen zu beobachten. Genauso denkbar wäre jedoch z.B. der Vergleich zwischen Textsorten, Themen, Sprecher/innen etc.

3.3 Komplexe Formen

Die Berechnung von n-Grammen auf der Basis von Wortformen führt mitunter zu unbefriedigenden Ergebnissen. So gehen die n-Gramme „gegen den Terrorismus“, „Kampf gegen den“, „Kampf gegen Terrorismus“ und „Krieg gegen den“ auf ein gemeinsames, abstrakteres Muster „[Kampf/Krieg] gegen [den] [Terror/Terrorismus]“ zurück. Komplexere Formen der Berechnung von n-Grammen versuchen, dieses Problem anzugehen. Eine Möglichkeit besteht darin, syntaktische Restriktionen zu definieren, ähnlich wie im Fall der syntaktischen Kookkurrenz bei der Berechnung von Kollokationen. Solche Ansätze werden etwa vor dem Hintergrund konstruktionsgrammatischer Interessen angewendet (Stefanowitsch/Gries 2003).

Ein stärker datengeleiteter Ansatz ist die Berechnung sog. „komplexer n-Gramme“ (Scharloth/Bubenhofer 2011; Hein/Bubenhofer im Druck), wie bereits oben in Tabelle 2 und Tabelle 3 vorgestellt. Noch kaum ausgelotet sind zudem Möglichkeiten, ausgehend von n-Grammen deren typischen Anordnungen nacheinander in Texten zu berechnen. Einen ersten Ansatz demonstrieren Bubenhofer et al. (im Druck) mit der datengeleiteten Berechnung von Mustern in Narrativen auf der Basis von hierarchischen n-Gramm-Kollokationsgraphen, um typische Erzählmuster aufzudecken.

Erwähnenswert sind in diesem Zusammenhang aber auch Ansätze, die zwar von einer datengeleiteten Korpusanalyse ausgehen, danach jedoch stärker auf qualitativ-interpretative Analysen setzen. Ein Beispiel sind die Arbeiten zu Usuellen Wortverbindungen und Wortverbindungsmustern von Steyer (2013). Sie geht von Kollokationsanalysen aus, die dann aber systematisch qualitativ (mit regelmäßigem quantitativem Rückgriff) untersucht werden, um abstrakte Wortverbindungsmuster (z.B. „aus welchen [SUBSTANTIV-GRUPPE] auch immer“) zu erarbeiten (Steyer 2013, 332).

Zuletzt soll noch auf die Berechnung von Kollokationsgraphen wie in Abbildung 1 eingegangen werden. Hier werden für ein Gesamtkorpus alle signifikanten Kollokationen berechnet – ggf. mit einer Beschränkung auf Basen, die im Vergleich zu einem Referenzkorpus typisch für das Untersuchungskorpus sind – und als Netz visualisiert. Damit zeigen sich Verdichtungsbereiche von Wörtern, die (durch ihr Kollokationsverhalten) besonders viele Verbindungen untereinander aufweisen. So kann die binäre Beschränkung von Kollokationen aufgehoben werden, indem ihre Position im Kollokationsnetz gezeigt wird. Anstelle von Wort- oder Grundformen kann es darüber hinaus zielführend sein, mit semantischen Taxonomien zu arbeiten. Ein Beispiel ist die Annotation von Texten mit einer Taxonomie wie dem „Wortschatz nach Sachgruppen“ (Dornseiff 2004) und die Berechnung von Kollokationen auf der Basis der Sachgruppen (Scharloth et al. 2013). Damit zeigen sich musterhafte Strukturen in den Daten auf einer semantischen Ebene.

4 Fazit

Aus korpuslinguistischer Perspektive sind zwei Forschungsparadigmen sichtbar, die sich für musterhafte Strukturen in Textdaten interessieren. Das eher korpusbasierte Paradigma geht von bestehenden linguistischen Analysekategorien aus und zielt darauf ab, diese korpuslinguistisch zu formalisieren, um empirische Evidenz dafür zu finden. Mit Musterhaftigkeit sind damit Phänomene gemeint, die sich an ebendiesen linguistischen Kategorien orientieren. Produktiv für Musterhaftigkeit sind Theorien wie beispielsweise die Phraseologie (Burger 2003), die seit mehreren Jahrzehnten auf musterhafte Strukturen im Sprachgebrauch aufmerksam macht. Ebenso wichtig sind verschiedene Grammatiktheorien, was sich z.B. im Fall der Konstruktionsgrammatik in neuester Zeit zeigt (Lasch/Ziem 2011).

Auf der anderen Seite werden mit dem datengeleiteten Paradigma andere Ziele verfolgt. Auch hier wird mit der Prämisse gearbeitet, dass sich im Sprachgebrauch Musterhaftigkeit feststellen lässt. Musterhaftigkeit wird hier aber als Emergenzphänomen wahrgenommen. Musterhaftigkeit zeigt sich, wenn sehr große Datenmengen datengeleitet analysiert werden. Die gefundenen musterhaften Strukturen sind zunächst statistische Auffälligkeiten, die sich nicht immer leicht in bestehende linguistische Kategorien integrieren lassen. Dadurch ergibt sich aber die Chance, bestehende Kategoriensysteme zu überdenken und aufgrund der empirischen Evidenzen neue Kategorien zu bilden. Zudem lenken datengeleitete Verfahren den Blick auf unauffällige Konstruktionen, die bisher in der linguistischen Analyse marginalisiert worden sind. Beispielhaft sei die Arbeit von Steyer (2013) erwähnt, die den Konstruktionen mit dem Lemma „Grund“ 150 Buchseiten datengeleitete Analyse widmet. Parallel dazu ist zu beobachten, wie sich die disziplinären Grenzen verschieben – etwa bei der Phraseologie, die ihren Gegenstandsbereich von primär idiomatischen und nicht-kompositionellen Wendungen in letzter Zeit massiv ausweitet und genauso nicht-idiomatische, kompositionelle und mehrgliedrige Einheiten und ihren weiteren syntaktischen Kontext in den Blick nimmt (vgl. Steyer 2013, 36).

Das datengeleitete Paradigma könnte in Zukunft noch bedeutender werden. Denn in einer digitalen Welt ist eine datengeleitete Korpuslinguistik gleichsam Chance und Notwendigkeit, um emergente Strukturen auf der Performanz-Ebene von Sprache freizulegen.

5 Literatur

Banerjee, Satanjeev/Pedersen, Ted (2003): The design, implementation, and use of the ngram statistic package. In: Proceedings of the Fourth Internatio-

nal 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.

Belica, Cyril (2007): Kookkurrenzdatenbank CCDB - V3. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Abgerufen am 03.03.2014 von http://corpora.ids-mannheim.de/ccdb/.

Belica, Cyril/Steyer, Kathrin (2008): Korpusanalytische Zugänge zu sprachlichem Usus. In: Vachková, Marie (Hg.): Beiträge zur bilingualen Lexikographie. Prag, 7–24.

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

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

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

Bubenhofer, Noah/Müller, Nicole/Scharloth, Joachim (im Druck): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. In: Zeitschrift für Semiotik, Methoden der Diskursanalyse.

Bubenhofer, Noah/Scharloth, Joachim (2013): Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In: Warnke, Ingo/Meinhof, Ulrike/Reisigl, Martin (Hg.): Diskurslinguistik im Spannungsfeld von Deskription und Kritik. Berlin (Diskursmuster – Discourse Patterns, 1), 147–168.

Bubenhofer, Noah/Scharloth, Joachim (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel/Kamber, Alain (Hg.): La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretatino. Neuchâtel (Travaux neuchâtelois de linguistique, 55), 241–259.

Bubenhofer, Noah/Schröter, Juliane (2012): Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalye. In: Maitz, Péter (Hg.): Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. Berlin/Boston (Studia Linguistica Germanica, 110), 263–287.

Bubenhofer, Noah/Volk, Martin/Klaper, David et al. (Hg.) (2013): Text+Berg-Korpus (Release 147_v03). Abgerufen am 4.3.2014 von http://www.textberg.ch.

Burger, Harald (2003): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin.

Cheng, Winnie/Greaves, Chris/Warren, Martin (2006): From n-gram to skipgram to concgram. In: International Journal of Corpus Linguistics 11 (4), 411–433.

Dornseiff, Franz (2004): Der deutsche Wortschatz nach Sachgruppen. Berlin, New York.

Evert, Stefan (2009): 58. corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus Linguistics. Berlin, New York (Handbücher zur Sprach- und Kommunikationswissenschaft, 29), 1212–1248.

Evert, Stefan (2005): The statistics of word cooccurrences. word pairs and collocations. Stuttgart. Abgerufen am 4.3.2014 von http://elib.uni-stuttgart.de/opus/volltexte/2005/2371.

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

Geyken, Alexander (2007): The DWDS corpus: a reference corpus for the german language of the 20th century. In: Fellbaum, Christiane (Hg.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London, 23–42.

Geyken, Alexander/Didakowski, Jörg/Siebert, Alexander (2008): Generation of word profiles on the basis of a large and balanced german corpus. In: Proceedings of the XIII EURALEX International Congress. 371–385.

Gries, Stefan Thomas (2005): Null-hypothesis significance testing of word frequencies: a follow-up on kilgarriff. In: Corpus Linguistics and Linguistic Theory 1 (2), 277–294.

Günthner, Susanne/Knoblauch, Hubert (1994): „Forms are the Food of Faith." Gattungen als Muster kommunikativen Handelns. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 46 , 693–723.

Häcki Buhofer, Annelies (2011): Lexikografie der Kollokationen zwischen Anforderungen der Theorie und der Praxis. In: Engelberg, Stefan/Holler, Anke/Proost, Kristel (Hg.): Sprachliches Wissen zwischen Lexikon und Grammatik. Berlin, 505–531.

Halliday, Michael Alexander Kirkwood/Teubert, Wolfgang/Yallop, Colin et al. (2004): Lexicology and corpus linguistics. an introduction. London/New York.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, H./Mugdan, J. (Hg.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984. Tübingen (Lexicographica Series Maior), 118–129.

Hausmann, Franz Josef (2004): Was sind eigentlich Kollokationen? In: Steyer, Kathrin (Hg.): Wortverbindungen – mehr oder weniger fest. Berlin/New York, 309–334.

Hein, Katrin/Bubenhofer, Noah (im Druck): Korpuslinguistik konstruktionsgrammatisch. Diskursspezifische n-Gramme zwischen statistischer Signifikanz und semantisch-pragmatischem Mehrwert. In: Lasch, Alexander/Ziem, Alexander (Hg.): Konstruktionsgrammatik IV: Konstruktionen als soziale Konventionen und kognitive Routinen. Tübingen.

Kilgarriff, Adam (2005): Language is never, ever, ever, random. In: Corpus Linguistics and Linguistic Theory 1 (2), 263–276.

Kupietz, Marc/Belica, Cyril/Keibel, Holger et al. (2010): The german reference corpus dereko: a primordial sample for linguistic research. In: Proceedings of the 7th conference on International Language Resources and Evaluation. Valletta, Malta, 1848–1854.

Lasch, Alexander/Ziem, Alexander (Hg.) (2011): Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze. Tübingen.

Lemnitzer, Lothar (1997): Extraktion komplexer Lexeme aus Textkorpora. Tübingen (Reihe Germanistische Linguistik, 180).

Manning, Christopher D/Schütze, Hinrich (2002): Foundations of statistical natural language processing. 5. Aufl. Cambridge, Massachusetts.

Oakes, Michael (1998): Statistics for corpus linguistics. Edinburgh (Edinburgh Textbooks in Empirical Linguistics).

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

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Stuttgart.

Risch, John/Kao, Anne/Poteet, Stephen et al. (2008): Text visualization for visual text analytics. In: Simoff, Simeon/Böhlen, Michael/Mazeika, Arturas (Hg.): Visual Data Mining. Berlin, Heidelberg (Lecture Notes in Computer Science), 154–171.

Rohrdantz, Christian/Koch, Steffen/Jochim, Charles et al. (2010): Visuelle Textanalyse. In: Informatik-Spektrum 33 (6), 601–611, doi: 10.1007/s00287-010-0483-x.

Scharloth, Joachim/Bubenhofer, Noah (2011): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York, 195–230.

Scharloth, Joachim/Eugster, David/Bubenhofer, Noah (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Busse, Dietrich/Teubert, Wolfgang (Hg.): Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden, 345–380.

Schiller, Anne/Teufel, Simone/Thielen, Christine (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart.

Scott, Mike/Tribble, Chris (2006): Textual patterns: key words and corpus analysis in language education.

Silva, Joaquim Ferreira da/Lopes, Gabriel Pereira (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, Florida, 369–381.

Sinclair, John (2004): Trust the text. language, corpus and discourse. London.

Stefanowitsch, Anatol/Gries, Stefan Thomas (2003): Collostructions: Investigating the Interaction of Words and Constructions. In: International Journal of Corpus Linguistics 8 (2), 209–243.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hg.): Wortverbindungen – mehr oder weniger fest. Berlin, New York (Institut für Deutsche Sprache. Jahrbuch 2003), 87–116.

Steyer, Kathrin (2013): Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen.

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 (Studies in Corpus linguistics, 6).

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