Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse*

Gerhard Heyer, Martin Läuter, Uwe Quasthoff und Christian Wolff

7.1. Einführung

Der vorliegende Beitrag befaßt sich mit einem mehrstufigen Ansatz zur Analyse großer Textcorpora. Dabei liegt der Schwerpunkt auf der Untersuchung sekundärer, insbesondere linguistischer Filter für die Optimierung statistischer Analyseverfahren. Neben Beispielen für solche Filterverfahren werden auch praktische Anwendungen aufgezeigt.

7.2. Textcorpora und statistische Analysen

Durch Analyse sehr großer Textdatenbestände, die sowohl auf Datenträgern (CD ROM) bereitstanden, als auch durch Suchagenten aus dem World Wide Web zusammengestellt wurden, konnten in den vergangenen Jahren im Rahmen des Projektes „Deutscher Wortschatz“ am Institut für Informatik der Universität Leipzig mehrere umfangreiche monolinguale Corpora aufgebaut werden (vgl. Quasthoff, 1998b,a). Dabei lag bei der Quellenauswahl für die Corpora der Schwerpunkt zunächst auf allgemeinsprachlichen Textsorten (Zeitungstexte) sowie auf strukturierten elektronischen Lexika. Neben einem Corpus des Deutschen (Umfang: ca. 300 Mio. laufende Wortformen, ca. 6 Mio. unterschiedliche flektierte Formen und etwa 15 Mio. Beispielsätze) entstanden auch monolinguale Corpora für die Sprachen *Englisch, Französisch, Niederländisch* und *Sorbisch*. Ein wesentliches Anliegen war dabei die Entwicklung einer Verarbeitungsinfrastruktur, die einerseits den CorpusaufbauKorpus!Korpuserstellung weitgehend automatisiert und andererseits eine in das WWW integrierte Zugriffsschnittstelle für die Corpora bereitstellt (vgl. Quasthoff und Wolff, 2000).

Nachdem diese Infrastruktur seit einiger Zeit online im World Wide Web bereitsteht[1] und auch hinsichtlich unterschiedlicher Ausgangsformate für die Quellentexte genutzt werden kann (u. a. ASCII-Text, HTML, PDF, können mittlerweile auch monolinguale *Fachcorpora* erstellt werden, die sich in *inhaltlicher* Hinsicht (Fachtexte einer bestimmten *Domäne*) oder mit Blick auf einen *temporalen* Bezug (Texte eines bestimmten *Zeitraums*) vom umfangreicheren allgemeinsprachlichen Corpus unterscheiden und Ausgangspunkt vergleichender Analysen des Sprachgebrauchs sein können (vgl. Heyer et al., 2000b).

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 71–83. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`

[1] vgl. `http://wortschatz.uni-leipzig.de`

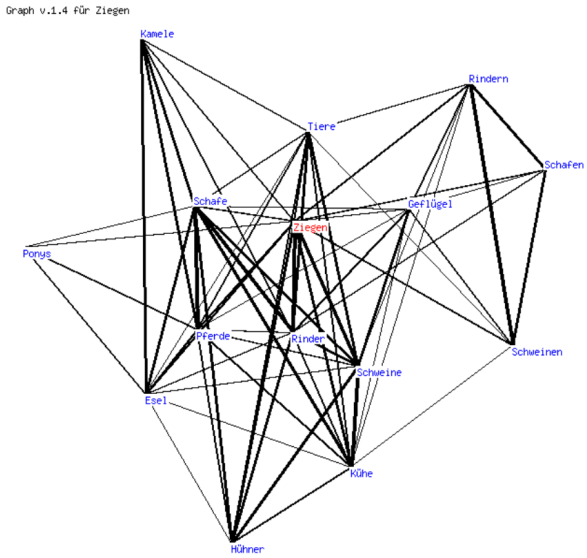

Abbildung 7.1.: Visualisierung der Kollokationen zu „Ziegen“ (deutsches Corpus)

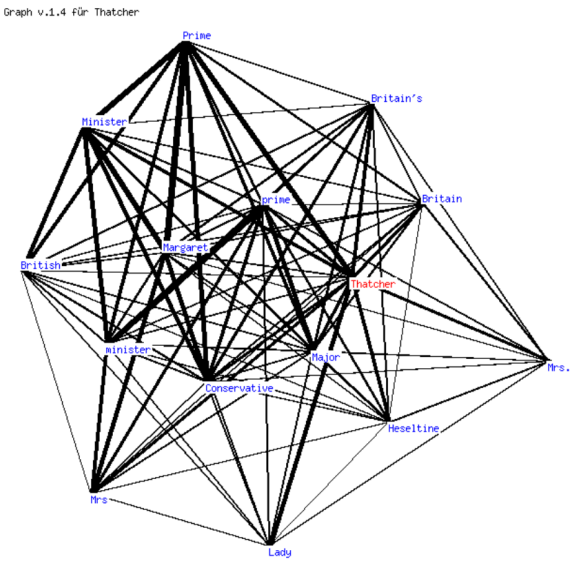

Abbildung 7.2.: Visualisierung der Kollokationen zu „Thatcher“ (englisches Corpus)

7.2.1. Statistische Ermittlung signifikanter Relationen

Aufbauend auf der Erfassung und Analyse der Ausgangstexte werden statistische Auswertungsmethoden vor allem für die Ermittlung signifikanter Kollokationen eingesetzt. Das dabei verwendete Maß greift den aus der Statistik bekannten G-Test auf und liefert Ergebnisse, die dem *mutual information index* vergleichbar sind.[2] Kollokationen werden dabei als signifikantes Vorkommen zweier Muster (Wortformen) in einem gemeinsamen Kontext verstanden. Dabei wird als Kontexteinheit entweder ein Satz („Satzkollokationen") oder die unmittelbare Nachbarschaft zweier Muster (Wortformen) gewählt („Nachbarschaftskollokationen"). Das Maß verhält sich relativ neutral gegenüber der absoluten Häufigkeit des Vorkommens der betrachteten Wortformen, d. h. es ist geeignet, auch für vergleichsweise seltene Begriffe signifikante Kollokationen zu ermitteln. Dabei ist zu betonen, daß die Kollokationsanalyse *automatisch* und *vollständig*, d. h. für alle Token eines Corpus, durchgeführt wird und alle Kollokationsmengen über eine Abfrageschnittstelle ohne weitere Berechnung zur Weiterverarbeitung zur Verfügung stehen. Zusätzlich zur Berechnung von Kollokations*paaren* ist auch die Möglichkeit gegeben, Kollokationsmengen zu n-stelligen Begriffsmengen zu berechnen (vgl. Quasthoff und Läuter, 1999). Neben der Darstellung von Kollokationsmengen als Begriffslisten umfaßt die WWW-Schnittstelle auch eine (*real time-*)Visualisierung von Kollokationsmengen als Begriffsnetzwerk, in dem die stärksten Kollokationen eines Ausgangsbegriffs dargestellt werden, soweit sie auch untereinander in Beziehung stehen (Abb. 7.1 und 7.2). Hierfür kommen bewährte Algorithmen für die planare Darstellung von Graphen zum Einsatz (vgl. Davidson und Harel, 1996).

7.2.2. Offene Probleme

Die durch die statistischen Verfahren ermittelten Kollokationsmengen lassen sich zwar *unmittelbar* nutzen, um z. B. im Rahmen einer Suchmaschine zusätzliche Rechercheterme zu liefern (*query expansion*) oder – in Form der Visualisierung – ein für den Leser interpretierbares Begriffsnetz zu beliebigen Konzepten zu generieren, es ergeben sich aber Probleme qualitativer wie quantitativer Natur: Zum einen steigt die Anzahl signifikanter Kollokationen mit der absoluten Häufigkeit des Begriffs, was eine weitere *Filterung* sinnvoll erscheinen läßt. Zum anderen sagt das statistische Maß nichts über den Typ der Beziehung zwischen zwei Begriffen, die ein Kollokationspaar bilden, aus (Problem der Benennung von Kanten in Begriffsnetzwerken). Dies ist ein starkes Indiz für den Einsatz zusätzlichen Wissens, um umfangreiche Kollokationsmengen nach verschiedenen Kriterien filtern zu können. Am Beispiel der Kollokationsmengen zu Auto sei dies verdeutlicht (Tab. 7.1, 7.2, S. 74 f.).[3] Nachfolgend wird gezeigt, wie sich durch linguistische und statistische Filter sowie eine differenzierende Corpusanalyse solche Ergebnismengen weiterverarbeiten lassen.

7.3. Sekundäre Filterung

Der Ansatz einer *sekundären Filterung* baut einerseits darauf auf, daß in den Corpora weitere linguistische oder lexikalische Informationen zur Verfügung stehen bzw. die Corpora auch

[2] Eine Übersicht zu verschiedenen gebräuchlichen Kollokationsmaßen geben Lemnitzer (1998) und Krenn (2000). Zur corpusbasierten Berechnung von Kollokationen vgl. auch die Websites des Instituts für Deutsche Sprache (IDS, `http://www.ids-mannheim.de/kt/corpora.shtml`) sowie des Linguistic Data Consortium (LDC, `http://www.ldc.upenn.edu/LOL/`).

[3] In den Beispielen stehen die numerischen Werte für die Höhe des Signifikanzmaßes.

Begriff	Auto
Anzahl Satzkollokationen:	3 635
Anzahl signifikante linke Nachbarn:	400
Anzahl signifikante linke Nachbarn:	541

Tabelle 7.1.: Übersicht Kollokationen zu Auto

hinsichtlich bestimmter informationeller Kategorien annotiert sind (z. B. *named entities*)[4], andererseits lassen sich auch musterbasierte statistische Verfahren für einen zweiten Analyse- bzw. Filterungsschritt nutzen. Nachfolgend sollen folgende Filterungsverfahren unterschieden werden:

- *Kategorienfilter*, bei denen aus Kollokationsmengen diejenigen Teilmengen extrahiert werden, deren Mitglieder jeweils eine bestimmte grammatikalische Kategorie aufweisen,
- Filterung durch *named entities* bzw. bekannte semantische Attribute,
- *Kollokationsfilter*, bei denen Kollokationsmengen für mehrere Begriffe analysiert werden und
- *vergleichende Analyse* von Begriffen bei Vorliegen unterschiedlicher Teilcorpora.

7.3.1. Kategorienfilter

Unter einem Kategorienfilter ist die Auswahl derjenigen Mitglieder einer Kollokationsmenge zu verstehen, die derselben grammatikalischen Kategorie angehören. Voraussetzung ist dabei, daß entsprechende Informationen für eine hinreichende Menge types im Corpus zur Verfügung stehen. Es existieren Abfrageschnittstellen für

- die Ermittlung von Adjektiven als linker Nachbarn zu einem Substantiv,
- die Ermittlung von Verbformen als rechte Nachbarn zu einem Nomen und die
- Ermittlung von Substantiven, die zusammen mit einem Adjektiv auftreten

Tab. 7.3 (S. 76) zeigt die so ermittelten Adjektive und Verbformen zum Nomen Politiker sowie die zusammen mit dem Adjektiv (Adverb) offensichtlich gebrauchten Nomen gezeigt.

Interpretiert man diese Teilmengen, so lassen sich Adjektive als typische *Eigenschaften* und Verben als typische *Tätigkeiten* zu Begriffen bzw. Konzepten auffassen. Dies ist bereits bei Betrachtung eines unspezifischen allgemeinsprachlichen Corpus von Interesse. Die Beschreibungsgenauigkeit läßt sich noch weiter erhöhen, wenn als Grundlage der Analyse ein fachspezifischer Corpus verwendet wird. Auf diesem Weg kann z. B. eine solche Analyse einer umfangreichen

[4] Insofern nehmen die beschriebenen Textcorpora eine Zwitterrolle ein, da sie neben den eigentlich Corpusdaten auch lexikalische Informationen (Häufigkeiten, grammatikalische Kategorien, Sachgebietsangaben) bereitstellen.

Satzkollokationen für „Auto" (top/last 20)	linke Nachbarn von „Auto" (top/last 10)	rechte Nachbarn von „Auto" (top/last 10)
ein (2115), fahren (1396), mit (1283), einem (1279), dem (1237), das (1227), Wagen (979), prallte (914), Fahrer (809), seinem (723), fuhr (709), fährt (638), Polizei (609), erfaßt (587), auf (558), gefahren (485), verletzt (472), geparktes (471), Straße (454), Fahrbahn (440), [...] zugestürmt (5), zuparken (5), zurückgelegten (5), zurücklegt (5), zurückzulegen (5), zusammenzustoßen (5), zuverlässig (5), zünden (5), zündete (5), Ältere (5), Älteren (5), ÖPNV-Angebot (5), Öffentlichem (5), Öffentlichen (5), Ökomobil (5), Überdachte (5), Überfalls (5), Überholer (5), öffnete (5), übersteuerte (5)	das (2731), dem (2054), ein (1779), seinem (1173), einem (1135), sein (870), im (857), geparktes (556), ihr (500), Das (446), [...] sparsamen (3), streift (3), tolles (3), ungewöhnliches (3), unterm (3), verdächtigen (3), verlassenen (3), vermißte (3), versicherten (3), wessen (3)	erfaßt (702), fahren (565), angefahren (402), fährt (192), gefahren (158), gezerrt (141), überfahren (138), geschleudert (128), unterwegs (128), gestohlen (125), [...] umzugehen (3), unentbehrlich (3), verunglückten (3), vorziehen (3), wartenden (3), wegnehmen (3), wiederfinden (3), zuverlässiger (3), übernachten (3), überquert (3)

Tabelle 7.2.: Satz- und Nachbarschaftskollokationen zu „Auto“ (deutsches Corpus)

Adjektive als signifikante linke Nachbarn zu „Politiker“	Verbformen als signifikante rechte Nachbarn zu „Politiker“	Nomen als signifikante rechte Nachbarn zu „offensichtlich“
verantwortlichen (428), führende (404), Bonner (399), konservative (283), führenden (240), deutsche (217), deutscher (198), beide (177), prominente (157), führender (153), hochrangige (146), viele (138), westliche (114), prominenter (98), liberale (91), korrupte (87), westlicher (85), grüne (78), westlichen (76), sozialdemokratische (73), vieler (68), populärste (64), amerikanischer (60), britischer (60), sozialistische (59), ostdeutsche (55), italienischer (55), meisten (54), zahlreiche (53), französischer (52)	sollten (75), fordern (66), hätten (58), müßten (38), forderten (37), reagierten (32), sprachen (27), antworten (23), reden (21), streiten (20), glauben (20), denken (15), scheinen (15), warnten (15), appellierten (15), betonten (14), mögen (14), äußerten (14), befürchten (14), warfen (13), wollten (13), sprechen (13), verwickelt (13), fürchten (12), widersprachen (11), kritisierten (11), mahnen (11), appelliert (11), plädieren (11), kritisieren (11)	Benachteiligung (28), Widerspruch (23), Fehlern (21), Fehler (20), Gründen (20), Unfähigkeit (18), Unrecht (17), Probleme (13), Verstoß (13), Demokratiedefizit (12), Fehlentscheidungen (12), Schwäche (12), Zahlungsunfähigkeit (12), Diskrepanz (11), Ignoranz (11), Mißachtung (11), Mißbrauchs (11), Mißstände (11), Desinteresse (10), Diskriminierung (10), Vorteile (10), Widersprüche (10), Angst (9), Bemühens (9), Versuch (9), Wirkungslosigkeit (9), Fehlentscheidung (8), Schwachstelle (8), Torchance (8), Uneinigkeit (8), Ungerechtigkeiten (8)

Tabelle 7.3.: Beispiele für Kategorienfilter (Adjektive, Verbformen und Nomen als unmittelbare Nachbarn)

Software-Dokumention Hilfsmittel beim Software-Reengineering sein, da für den zu modellierenden Gegenstandsbereich automatisch Verben (Methoden im Sinne der Objektorientierung) und Adjektive (Eigenschaften im Sinne der Objektorientierung) ermittelt werden können (vgl. dazu auch sprachbasierte Ansätze im Software Engineering, Ortner, 1997, und Heyer et al., 2000a).

7.3.2. Filterung mit Hilfe von „Named entities" und semantischen Relationen

Grammatikalische Kategorien eignen sich für die Filterung, da sie nicht an bestimmte Domänen gebunden sind und diese Art der Information in ausreichendem Maße zur Verfügung steht. Daneben kann aber prinzipiell jede andere informationelle Kategorie verwendet werden, sei es, daß sie aus einem externen Wissensspeicher zur Verfügung steht (z. B. Zuordnung von Begriffen zu Sachgebieten in einer Ontologie), sei es, daß sie aus dem Corpus selbst ermittelt werden kann (z. B. durch automatische Erkennung von Eigennamen). Für den vorliegenden deutschen Corpus wurden typische Eigennamen auf der Basis von Vor- und Nachnamenslisten (teilautomatisch) ermittelt. Mit Hilfe dieser Daten lassen sich etwa Relationen mit der Struktur *generische Eigenschaft – typische Vertreter* extrahieren. Für die Relation *Berufsbezeichnung (Klasse) – Vertreter des Berufs (Instanz)* zeigt dies Tab. 7.4 (S. 78) mit den Eigennamen in der Kollokationsmenge zum Begriff *Liedermacher*.

7.3.3. Kollokationsfilter

Eine dritte Filterungsmöglichkeit besteht in der Auswertung von Kollokationsmengen zu unterschiedlichen Begriffen. Durch Bildung der Schnittmenge der Kollokationsmengen zu zwei Begriffen lassen sich Begriffe extrahieren, die zu *beiden* Ausgangsbegriffen in einer signifikanten Beziehung stehen. Diese Vorgehensweise läßt sich auch als Ermittlung eines Platzhalters in einem zweistelligen Prädikat auffassen (*Prädikat(Argument1, Argument2)*), bei dem wahlweise das Prädikat selbst oder eines der beiden Argumente als Platzhalter aus den Kollokationsmengen ermittelt werden sollen. Es ist offensichtlich, daß diese Form der Informationsextraktion z. B. eine Hilfestellung für Frage-Antwort-Systeme geben kann. Das nachfolgende Beispiel soll dies für das Beispiel *Oberbürgermeister(Name, Stadt)* verdeutlichen: Es ergeben sich drei Anfragevarianten, deren Ergebnisse (als Kollokationsschnittmengen) in Tab. 7.5, S. 80 dargestellt sind.

Anders als bei einer deklarativen Wissensbasis und einem über ihr operierenden Inferenzmechanismus besteht hier nicht die Sicherheit, daß in allen Fällen ein korrektes Ergebnis geliefert wird; zudem umfassen die Ergebnismengen jeweils mehrere, nach Signifikanz geordnete Einträge. In der Regel wird die „korrekte Anwort" auch die höchste Signifikanz aufweisen oder wenigstens bei Einschränkung auf eine bestimmte Kategorie (z. B. Nomen, Eigennamen) einen vorderen Listenplatz einnehmen. Das Beispiel zeigt zudem, daß eine wiederholte Anwendung von Filtern sinnvoll sein kann: Ist etwa bekannt, welcher Kategorie der Platzhalter angehören muß, so würde etwa in der mittleren Spalte der Eigenname nach vorne rücken. In der letzten Spalte erhält man indirekt auch Antwort auf die Frage nach der Parteizugehörigkeit.

Dasselbe Verfahren – Schnittmengenbildung von Kollokationsmengen – kann auch angewandt werden, um bei mehrdeutigen Begriffen die Kollokationsmenge hinsichtlich der verschiedenen Bedeutungsvarianten zu segmentieren. Voraussetzung ist dann allerdings, daß innerhalb der Kollokationsmenge des Ausgangsbegriffs je ein Begriff bestimmt werden kann, der repräsentativ für eine Bedeutungsvariante ist. Die Menge der Kollokationen zu *Zylinder* umfaßt (u. a.)

Eigenname	Kollokationssstärke
Konstantin Wecker	235
Wolf Biermann	188
Hans Söllner	40
Gerhard Schöne	35
Hannes Wader	28
Clemens Bittlinger	27
Stephan Krawczyk	27
Reinhold Andert	23
Ludwig Hirsch	21
Stephan Krawcyk	17
Sergio Vesely	16
Piet Atten	12
Reinhard Lakomy	12
Oliver Ziegler	11
Georg Danzer	11
Siggi Liersch	11
Herbert Stettner	11
Alexander Dolsky	11
Klaus Neuhaus	11
Franz Josef Degenhardt	11

Tabelle 7.4.: Eigennamen als signifikante rechte Nachbarn zu „Liedermacher“

signifikante Kollokationen aus den Bedeutungsvarianten *Kopfbedeckung* und *Teil eines Verbrennungsmotors*, die beiden stärksten Kollokationen sind *Frack* und *Kolben*. Bildet man jeweils die Kollokationsschnittmengen zu *Zylinder* und *Frack* bzw. *Zylinder* und *Kolben*, so läßt sich die Gesamtmenge der Kollokationen zu *Zylinder* in bedeutungsbezogene Teilmengen separieren (Tab. 7.6, S. 81).

7.3.4. Corpusvergleich

Eine weitere Filterungsmöglichkeit besteht im Vergleich der Begriffsverwendung mit Bezug auf unterschiedliche Teilcorpora, insbesondere der Vergleich von allgemeinsprachlichen und fachsprachlichen Corpora. Neben der Analyse von Kollokationsmengen kann bereits der Vergleich von *Häufigkeitsklassen* Hinweise auf den Status eines Begriffs ergeben und damit z. B. die automatische Extraktion von Fachterminologie unterstützen. Tab. 7.7 (S. 81) gibt einen Vergleich von Häufigkeitsklassen für ausgewählte Fachbegriffe mit Bezug zum allgemeinsprachlichen Corpus bzw. zu einem Fachcorpus (mehrere Jahrgänge einer Automobilfachzeitschrift). Eine Differenz von 8 Häufigkeitsstufen entspricht dabei einem Faktor 256 (d. h. 2^8) bei den absoluten Häufigkeiten (Tab. 7.7, S. 81).

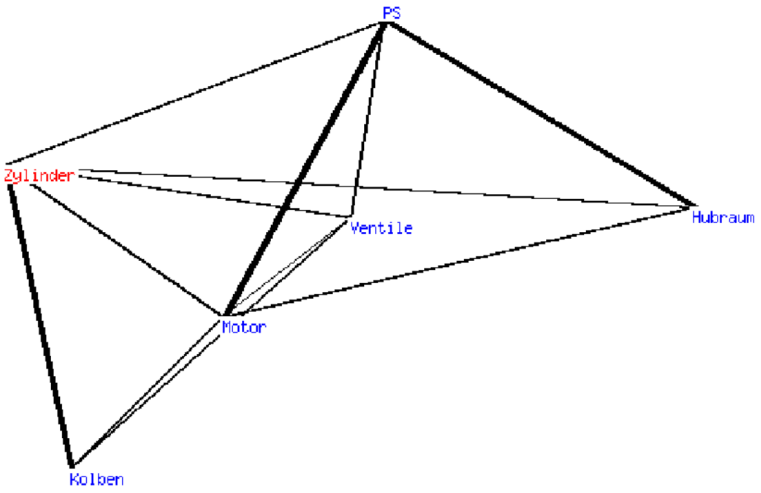

Abbildung 7.3.: Visualisierung von „Zylinder“ (allgemeinsprachliches Corpus)

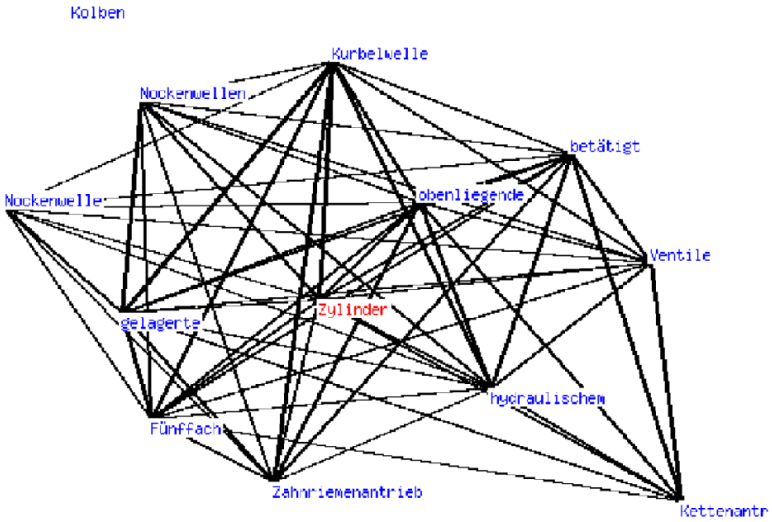

Abbildung 7.4.: Visualisierung von „Zylinder“ (fachsprachliches Corpus)

X(Christian Ude, München)	Oberbürgermeister(X München)	Oberbürgermeister (Christian Ude, X)
Oberbürgermeister (3496), OB (2139), Bayern (1185), bayerische (534), Oberbayern (489), Landeshauptstadt (476), CSU (449), Stadt (446), Bayerns (376), bei (373)	in (8345), Christian Ude (3496), Frankfurt (1075), nach (1024), Stuttgart (853), Bayerischen (832), am (820), bayerischen (744), Köln (732), Stadt (552)	OB (2139), SPD (887), Stadt (552), Peter Menacher (263), Rathaus (240), CSU (206), Münchens (200), hat (175), Münchner (173), Stadtrat (157)

Tabelle 7.5.: Schnittmengenbildung zur Bestimmung von Variablen in semantischen Relationen

Neben eine solche numerische Auswertung kann dabei auch die Interpretation der für jeden Begriff und jedes Corpus generierten Visualisierungen der Kollkationsmengen treten, die im folgenden Beispiel für *Zylinder* eine wesentlich stärkere Feindifferenzierung bei dem Begriffsnetz für das Fachcorpus ergibt.[5]

7.4. Anwendungen

Die voranstehenden Überlegungen und Beispiele zum Postprocessing von Kollokationsmengen sind methodisch als Versuch zu bewerten, corpusbasierte Verfahren der Wissensextraktion mit Hilfe statistischer und linguistischer Methoden einzuführen. In der Darstellung sind bereits Anwendungsmöglichkeiten für solche Verfahren genannt worden. Sie sollen hier abschließend noch einmal verdeutlicht werden.

7.4.1. Suchmaschinen

Da die Kollokationsmengen online für beliebige Begriffe abgefragt werden können, eignen sie sich zur Rechercheunterstützung bei Suchmaschinen. Da aus empirischen Studien bekannt ist, daß in der Regel wenige und zu unspezifische Begriffe Verwendung finden (vgl. Wolff, 2000), kann sowohl die textuelle Ausgabe einer Liste relevanter Kollokationen als auch die Anzeige (Visualisierung eines Begriffsnetzes) bei der Anfragepräzisierung hilfreich sein. Für diese Anwendung wurde eine Schnittstelle im WWW entwickelt, bei der zu jedem Suchbegriff automatisch Kollokationen, ein Begriffsnetz sowie weitere relevante Informationen (z. B. Synonyme) angezeigt werden, die sich ergänzend in die Anfragemaske übertragen lassen. Weitergehend ist auch an die Nutzung von Kollokationsschnittmengen für Frage-Antwort-Systeme wie *Ask Jeeves* zu denken, bei denen bisher einfache *keyword spotting*-Verfahren vorherrschen (Breck, 2000, Voorhees und Tice, 2000).

[5] Interessant ist dabei, daß sich die bekannten Bedeutungsvarianten für Zylinder im allgemeinsprachlichen Corpus (*Kopfbedeckung; Arbeitsgerät im Labor; mathematischer Körper*) nicht niederschlagen, da sie offensichtlich hinter der Begriffsvariante *Teil eines Verbrennungsmotor* auch allgemeinsprachlich deutlich zurückstehen.

Kollokationsschnittmenge „Zylinder" und „Frack"	Kollokationsschnittmenge „Zylinder" und „Kolben"
im (97), und (74), mit (61), Herren (46), schwarzen (31), Damen (31), ein (27), schwarzem (25), Vorzeigereiter (25), trägt (24), tragen (23), Bühne (20), in (19), bewegt (17), Hose (15), Cut (13), Herr (13), Augenglas (11), schwarzer (10), roter (10)	und (72), mit (61), Ventile (58), einem (54), Motor (43), Dampf (35), aus (27), ein (27), Luft (27), Frischgase (26), Auslasskanal (23), Schwungrad (22), einen (22), Ventil (22), bewegt (22), Kugel (20), den (19), Zylinders (18), strömt (17), Maschine (16)

Tabelle 7.6.: Kollokationsschnittmengen zur Separierung von Bedeutungen (jeweils top 20)

Fachbegriff	Häufigkeitsklasse im Fachcorpus	Häufigkeitsklasse im allgemeinsprachlichen Corpus	Differenz
Hubraum	6	14	8
Nockenwelle	9	18	9
Fahrgeräusch	11	19	8
Zylinder	8	13	5

Tabelle 7.7.: Vergleich von Fachtermen mit Bezug auf ein allgemeinsprachliches und ein fachbezogenes Corpus

7.4.2. Lexikographie und Sprachenlernen

Ein weiteres Anwendungsfeld für die Verarbeitung von Kollokationsmengen ist die deskriptive Lexikographie und damit verbunden das Sprachenlernen: Zum einen lassen sich für umfangreiche Begriffsmengen die allgemeinsprachlich relevanten Beziehungen ermitteln und durch das linguistische Postprocessing nach Kategorien ordnen. Für den angelsächsischen Bereich liegen entsprechende Kollokationslexika bereits seit einiger Zeit vor (vgl. Benson et al., 1993). Die hier gewählte Vorgehensweise hat den Vorteil, daß durch die verschiedenen Softwareschnittstellen sowie durch den Zugang im WWW gegenüber einem traditionellen Lexikon vielfältigere Nutzungsmöglichkeiten gegeben sind. Ihre Anwendung kann unter anderem auf dem Gebiet des Sprachenlernens liegen, da typische Assoziationen zu Begriffen dargestellt werden können und diese vom Sprachlerner auch zur Bereinigung von Unklarheiten (Begriffswendung, gebräuchliche Adjektiv-Nomen-Kombinationen etc.) genutzt werden können. Voraussetzung ist dabei die gezielte Auswahl hochwertiger und repräsentativer Quellen. Einzelne Fehler in den Quellen wir-

ken sich aufgrund des statistischen Ansatzes allerdings kaum aus (z. B. orthographische Fehler in Zeitungstext).

7.4.3. Semiometrie und Terminologieextraktion

Ein drittes Anwendungsbeispiel sind Semiometrie und Terminologieextraktion. Im einfachsten Fall kommt dabei die bereits gezeigte häufigkeitsbezogene Ermittlung signifikanter Fachbegriffe zum Einsatz. Durch die Analyse von Kollokationsmengen lassen sich zu Begriffen aber auch typische Eigenschaften bestimmen, wie sie sich in den Ausgangstexten widerspiegeln. Dabei kann man grundsätzlich auch die temporale Dimension berücksichtigen, d. h. als weiterer Parameter der Filterung kann der Entstehungszeitraum der Textquellen Berücksichtigung finden. Damit erhält man Teilcorpora, die für identische Begriffe jeweils unterschiedliche Kollokationsmengen aufweisen und damit Bedeutungsentwicklungen von Begriffen nachweisen.

7.5. Fazit

Der vorliegende Beitrag zeigt einige Ansatzpunkte für das Postprocessing von mit Hilfe statistischer Verfahren extrahierter Begriffsrelationen (Kollokationen) aus großen Textcorpora. Ein wesentliches Anliegen ist es dabei, zu zeigen, wie sich unterschiedliche Wissensquellen und Extraktionsverfahren bei der Analyse von Texten kombinieren lassen. Dabei gelten als weitere methodische Randbedingungen einerseits das Ziel, diese Analysen vollständig, d. h. auch bei sehr großen Corpora für alle auftretenden Begriffe durchführen zu können, und andererseits die weitestgehende Automatisierung der Analyseverfahren. Es konnte gezeigt werden, daß solche Analyseverfahren für eine Vielzahl praktischer Anwendungen geeignet sind.

Literaturverzeichnis

Benson, M.; Benson, E. und Ilson, R. (1993): *The BBI Combinatorial Dictionary of English*. Amsterdam, Philadelphia: John Benjamin.

Breck, E. J. (2000): "How to Evaluate Your Question Answering System Every Day . . . and Still Get Real Work Done". In: *Proc. LREC-2000. Second International Conference On Language Resources and Evaluation*. Athen, Band 3, S. 1495–1500.

Davidson, R. und Harel, D. (1996): "Drawing Graphs Nicely Using Simulated Annealing". *ACM Transactions on Graphics* 15 (4): S. 301–331.

Heyer, G.; Läuter, L.; Quasthoff, U. und Wolff, C. (2000a): "Texttechnologische Anwendungen für Inter- und Intranet". In: *Sprachtechnologie für eine dynamische Wirtschaft im Medienzeitalter. Tagungsakten der XXVI. Jahrestagung der Internationalen Vereinigung Sprache und Wirtschaft, Köln, November 2000*, herausgegeben von Schmitz, K.-D., Wien: TermNet, S. 203–209.

Heyer, G.; Quasthoff, U. und Wolff, C. (2000b): "Aiding Web Searches by Statistical Classification Tools". In: *Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposium für Informationswissenschaft, ISI 2000, Darmstadt*, herausgegeben von Knorz, G. und Kuhlen, R. Konstanz: UVK, S. 163–177.

Krenn, B. (2000): "Distributional and Linguistic Implications of Collocation Identification". In: *Proc. Collocations Workshop, DGfS Conference*. Marburg.

Lemnitzer, L. (1998): "Komplexe lexikalische Einheiten in Text und Lexikon". In: *Linguistik und neue Medien*, Wiesbaden: Dt. Universitätsverlag, S. 85–91.

Ortner, E. (1997): *Methodenneutraler Fachentwurf*. Stuttgart, Leipzig: Teubner.

Quasthoff, U. (1998a): "Projekt der deutsche Wortschaft". In: *Linguistik und neue Medien*, Wiesbaden: Dt. Universitätsverlag, S. 93–99.

Quasthoff, U. (1998b): "Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values". In: *Proc. First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998*. Band 2, S. 853–856.

Quasthoff, U. und Läuter, M. (1999): "Kollokationen und semantisches Clustering". In: *Multilinguale Corpora. Codierung, Strukturierung, Analyse*, herausgegeben von Gippert, J. Prag: Enigma, S. 34–41.

Quasthoff, U. und Wolff, C. (2000): "An Infrastructure for Corpus-Based Monolingual Dictionaries". In: *Proc. LREC-2000. Second International Conference On Language Resources and Evaluation, Athen, May/June 2000*. Band 1, S. 241–246.

Rieger, B. (2000): "Computing Granular Word Meanings. A fuzzy linguistic approach in Computational Semiotics". In: *Computing with Words*, herausgegeben von Wang, P. P., New York: Wiley. Im Erscheinen.

Ruge, G. (1994): *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Hildesheim, New York: Olms.

Voorhees, E. M. und Tice, D. M. (2000): "The TREC-8 Question Answering Track". In: *Proc. LREC-2000. Second International Conference On Language Resources and Evaluation, Athen, May/June 2000*. Band 3, S. 1501–1508.

Wolff, C. (2000): "Vergleichende Evaluierung von Such- und Metasuchmaschinen im World Wide Web". In: *Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposium für Informationswissenschaft, ISI 2000, Darmstadt*, herausgegeben von Knorz, G. und Kuhlen, R. Konstanz: UVK, S. 31–48.