Stefan Th. Gries

Zur Identifikation von Mehrwortausdrücken: ein Algorithmus, seine Validierung und weiterführende Überlegungen

Abstract: In diesem Artikel diskutiere ich Aspekte der datengetriebenen Identifikation von Mehrwortausdrücken aus Korpora sowie einen einfachen neuen Ansatz und vier kurze Validierungsstudien; letztere verwenden verschiedene Methoden, kontrastieren den hier vorgestellten Ansatz mit einem konkurrierenden und überprüfen seine Prädiktivität für Spracherwerbsdaten. Abschließend bespreche ich einige Desiderata für zukünftige Forschung zu Mehrwortausdrücken, die bisher oft vernachlässigt wurden.

Keywords: KORPUSLINGUISTIK, MEHRWORTAUSDRÜCKE, HÄUFIGKEIT, ASSOZIATION, DISPERSION.

1 Einleitung

Der vorliegende Beitrag beschäftigt sich mit der datengetriebenen Identifizierung von Kollokationen, Phraseologismen und Mehrwortausdrücken, ein viel untersuchtes Phänomen in der Korpuslinguistik. Zur Identifizierung solcher Ausdrücke sind in der Vergangenheit eine Reihe von Verfahren und verschiedenen statistischen Maßen vorgeschlagen worden, die erwartbarerweise auch unterschiedliche Ergebnisse lieferten. Die verwendeten Verfahren und Maße lassen sich „extern“ evaluieren, indem man etwa ihre Vorhersagen mit dem vergleicht, was Sprecher als Mehrwortausdrücke wahrnehmen. Auf diese Weise erhält man dann eine Einschätzung darüber, inwieweit ein gegebenes statistisches Maß (oder Verfahren) tatsächlich das Konzept („Mehrwortausdruck“) abbildet, für das man es einsetzt. In diesem Beitrag werden einige der vorgeschlagenen Verfahren und statistischen Maße problematisiert und Alternativen umrissen.

Korpuslinguistik ist eine inhärent quantitative/statistische Disziplin, da Korpora streng genommen nur Häufigkeiten von Elementen als Datengrundlage anbieten (so dass alle anderen für Linguisten interessanten Begrifflichkeiten über (Kookkurrenz-)Häufigkeiten operationalisiert werden müssen. Der vorliegende Beitrag möchte einen

Stefan Th. Gries: University of California, Santa Barbara; Department of Linguistics; University of California, Santa Barbara; Santa Barbara, CA 93106-3100; U.S.A., stgries@linguistics.ucsb.edu

Bereich fokussieren, der ein viel untersuchtes Problem in der Korpuslinguistik und für den Status statistischer Maße ist; behandelt wird die themenspezifische Relevanz statistischer Maße anhand der Identifikation von Kollokationen, Phraseologismen und Mehrwortausdrücken.

Mehrwortausdrücke sind aus mehreren Gründen relevant: Im Bereich der maschinellen Sprachverarbeitung sind Mehrwortausdrücke zum Beispiel wichtig, da sie oft keine kompositionelle Bedeutung haben und daher maschinelles Sprachverständnis oder Übersetzen erschweren können; im Bereich der angewandten Linguistik sind sie zum Beispiel relevant, da es für Sprachlerner oft schwierig ist, muttersprachliche Selektion (*native-like selection*, Pawley & Syder 1983) zu erwerben; im Bereich der theoretischen Linguistik sind sie unter anderem relevant, da sie die Frage problematisieren, wie viel das mentale Lexikon enthält und wie viel erst ad hoc / bei Bedarf zusammengesetzt wird.

Kollokationsforschung hat sich für Jahrzehnte damit befasst, sogenannte Assoziationsmaße zu entwickeln, die die Kookkurrenz von zwei Einheiten quantifizieren; mittlerweile sind Dutzende solcher Maße vorgeschlagen worden, die üblicherweise verwendet werden, um potentielle Kollokationen/Phraseologismen einer Rangfolge nach sortieren zu können, die 'Kollokationsstatus' widerspiegelt; intuitiv würde man wahrscheinlich erwarten, dass solch ein Ausdruck wie *hermetisch verriegelt* hohe Assoziationsmaße erzielt, während ein Ausdruck wie *der Tisch* nur vergleichsweise geringe Werte erzielt. Die allermeisten Assoziationsmaße basieren dabei auf einer Kreuztabelle von Häufigkeiten, wie sie in Tabelle 1 dargestellt ist. Die Kookkurrenzhäufigkeit der beiden Ausdrücke *x* und *y*, deren Assoziation quantifiziert werden soll, ist in Zelle *a*, die Zellen *a+b* sowie *a+c* enthalten die Häufigkeiten der beiden Ausdrücke *x* und *y* im Korpus; die Summe *a+b+c+d* ist die Korpusgröße.

Tabelle 1: Schematische Kookkurenztabelle für die meisten korpuslinguistischen Assoziationsmaße

	Ausdruck *y*	andere Ausdrücke	Summe
Ausdruck *x*	beobachtet: *a* erwartet: $^{(a+b)\times(a+c)}/_{n}$	beobachtet: *b* erwartet: $^{(a+b)\times(b+d)}/_{n}$	*a+b*
andere Ausdrücke	beobachtet: *c* erwartet: $^{(c+d)\times(a+c)}/_{n}$	beobachtet: *d* erwartet: $^{(c+d)\times(b+d)}/_{n}$	*c+d*
Summe	*a+c*	*b+d*	*a+b+c+d=n*

Wie in Tabelle 1 dargestellt, werden oft diejenigen Häufigkeiten berechnet, die man gemäß der Nullhypothese erwarten würde, wenn Ausdruck *x* und *y* nicht kollokieren; danach werden üblicherweise Assoziationsmaße berechnet, von denen *pointwise*

mutual information MI, *t* und *loglikelihood G* (siehe Dunning 1993) wohl die am weitesten verbreiteten Maße sind, dargestellt in (1).

(1) a. $MI = log_2 {}^{a_{beobachtet}}/_{a_{erwartet}}$
b. $t = \frac{a_{beobachtet} - a_{erwartet}}{\sqrt{a_{erwartet}}}$
c. $G = 2 \times \sum_a^d beobac\ \ tet \times log \frac{beobachtet}{erwartet}$

Diese und andere Maße wurden in vielerlei Studien auf Zweiwort-Kollokationen (und teilweise auf die Kookkurrenz von Wörtern und/in Konstruktionen im konstruktionsgrammatischen Sinn) angewendet und auf Brauchbarkeit und Vorhersagekraft getestet, vgl. Krenn (2000), Evert & Krenn (2001), Krenn & Evert (2001), Evert (2004) und Pecina (2010) sowie Wiechmann (2008) für Wörter und Konstruktionen.

Im Vergleich zur Erforschung von Zweiwort-Kollokationen, ist die korpuslinguistische *bottom-up* Identifikation von Mehrwortausdrücken weniger weit fortgeschritten. Während es inzwischen viele Studien in der angewandten Linguistik gibt, die sich mit N-Grammen – Sequenzen von *n* Wörtern – beschäftigen, vgl. (2), so sind diese oft problematisch in der Hinsicht, dass die relevanten *N*-Gramme nicht komplett datengetrieben erhoben wurden, indem solche Studien beispielsweise nur bestimmte *N*-Gramm-Längen behandeln; viele Studien zu *lexical bundles* beispielsweise betrachten 4-Gramme. Dies ist problematisch, weil natürlich viele *N*-Gramme keine 4-Gramme sind: *N* kann viele verschiedene Werte annehmen, mindestens die in (2) exemplifizierten (und diese berücksichtigen noch nicht die Möglichkeit von *N*-Grammen mit Lücken).

(2) a. *because of* *n*=2
b. *in spite of* *n*=3
c. *on the other hand* *n*=4
d. *be that as it may* *n*=5
e. *the fact of the matter is* *n*=6

Das bedeutet: Selbst nur die Extraktion von Mehrwortausdrücken ist deutlich komplizierter als die von 'einfachen Kollokationen', weil zur Identifikation eines geeigneten Assoziationsmaßes die Fragen hinzukommen, (i) ob/wie das entsprechende Assoziationsmaß auf mehr als zwei Ausdrücke ausgedehnt werden kann, wie man entscheidet, was die optimale Länge eines angenommenen Mehrwortausdrucks ist und (ii) wie die vielen verschiedenen Kombinierungsmöglichkeiten von mehr als zwei Ausdrücken analysiert werden sollen. Ist *in spite of* bspw. eine Kombination von

(3) *a.* *in* und *spite* und *of*?
b. *in spite* und *of*?
c. *in* und *spite of*?
d. *in* ___ *of* und *spite*?

Ein zentrales Ziel für die Erforschung von *N*-Grammen ist daher, einen Algorithmus zu entwickeln, der *N*-gramme aus Korpora extrahieren kann, ohne dass die Länge der zu erhaltenden *N*-gramme von vorneherein festgelegt wird. In diesem Artikel verfolge ich zwei mit diesem Ziel verwandte Absichten: Nach einem kurzen und sehr selektiven Überblick über einige frühere Studien (Abschnitt 2) diskutiere ich zuerst einen neuen Algorithmus zur Identifikation von Mehrwortausdrücken und zeige, wie dieser in einigen kleinen Validierungsstudien abgeschnitten hat (Abschnitt 3). Zum anderen diskutiere ich zentrale problematische Eigenschaften von Algorithmen zur Identifikation von Mehrwortausdrücken und schlage Erweiterungen vor, die die Identifikation von Mehrwortausdrücken verbessern sollte (Abschnitt 4).

2 Eine Auswahl bisheriger Studien

Viele der bisherigen Studien zur Extraktion von Mehrwortausdrücken versuchen, sich der Komplexität des Problems über einen iterativen Ansatz anzunähern und auf diese Weise die Bestandteile langer *N*-Gramme zu bestimmen. In einer der ersten und einflussreichsten Studien schlug Jelinek (1990) vor:

1. einen Minimalwert *m* für das zu verwendete Assoziationsmaß zu definieren;
2. alle 2-Gramme eines Korpus zu extrahieren;
3. alle Kollokationen zu finden, deren Assoziationsmaß einen Wert von *m* überschreitet und diese zu neuen Einheiten im Korpus zu kombinieren;
4. diesen Prozess iterativ zu wiederholen.

Während sich spätere Arbeiten natürlich in vielen Details von Jelineks Ansatz unterscheiden, so haben viele doch eine ähnliche iterative Struktur. Eine hier wichtige Unterscheidung von verschiedenen Ansätzen ist, ob sie auf Tokenhäufigkeiten basieren oder auf Tokenhäufigkeiten und Assoziation; ich gebe in den folgenden Abschnitten einige Beispiele.

2.1 Frequenz

Wie Jelinek (1990) stellt auch die Studie von Kita et al. (1994) einen iterativen Ansatz vor. Dieser Ansatz basiert darauf zu berechnen, wieviel Aufwand das Prozessieren aller Formen eines Korpus benötigt. Im einfachsten Fall würde jedes Wort einen Aufwand in der Höhe von 1 verursachen, aber wenn *N*-Gramme mit *n*>1 als einfache Wörter rekonzeptualisiert werden, sinkt der Prozessieraufwand in Abhängigkeit zu der Häufigkeit des N-Gramms. Die Berechnung des dort vorgeschlagenen Kosten-Kriteriums (*Kita's cost criterion*) involviert nur die Multiplikation des Häufigkeitsunterschieds zweier *N*-Gramme (von denen eines ein Ein-Wort kürzeres Subgramm des anderen ist) mit der Länge des kürzeren *N*-Gramms), wobei die Assoziationsstärke zwischen den Einheiten des *N*-Gramms keine Rolle spielt.

Die Studie von O'Donnell (2011) ist sehr ähnlich (zitiert Jelinek oder Kita et al. allerdings nicht). O'Donnell's *Adjusted Frequency List* (AFL) extrahiert erst alle *N*-Gramme bis zu einer bestimmten Länge aus einem Korpus und rechnet dann die Häufigkeiten von längeren *N*-Grammen sukzessive und rekursiv aus den Häufigkeiten von kürzeren *N*-Grammen heraus. Auch hier wird die Assoziationsstärke der Elemente, die in einen neuen Mehrwortausdruck eingehen, nicht berücksichtigt.

2.2 (Frequenz und) Assoziation

Die wahrscheinlich erste einflussreichere Diskussion zur Identifikation von Mehrwortausdrücken, die nicht nur Frequenz sondern auch Assoziationsstärke berücksichtigt, ist die bereits erwähnte Studie von Jelinek (1990), in der die Identifikation auf der Basis von *MI* durchgeführt wird. Andere Studien gehen teilweise ähnlich vor, verwenden jedoch andere Assoziationsmaße. Ein Beispiel soll hier zunächst genügen, Wible et. al. (2006) sowie Wible & Tsao (2011). Diese Studien generieren keine Liste aller Mehrwortausdrücke in einem Korpus sondern alle Mehrwortausdrücke mit einem bestimmten Suchwort (und verwenden außerdem Wortklasseninformation); das statistische Assoziationsmaß, das dort verwendet wird, ist ebenfalls *MI*.

Ein sehr interessanter Ansatz, der leider viel zu wenig aufgegriffen wurde, ist Daudaravičius & Murcinkevičienės (2004) Maß der *lexical gravity G*. Dieser Ansatz ist die Generalisierung eines ohnehin schon sehr innovativen Assoziationsmaßes, welches sich von allen anderen dahingehend unterscheidet, dass es nicht nur auf den Tokenfrequenzen a bis d in Tabelle 1 basiert, sondern stattdessen auch die Anzahl an Typen berücksichtigt, die in den Zellen b und c repräsentiert sind. *Lexical gravity G* wird wie folgt berechnet:

(4) $Lexical\ Gravity\ G = log \frac{a \times type\ freq.von\ a+b}{a+b} + log \frac{a \times type\ freq.von\ a+c}{a+c}$

Wie aus (4) ersichtlich ist, nimmt dieses Maß zu, wenn *a* und/oder die Typenfrequenz der Kollokate von Ausdruck *x* und/oder die Typenfrequenz der Kollokate von Ausdruck *y* zunehmen; das Maß nimmt ab, wenn die Tokenfrequenzen *a+b* und/oder *a+c* zunehmen. Obwohl dieses Maß an sich schon als Assoziationsmaß sehr interessant ist, weil es als einziges nicht nur Token- sondern auch Typenfrequenz berücksichtigt, bauen Daudaravičius & Murcinkevičienė es außerdem für *N*-Gramme aus (von ihnen Kollokationsketten, *collocational chains*, genannt), indem sie vorschlagen, solche Ausdrücke als *N*-Gramm zu betrachten, die aus mehreren 2-Grammen bestehen, von denen jedes einen *G*-Wert hat, der >5,5 ist, wobei die Motivation für genau diesen Grenzwert von 5,5 nicht offensichtlich ist.

Gries & Mukherjee (2010) verwenden eine Variante von Daudaravičius & Murcinkevičienės Maß, in der sie (i) für jedes *N*-Gramm den Durchschnitt der *G*-Werte berechnen und dann (ii) für alle *N*-Gramme, deren durchschnittlicher G-Wert >5,5 ist, testen, ob es ein längeres *N*-Gramm gibt, dass das erstere enthält, aber einen höheren Durchschnitt von *G*-Werten hat. Wenn kein solches längeres *N*-Gramm existiert, wird das kürzere behalten, ansonsten das längere. Mit diesem 'Reinigungsprozess' versuchen Gries & Mukherjee, der Tatsache Rechnung zu tragen, dass auch *N*-Gramme mit hohen *G*-Werten theoretisch auch noch in weiteren längeren *N*-Grammen mit ebenfalls hinreichender Assoziationsstärke vorkommen können.

Dieser Abschnitt konnte nur einen kurzen Überblick geben, sollte aber deutlich machen, dass (i) bisherige Arbeiten oft Jelinek (1990) folgen und iterativ vorgehen und dass (ii) existierende Arbeiten sich u. a. dahingehend unterscheiden, ob sie ausschließlich auf Häufigkeiten basieren oder auch Assoziationsstärke berücksichtigen; für weitere Studien vgl. auch das Sonderheft von *Language Resources and Evaluation* zu Mehrwortausdrücken sowie u.a. Nagao & Mori (1994), Ikehara, Shirai, & Uchino (1996), Shimohata, Sugio, & Nagata (1997), da Silva et al. (1999), da Silva & Lopez (1999).

3 MERGE und seine Validierung

Der Ansatz, der hier vorgestellt wird, wird als MERGE bezeichnet (kurz für Multi-word Expressions from the Recursive Grouping of Elements), siehe Wahl (2015), Wahl & Gries (eingereicht a, b). MERGE basiert ebenfalls auf der Logik von Jelinek (1990) und beinhaltet folgende Schritte:

1. Ermittle die Häufigkeiten aller 1-Gramme eines Korpus (und damit die Korpusgröße);
2. ermittle alle 2-Gramme eines Korpus und ihre Häufigkeiten (diese können Lücken enthalten, in der Anwendung hier beschränke ich mich auf *N*-Gramme ohne Lücken);
3. berechne für jedes 2-Gramm das Assoziationsmaß *loglikelihood ratio G*, welches wie ein Signifikanztest sowohl auf Frequenz als auch auf Assoziation anspricht;
4. ermittle das 2-Gramm mit dem höchsten *G*-Wert und verwandle es in ein neues 1-Gramm;
5. aktualisiere das Korpus, indem alle Vorkommnisse des neu vereinigten 1-Gramms ein neues Wort werden, und aktualisiere alle Häufigkeitslisten entsprechend, und beginne erneut bei Schritt 1.

Das bedeutet, MERGE könnte z. B. *in spite of* dadurch finden, indem in einem Iterationsschritt *in* und *spite* zu *in spite* vereinigt werden (weil dies den höchsten *G*-Wert erzielt hat) und in einem späteren Schritt werden *in spite* und *of* zu *in spite of* verbunden. Dieser Prozess kann entweder enden, weil eine benutzerdefinierte Zahl von Iterationen durchlaufen wurde (beispielsweise 10.000) oder, weil ein benutzerdefinierter Schwellenwert von *G* nicht mehr überschritten wird.

Genau wie andere Studien auch muss auch dieser methodologische Vorschlag sich die Frage gefallen lassen, ob er effektiv ist und valide Ergebnisse generiert. Um diese Frage zu beantworten, haben Wahl & Gries (eingereicht a, b) mehrere kleine Validierungsstudien durchgeführt.

3.1 Validierung 1: Mehrwortausdruckbewertungen linguistisch untrainierter Leser

Wir wendeten MERGE auf die Kombination zweier Korpora an, das Santa Barbara Corpus of Spoken American English (ca. 250K Wörter, Du Bois, Chafe, Meyer, and Thompson 2000; Du Bois, Chafe, Meyer, Thompson & Martey 2003; Du Bois & Englebretson 2004; 2005) und den gesprochenen Teil des ICE-Canada (ca. 450K Wörter, Newman & Columbus 2010). Die beiden Korpora wurden vorbereitet, indem Tags, Transkriptionszeichen und andere Annotationen entfernt wurden; danach wurde MERGE auf das gemeinsame Korpus angewendet (für 20.000 Iterationen). Danach wurden die ersten 40 und die letzten 40 Mehrwortausdrücke ermittelt und in unterschiedlichen (randomisierten) Auswertungsfragebogen zusammen mit einer Anleitung und weiteren konkreten Beispielen an 20 Studierende der University of

California, Santa Barbara verteilt. Sie wurden gebeten, auf einer 7-stufigen Skala anzugeben, wie sehr jeder Mehrwortausdrucke ein "common resuable chunk" sei.

Die Wertungen wurden mit einem gemischten Regressionsmodell ausgewertet, das als zentralen Prädiktor die binäre Variable *Rang* (*die ersten 40* vs. *die letzten 40*) enthielt, die Länge der Mehrwortausdrücke als Kontrollvariable und die größtmöglichen Zufallsfaktorenstruktur (also Achsenabschnitte und Steigungen für Stimuli und Versuchspersonen). Das finale Modell ergab eine hoch signifikante Korrelation ($p<10^{-15}$, $R^2_{marginal}=0,64$) und zeigte, dass die von MERGE als gute/frühe Mehrwortausdrücke generierten Ausdrücke in der Tat signifikant bessere Wertungen erhielten als schlechtere/spätere Mehrwortausdrücke, selbst wenn man Länge und Zufallsfaktoren kontrolliert; der Unterschied betrug 3,87 Punkte auf der 7-Punkt-Skala.

3.2 Validierung 2: MERGE vs. AFL im Vergleich linguistisch untrainierter Leser

In einem zweiten Schritt verglichen wir die Resultate von MERGE mit denen der AFL. Beide Algorithmen wurden auf die selben Korpora wie oben angewendet; MERGE durchlief dieses Mal 1000 Iterationen, für den AFL-Algorithmus setzten wir den erlaubten Minimalfrequenzwert auf 5 und nahmen nach dem Durchlauf des Algorithmus die häufigsten 1000 Mehrwortausdrücke. Von beiden Listen ermittelten wir dann diejenigen Mehrwortausdrücke, die nur von einem der beiden Algorithmen gefunden wurden und zogen eine stratifizierte Zufallsstichprobe von je 180 Mehrwortausdrücken, die dann in randomisierten Fragebögen aufbereitet 20 weiteren Studierenden zur Beurteilung vorgelegt wurden.

Auch diese Beurteilungen wurden mit einem gemischten Regressionsmodell ausgewertet, das als zentralen Prädiktor die binäre Variable Quelle (MERGE vs. AFL) enthielt, die Länge der Mehrwortausdrücke als Kontrollvariable und die größtmöglichen Zufallsfaktorenstruktur (wie oben). Das finale Modell ergab eine signifikante aber sehr schwache Korrelation ($p_{einseitig}=0,022$, $R^2_{marginal}=0,02$). Sie zeigte allerdings, dass die von MERGE als Mehrwortausdrücke generierten Ausdrücke in der Tat signifikant bessere Wertungen erhielten als die von der AFL generierten Mehrwortausdrücke, selbst dann, wenn man Länge und Zufallsfaktoren kontrolliert; der Unterschied betrug 0,6 Punkte auf der 7-Punkt-Skala. Dieses Ergebnis ist trotz der Schwäche des Effekts insofern interessant, als dass es zeigt, dass der Ansatz der Frequenz und Assoziation vereint – wie MERGE – besser abschneidet als AFL, der nur Frequenzinformationen verwendet.

3.3 Validierung 3: MERGE vs. AFL für getaggte Mehrwortausdrücke im BNC

In einer dritten Fallstudie wendeten wir MERGE und AFL auf alle gesprochenen Daten im BNC an, um zu testen, welcher der beiden Ansätze bessere Ergebnisse in der Auffindung derjenigen Mehrwortausdrücke zeigt, die die BNC-Kompilierer als Mehrwortausdrücke getaggt haben (mit dem <mw>…</mw> Tag). Wir wendeten MERGE und AFL an und identifizierten die ersten 10.000 Mehrwortausdrücke beider Algorithmen. Dann prüften wir, wie viele der 388 Mehrwortausdrücke jeder der beiden Algorithmen identifiziert hatte: die AFL fand 93, MERGE 112, ein Unterschied von 20,4%, der gemäß einseitiger Binomialtests (für beide Kontrastrichtungen) signifikant ist (beide $p_{einseitig}<0.018$); das bedeutet, die Performanz von MERGE ist signifikant besser als die der AFL.

3.4 Validierung 4: Mehrwortausdrücke im Erstspracherwerb

Die letzte Fallstudie untersucht, ob die Ergebnisse von MERGE damit korrelieren, welchen *N*-Gramme Kinder anhand des elterlichen Inputs erwerben. Für diese Fallstudie verwendeten wir die Lara- und Thomas-Korpora (Rowland & Fletcher 2006, Lieven et al. 2009). Beide Korpora wurden in ein Trainings- und ein Testkorpus im Verhältnis 2:1 aufgeteilt, danach wurden Kinder- und Erwachsenendaten getrennt. MERGE wurde angewendet auf die Erwachsenendaten des Trainingskorpus (bis der maximale *G*-Wert nicht mehr positiv war). Dann extrahierten wir alle 2-5-Gramme aus dem Testkorpus der beiden Kinder, löschten diejenigen *N*-Gramme, die auch schon im Trainingskorpus der Kinder vorkamen und teilten die verbleibenden *N*-Gramme auf in (i) die, die die Erwachsenen verwendeten und die Kinder später auch und (ii) die, die die Erwachsenen verwendeten, die Kinder später jedoch nicht; dies wurde getan, damit wir testen konnten, ob die Mehrwortausdrücke, die die Kinder erworben hatten, sich durch höhere MERGE-Werte auszeichneten als die, die sie nicht erworben hatten. Wir gruppierten entsprechend die Mehrwortausdrücke in *G*-Wert-Klassen und berechneten für jede Klasse den Prozentsatz der Mehrwortausdrücke, die das Kind gelernt hatte; dies war die abhängige Variable in unserer statistischen Analyse (wurzeltransformiert, um Verteilungsannahmen der Regression nicht zu verletzen). Diese Analyse war ein Regressionsmodell, in dem Kind (Lara vs. Thomas), *G*-Wert-Gruppe und die Länge der Mehrwortausdrücke sowie ihre potentiellen Interaktionen Prädiktoren waren.

Das finale Regressionsmodel ist hoch signifikant ($p<10^{-15}$) und erklärt die Variabilität der Mengen an erworbenen *N*-Grammen sehr gut (adj. $R^2=0.7801$). Das Modell zeigt, dass die Dreifachinteraktion aller Prädiktoren signifikant ist, aber eine

entsprechende Visualisierung zeigt, dass der wichtigste Befund der ist, dass Mehrwortausdrücke aus höheren *G*-Wert-Gruppen der Erwachsenen in der Tat die sind, die deutlich mehr von beiden Kindern geworben wurden.

3.5 Zusammenfassung

In vier Fallstudien konnte gezeigt werden, dass MERGE alles in allem vielversprechende Resultate erzielt:

1. Mehrwortausdrücke, die MERGE hoch bewertet, werden (i) von linguistisch untrainierten Lesern eher als solche wahrgenommen, (ii) besser bewertet als Mehrwortausdrücke, die MERGE niedrig bewertet oder die von dem AFL-Algorithmus erzeugt werden und (iii) werden häufiger erworben;
2. MERGE erzielt bessere Ergebnisse als AFL im Erkennen von Mehrwortausdrücken im BNC.

Diese Resultate zeigen meines Erachtens auch, welche verschiedenen Möglichkeiten zur Validierung Korpuslinguisten zur Verfügung stehen können, obwohl es bei der Identifikation von Mehrwortausdrücken noch Raum zur Verbesserung gibt; der folgende letzte Abschnitt diskutiert hierzu einige Ideen.

4 Vorschläge zur Verbesserung

Ein meiner Ansicht nach besserer Ansatz zur Identifikation von Mehrwortausdrücken erfordert einige Änderungen, die ich in verschiedenen anderen Kontexten separat besprochen habe, aber bisher wenig oder gar nicht in Verbindung mit der Identifikation von Mehrwortausdrücken gebracht wurden. Diese Änderungen haben mit Faktoren zu tun, auf denen nahezu jegliche korpuslinguistische Analyse oder datengetriebene Identifikation von Zweiwort-Kollokationen oder längeren Mehrwortausdrücken basieren: die Wahl des Assoziationsmaßes, die Rolle von Typenfrequenz und die Rolle von Dispersion (und andere, die ich hier nicht besprechen werde).

4.1 Die Wahl des Assoziationsmaßes

Zwei Faktoren sind relevant in Bezug auf die Wahl eines Assoziationsmaßes: erstens der Faktor Direktionalität. Nahezu alle Assoziationsmaße, die bisher verwendet

wurden – für Kollokationsforschung, aber auch für die Identifikation von Mehrwortausdrücken – sind bidirektional, was bedeutet, dass sie die Assoziation zweier Einheiten x und y zueinander ausdrücken, aber nicht differenzieren, ob diese Assoziation wirklich bidirektional ist oder nur jeweils monodirektional vorliegt. Beispiele für 2-Gramme (im gesprochenen Teil des BNCs), die eine maximale bidirektionale Assoziation (gemessen durch Δ*P*-Werte, i.e. die Differenzen von $p(\text{Wort}_1|\text{Wort}_2)$ - $p(\text{Wort}_1|\text{Wort}_2)$) haben, sind die folgenden Ausdrücke (aus bisher nicht veröffentlichten Daten von Gries 2013): *papier mâché*, *fromage frais*, *spina bifida*, *tittle tattle*, *avant garde*, *higgledy piggledy*, *hocus pocus*, *lingua franca*, *modus operandi*, *rigor mortis*, und *terra firma*; alle diese Ausdrücke bestehen aus zwei Teilen, die in diesem Korpus nur miteinander und niemals alleine sonstwo vorkommen: die Assoziation ist perfekt sowie bidirektional. Ausdrücke wie die folgenden sind dagegen Beispiele, bei denen das erste Wort das zweite sehr stark vorhersagt, aber nicht viz versa: *volte face*, *het up*, *insomuch as*, *insofar as*, *habeas corpus*, *upside down*, etc., während in den folgenden Beispielen die Assoziation vom zweiten zum ersten Ausdruck hin verläuft: *de rigueur*, *al fresco*, *agent provocateur*, *super duper*, *ad hominem*, *ad infinitum*. etc. Die allermeisten Assoziationsmaße unterscheiden Ausdrücke mit einer hohen Assoziation in einer Richtung nicht von Ausdrücke mit einer hohen Assoziation in der anderen Richtung, obwohl es durchaus möglich ist, dass eine derartige Unterscheidung einen Einfluss auf die Resultate eines Mehrwortalgorithmus haben könnte unabhängig davon, wie man die Information der beiden Assoziationsrichtungen bzw. Δ*P*-Werte letztendlich mathematisch verwendet.

Der zweite relevante Faktor hat damit zu tun, wie 'konzeptuell rein' das verwendete Assoziationsmaß ist. Werte wie *t* oder *G* reflektieren nicht nur Assoziation, sondern auch Häufigkeit: Wenn man für Tabelle 1 aus Abschnitt 1 *t* und *G* errechnet, dann erhält man andere Resultate, als wenn man alle Werte in Tabelle 1 mit 10 multipliziert und wieder *t* und *G* errechnet; dies gilt nicht für Werte wie die Odds Ratio oder Δ*P*. Es mag Situationen geben, in denen diese Verknüpfung von Assoziation und Häufigkeit in einen einzigen Wert vorteilhaft ist (siehe Gries 2012 im *collostructional analysis*-Kontext), aber dies muss von Fall zu Fall und auf der Basis von empirischen Daten entschieden werden; zur Identifikation von Mehrwortausdrücken gibt es m. E. noch keine systematischen Studien, die vergleichen, welche Art von Assoziationsmaß für einen bestimmten Zweck geeigneter ist. M.a.W., zukünftige Forschung sollte versuchen zu ermitteln, (i) ob Assoziationsmaße, die Assoziation *und* Häufigkeit widerspiegeln, nützlicher sind als solche, die nur Assoziation widerspiegeln, (ii) ob, falls wir Häufigkeit und Assoziation in einem einzigen Maß unterbringen, diese beiden Dimensionen gleich gewichtet werden sollten oder nicht oder (iii), ob wir einen Ansatz zu Assoziationsmaßen brauchen, der beide Dimensionen verwendet, aber eventuell getrennt als 2-Tupel.

4.2 Typenfrequenz

Ein weiterer wichtiger Punkt wurde mit dem Assoziationsmaß *lexical gravity G* von Daudaravičius & Marcinkevičienė (2004) bereits angesprochen. Die Frage ist, ob nicht nur die Tokenfrequenz von Kollokaten eine Rolle bei der Berechnung einer Assoziationsstärke spielen soll sondern auch die Typenfrequenz. Angesichts der Tatsache, dass Typenfrequenz generell ein wichtiges korpuslinguistisches Konzept ist mit Auswirkungen auf Produktivität, Lernbarkeit und Lerngeschwindigkeit, Sprachwandel etc., erscheint es sinnvoll zu versuchen, Typenfrequenz bei der Einschätzung von Assoziation miteinzubeziehen: die Tatsache, dass nach *in spite* fast ausschließlich *of* folgt, ist doch unzweifelhaft relevant und wahrscheinlich über Typenfrequenz direkter und intuitiver abgebildet als durch Tokenfrequenz.

4.3 Dispersion im Korpus

Ein vorerst letzter potentiell wichtiger Punkt ist Dispersion, die Verteilung von Mehrwortausdrücken im Korpus, die beispielsweise über das Maß *DP* (*Deviation of Proportions*, siehe Gries 2008, Biber et al. 2016) leicht gemessen werden kann. Auch wenn korpuslinguistische Arbeiten die Wichtigkeit von Dispersion noch immer unterschätzen, so gibt es doch inzwischen genügend Evidenz, dass die (Un-)Gleichmäßigkeit der Verteilung von Ausdrücken in einem Korpus einen massiven Einfluss auf alle möglichen quantitativen Resultate haben kann. Beispielsweise haben die Ausdrücke *enormous* und *staining* die gleiche Häufigkeit im Brown Korpus (37) und die gleiche Länge in Buchstaben; sie sind dennoch unterschiedlich, wenn man betrachtet, was Häufigkeit oft operationalisieren soll, nämlich die Wahrscheinlichkeit, ein Wort zu sehen/hören: *enormous* kommt in 36 der 500 Teile des Brown Korpus vor, *staining* dagegen in 1, aber dort 37 Mal. Es ist daher wenig verwunderlich, dass Dispersion sich in einigen Studien als ein psycholinguistisch gesehen informativeres Konstrukt oder Messinstrument als Häufigkeit herausstellte (vgl. Adelman, Brown, & Quesada 2006; Gries 2010; Baayen 2010). Angesichts solcher Befunde ist es sinnvoll, sich zu fragen, ob Dispersion für die Identifikation von Mehrwortausdrücken ebenfalls eine Rolle spielt oder spielen sollte, und falls diese Frage bejaht wird, ist die nächste Frage, wie das geschehen sollte: wieder durch eine Kombination in einen einzigen Wert oder als ein weiterer Beitrag zu einem Assoziations-Tupel?

4.4 Fazit

Auch wenn viele Fragen noch ungeklärt bleiben, so sollte klar geworden sein, dass wir bei der korpuslinguistischen und datengetriebenen Erforschung von Mehrwortausdrücken eigentlich noch ziemlich am Anfang stehen: Wenige Ansätze gehen über die Verwendung von Tokenhäufigkeit und größtenteils bidirektionale Assoziationsmaße, die Tokenhäufigkeit enthalten, hinaus. Anderweitig oder sogar generell wichtige Begriffe wie direktionale Assoziation, Typenfrequenz und Dispersion sind für Mehrwortausdrücke so gut wie gar nicht behandelt worden. Auch wenn noch unklar ist, wie diese Begriffe im Detail zusammengeführt werden können – als ein Maß, als Tupel? – so ist doch klar, dass diese Arten von Fragen behandelt werden müssen. In Abwesenheit besserer Lösungen könnte zumindest versucht werden, alle obigen Dimensionen entweder als Tupel oder zunächst über Schwellenwerte zu verwenden. Wenn die Dispersion zu gering wird, wird ein Mehrwortausdruck nicht akzeptiert, eine Logik, die ja schon oft für Häufigkeit angewendet wird, indem Mindesthäufigkeiten postuliert werden. Wie auch immer, es bleibt viel zu tun und ich hoffe, dass dieser Artikel, wenn auch vielleicht keine Lösungen, so doch Lösungswege für zukünftige Studien skizzieren konnte.

5 Literaturangaben

Adelman, James S., Gordon D.A. Brown, & Jose F. Quesada. 2006. Contextual Diversity, not word frequency, determines word-naming and lexical decision times. Psychological Science 19(9). 814-823.

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. The Mental Lexicon 5(3). 436-461.

Biber, Douglas, Randi Reppen, Erin Schnur, & Romy Graham. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. International Journal of Corpus Linguistics 21(4). 439-464.

Daudaravičius, Vidas & Rūta Marcinkevičienė. 2004. Gravity Counts for he boundaries of collocations. International Journal of Corpus Linguistics 9(2). 321-348.

Du Bois, John W., Wallace L. Chafe, Charles Meyers, & Sandra A. Thompson. 2000. Santa Barbara corpus of spoken American English, part 1. Philadelphia: Linguistic Data Consortium.

Du Bois, John W., Wallace L. Chafe, Charles Meyers, Sandra A. Thompson, and Nii Martey. 2003. Santa Barbara corpus of spoken American English, part 2. Philadelphia: Linguistic Data Consortium.

Du Bois, John W. & Robert Englebretson. 2004. Santa Barbara corpus of spoken American English, part 3. Philadelphia: Linguistic Data Consortium.

Du Bois, John W. & Robert Englebretson. 2005. Santa Barbara corpus of spoken American English, part 4. Philadelphia: Linguistic Data Consortium.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1). 61-74.
Evert, Stefan 2004. The statistics of word cooccurrences: word pairs and collocations. PhD dissertation, IMS, University of Stuttgart.
Evert, Stefan & Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, 188-195.
Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13(4). 403–437.
Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies (eds.), Corpus linguistic applications: current studies, new directions, 197-212. Amsterdam: Rodopi.
Gries, Stefan Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. Studies in Language 36(3). 477-510.
Gries, Stefan Th. 2013. 50-something years of work on collocations: what is or should be next …International Journal of Corpus Linguistics 18(1). 137-165.
Ikehara, Satoru, Satoshi Shirai, & Hajime Uchino. 1996. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. Proceedings of the 16th Conference on Computational linguistics, Vol. 1, 574-579.
Jelinek, Frederick. 1990. Self-organized language modeling for speech recognition. In Alex Waibel & Kai-Fu Lee (Eds.), Readings in Speech Recognition. San Mateo, CA: Morgan Kaufmann, 450-506.
Kita, Kenji, Yasuhiko Kato, Takashi Omoto, & Yoneo Yano. 1994. Automatically extracting collocations from corpora for language learning. Journal of Natural Language Processing 1(1). 21-33.
Krenn, Brigitte. 2000. The usual suspects: Data-oriented models for identification and representation of lexical collocations. PhD Thesis, Saarland University.
Krenn, Brigitte & Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. Proceedings of the ACL Workshop on Collocations, 39-46.
Lieven, Elena, Dorothé Salomo, & Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. Cognitive Linguistics 20(3). 481-507.
Nagao, Makoto & Shinsuke Mori. 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. Proceedings of the 15th Conference on Computational Linguistics, 611-615.
Newman, John and Georgie Columbus. 2010. The International Corpus of English – Canada. Edmonton, Alberta: University of Alberta.
O'Donnell, Matthew Brook. 2011. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. ICAME Journal 35. 135-169.
Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), Language and Communication, 191-225. London: Longman.
Pecina, Pavel. 2010. Lexical association measures and collocation extraction. Language Resources and Evaluation 44(1/2). 137-158.

Rowland, Caroline F. & Sarah L. Fletcher. 2006. The effect of sampling on estimates of lexical specificity and error rates. Journal of Child Language 33(4).859-877.

Shimohata, Sayori, Toshiyuki Sugio, & Junji Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 476-481.

da Silva, Joaquin F., Gaël Dias, Sylvie Guilloré, & José G. Pereira Lopes. 1999. Using LocalMaxs Algorithm for the extraction of contiguous and non-contiguous multiword lexical units. Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence, 113-132.

da Silva, Joaquin & Gabriel Pereira Lopes. 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. Proceedings of the 6th Meeting on the Mathematics of Language, 369-381.

Wahl, Alexander. 2015 The Distributional Learning of Multi-Word Expressions: A Computational Approach. Ph.D. dissertation, University of California, Santa Barbara.

Wahl, Alexander & Stefan Th. Gries. eingereicht a. Multi-word expressions: a novel computational approach to their bottom-up statistical extraction.

Wahl, Alexander & Stefan Th. Gries. eingereicht b. Computational extraction of formulaic sequences from corpora: two case studies of a new extraction algorithm.

Wiechmann, Daniel. 2008. On the computation of collostruction strength: testing measures of association as expressions of lexical bias. Corpus Linguistics and Linguistic Theory 4(2). 253-290.

Wible, David & Nai-Lung Tsao. 2011. Towards a new-generation of corpus-derived lexical resources for language learning. In Fanny Meunier, Sylvie De Cock, Gaetanelle Gilquin, & Magali Paquot (eds.), A taste for corpora: in honor of Sylviane Granger, 237-255. Amsterdam & Philadelphia: John Benjamins.