Computer- und korpuslinguistische Verfahren für die Analyse massenmedialer politischer Kommunikation: Humanitäre und militärische Interventionen im Spiegel der Presse

Peter Kolb[§], Amelie Kutter[§§], Cathleen Kantner[§§], Manfred Stede[§]

Zusammenfassung. *Dieser Beitrag zeigt am Beispiel eines aktuellen Forschungsprojekts auf, welche Potenziale computer- und korpuslinguistische Verfahren für eine multilinguale politikwissenschaftliche Medientext-Analyse bieten. Wir führen in die Forschungsfragen ein, die in der an der FU Berlin angesiedelten Medienanalyse zu militärischen und humanitären Interventionen bearbeitet werden, erläutern die Erfahrungen, die dabei mit vorhandenen computerlinguistischen Verfahren bisher gemacht wurden und diskutieren die Möglichkeiten, die innovative Werkzeuge der Computerlinguisten der Universität Potsdam erschließen.*

1. Einleitung

Mit der Verfügbarkeit umfangreicher elektronischer Daten eröffnet sich für Sozialwissenschaftler die Möglichkeit, ihre Fragestellungen auf größere Datenmengen anzuwenden, die Erfassung und Analyse teilweise zu automatisieren und den Aussagen damit größere (quantifizierte) Validität zu verleihen. Zugleich ergeben sich neue technische Herausforderungen für Datenmanagement und -analyse, die sich nur mit Hilfe der Computerlinguistik lösen lassen. Dieser Beitrag zeigt am Beispiel eines aktuellen Forschungsprojekts auf, welche Potenziale computer- und korpuslinguistische Verfahren für eine multilinguale politikwissenschaftliche *Medientext-Analyse* bieten. Wir führen in die Forschungsfragen ein, die in der an der FU Berlin angesiedelten Medienanalyse zu militärischen und humanitären Interventionen bearbeitet werden, erläutern die Erfahrungen, die dabei mit vorhandenen computerlinguistischen Verfahren bisher gemacht wurden und diskutieren die Möglichkeiten, die innovative Werkzeuge der Computerlinguisten der Universität Potsdam erschließen.

2. Die sozialwissenschaftliche Fragestellung

Im Zentrum der an der FU Berlin durchgeführten Medienanalyse zu militärischen und humanitären Interventionen steht die Frage, ob die unterschiedlichen nationalen Medienöffentlichkeiten in der Europäischen Union (EU) *Problemsichten* teilen, ähnliche politische Akteure – insbesondere die EU – als *Handlungsträger* sehen und ähnliche *normative Kriterien* zur Beurteilung von politischen Problemen heranziehen. Eine solche „Problemlösungsgemeinschaft“, so die grundlegende Annahme, würde die grenzübergreifende Diskussion und Entscheidungsfindung zu politischen Problemen erleichtern, die sich zunehmend auf EU-Ebene verlagern. Auch im Bereich der äußeren Sicherheit, darunter der militärischen und humanitären Interventionen, hat sich die EU als (zusätzliches) Steuerungszentrum etabliert. Auf welchen Grundlagen sie aber in Aktion treten soll ist umstritten und bedarf der politischen Kommunikation und Selbstverständigung innerhalb und zwischen den nationalen Öffentlichkeiten. Mediendebatten zu militärischen Interventionen eignen sich daher als Gegenstand, anhand dessen die mögliche Herausbildung einer transnationalen Problemlösungsgemeinschaft untersucht werden kann. Die Dekade nach dem

§ Universität Potsdam, Institut für Linguistik, AG Angewandte Computerlinguistik, Karl-Liebknecht-Str. 24-25, 14476 Golm. {kolb|stede}@ling.uni-potsdam.de

§ § Freie Universität Berlin, FB Politik und Sozialwissenschaften, Otto-Suhr-Institut für Politikwissenschaft, Arbeitsstelle Europäische Integration, Ihnestr. 22, 14195 Berlin. {kantnerc|akutter}@zedat.fu-berlin.de

Ende des Kalten Kriegs bietet sich als Untersuchungszeitraum an, da in dieser Zeit auch auf dem europäischen Kontinent gewaltsame Konflikte ausbrachen und die EU Instrumente für ein gemeinsames Management dieser Konflikte entwickelte.

Grundlage der ländervergleichenden Längsschnittanalyse ist ein bereinigtes Vollsample von 489.500 Zeitungsartikeln, die in den Jahren 1990-2006 in je zwei großen Tageszeitungen in acht Ländern erschienen. Sie wurden mit Hilfe einer komplexen Suchanfrage aus Medienarchiven wie LexisNexis und aus den Archiven einzelner Zeitungen gewonnen. Die Untersuchungsländer sind: Deutschland, Frankreich, Großbritannien, Irland, Niederlande, Österreich, Polen und – als außereuropäischer Vergleichsfall – die USA. Angesichts dieser Datenmenge ergaben sich zwei wesentliche Herausforderungen: einerseits eine automatisierte Aufbereitung der Rohdaten für ein auf die Forschungsfragen zugeschnittenes dynamisches Datenmanagement; andererseits die automatische Erfassung von semantischen Teilmengen, die entweder entfernt oder aber für nähere Analysen herangezogen werden sollten (z.B. alle fälschlicherweise erfassten Duplikate und Samplingfehler, Artikel zu Interventionen im engeren Sinne, Artikel mit EU-Referenzen etc.). Computerlinguistische Verfahren waren hierfür unerlässlich.

3. Unterstützung durch computerlinguistische Verfahren

Zwei grundlegende Methoden kamen in der ersten Phase zum Einsatz. Für die Datenaufbereitung nutzten wir die *automatische Musterkennung in (linguistisch nicht spezifizierten) Zeichenketten*, so etwa beim Parsen von Datenbank-relevanten Informationen aus den Rohdaten, für ein Vektorraum-Verfahren, das Dubletten identifizierte, sowie für die Extraktion und Kalkulation von Artikeln, die spezifische Suchworte enthielten. Dies geschah mit Hilfe der Software SPSS Clementine[1]. Zur Bestimmung inhaltlich relevanter Zeichenketten führten wir zusätzlich eine *Konkordanzanalyse* mit WordSmith[2] durch. Durch das manuelle Überprüfen der signifikanten Kollokate konnten wir bspw. bestimmen, in welcher Kombination mehrdeutige Begriffe wie „Europa“, „europäisch“, „Europäer“ eindeutig der EU zuzurechnen waren. Alle diese Verfahren erforderten jedoch viel Programmierung und inhaltliche manuelle Arbeit – sei es bei der Bestimmung, wann ein Artikel eine Dublette sei und wann eine zu berücksichtigende Wiederveröffentlichung, sei es bei der Disambiguierung von Wortgruppen oder bei der Bestimmung sämtlicher Flexionen eines Wortes. Zudem erforderten sie profunde Sprach- und Grammatikkenntnisse und manuellen Sprachvergleich in fünf Sprachen bzw. acht länderspezifischen Lexiken. Die genutzte Software bot teilweise Erleichterung (bspw. SPSS Clementine mit einem Tool zur Erkennung von „Konzepten“ – Lemmata) war aber so intransparent, dass sie für die Bedürfnisse des Projekts nicht angepasst werden konnte.

3.1 Das Text Mining Tool

In der zweiten Phase wurde dann ein erstes an der Universität Potsdam entstandenes Werkzeug für die Textmenge angepasst und erprobt: das Distributionsanalyse-System DISCO. DISCO ist Bestandteil eines „Text Mining Tools“ (TMT), das unterschiedliche Zugangsweisen zu einem Korpus in einer Anwendung integriert. TMT weist eine Client-Server-Architektur auf. Die gesamten Daten befinden sich auf einem Server-Rechner, der eine Webschnittstelle bereitstellt, über die mittels eines beliebigen Webbrowsers auf TMT zugegriffen werden kann. Dies bietet drei Vorteile: Erstens kann ein Client-Rechner von einem beliebigen Standort über das Internet auf das Tool zugreifen, zweitens spielt das jeweilige Betriebssystem, unter dem der Client läuft, keine Rolle, und drittens muss auf dem Client außer einem Webbrowser keine weitere Software installiert werden.
Das Text-Mining-System TMT umfasst drei Teilkomponenten: die Suchmaschine LiSCo, das Distributionsanalysewerkzeug DISCO und eine sogenannte „Themenmaschine“, die sich noch in der Entwicklungsphase befindet. Diese Bausteine werden im Folgenden beschrieben.

[1] http://www.spss.com/de/clementine/
[2] http://www.lexically.net/wordsmith/version4/

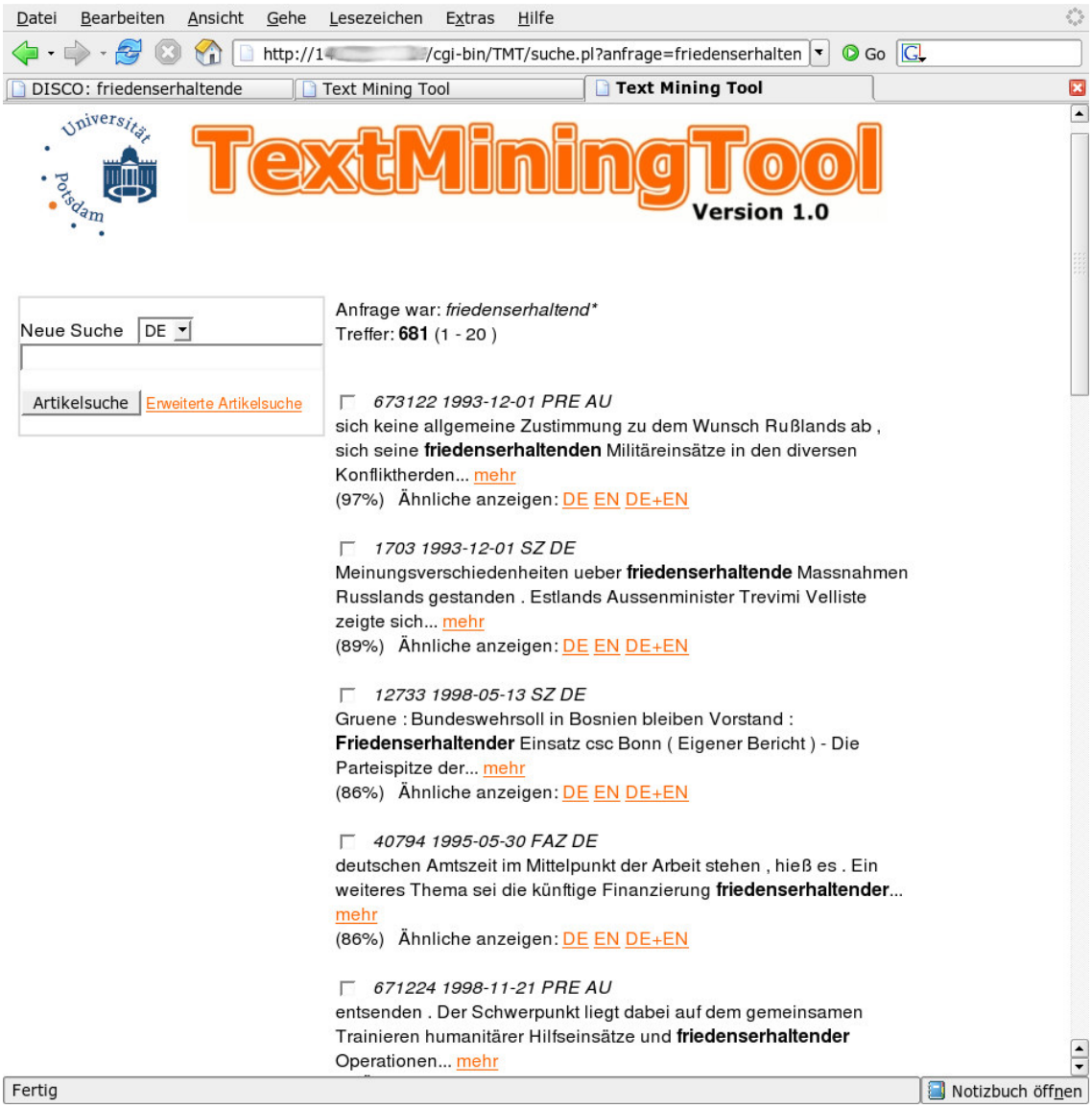


Abbildung 1: Trefferanzeige zur Suchanfrage 'friedenserhaltend'*

3.1.1 LiSCo

Die Suchmaschine LiSCo (Linguistische Suche in Corpora) indexiert das gesamte Korpus und stellt diverse Suchwerkzeuge bereit. LiSCo basiert auf dem Lucene-Index[3], einem in Java implementierten leistungsfähigen Volltextindex, der frei verfügbar ist.
Den ersten Schritt der Indexierung bildet die Aufbereitung und Vorverarbeitung des Korpus. Die Zeitungsartikel wurden mit Metadaten wie Quelle, Datum, Ursprungsland usw. versehen und in einem XML-Format gespeichert. Anschließend wurde mit Hilfe des Tree-Taggers [Schmid1995] ein PoS-Tagging und eine Lemmatisierung durchgeführt. Zum Schluss wurden alle Texte ggf. nach Unicode (UTF-8) konvertiert.
Die Zeitungsartikel wurden dann in den Lucene-Index eingelesen, wobei verschiedene durchsuchbare Felder für jedes Dokument gespeichert wurden. So kann sowohl nach den ursprünglichen Wortformen, als auch nach den Lemmata gesucht werden, außerdem nach Metadaten wie Datum, Land, Quelle usw. Der gesamte Volltext jedes Artikels wurde ebenfalls in den Index aufgenommen.

[3] http://lucene.apache.org

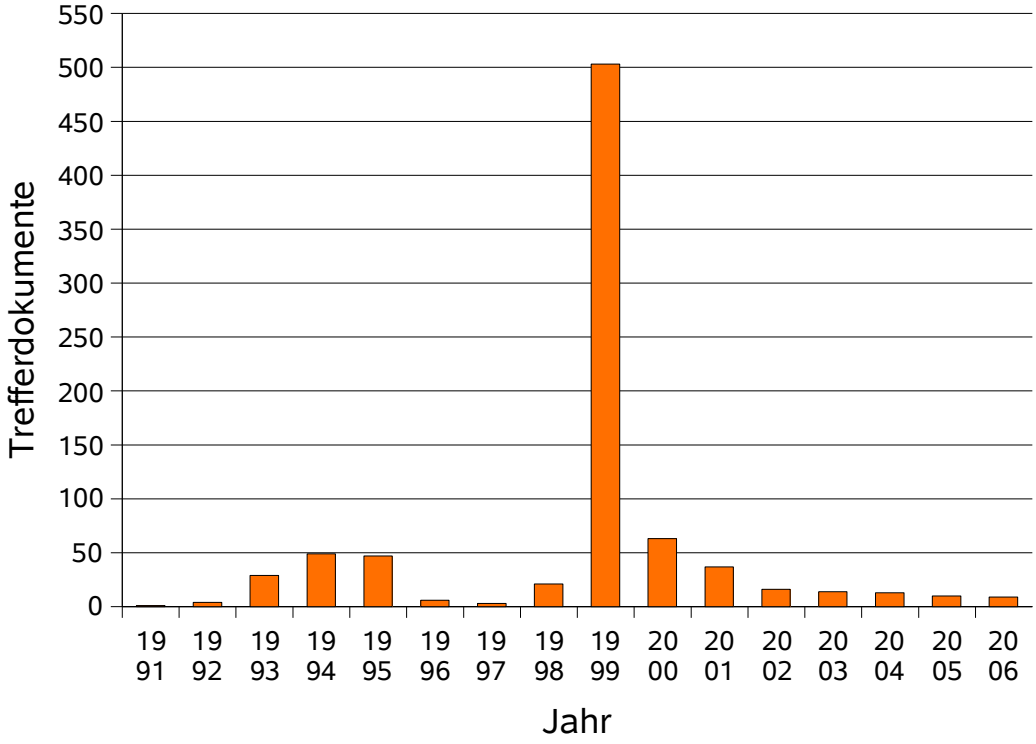


Abbildung 2: Anfrage: (Luftkrieg OR Luftangriffe OR Bombardierung) AND Nato AND (Serbien OR Serbiens)

Zur Suche im Index stellt Lucene eine leistungsfähige Abfrage-Syntax bereit. Eine Boolesche Suche mit den Operatoren AND, OR und NOT ist ebenso möglich wie eine Suche mit Wildcards, eine trunkierte Suche (*Tor** findet *Tor*, *Tore*, *Tors*, *Torpfosten*, ...), oder eine exakte Phrasensuche, bei der die Suchanfrage (wie bei Google) in doppelte Anführungszeichen eingeschlossen werden muss.
Lucene bietet außerdem die Möglichkeit, Treffer in ihrem Kontext anzuzeigen und hervorzuheben. Damit lässt sich bereits eine einfache Konkordanzanzeige bereitstellen.
Abbildung 1 zeigt die Trefferanzeige zur Suchanfrage *friedenserhaltend**. Durch einen Klick auf „mehr“ wird der Volltext des jeweiligen Artikels angezeigt.
Da zu jedem Artikel das Datum im Index gespeichert und somit abfragbar ist, konnte in einfacher Weise eine „Zeitleisten“-Funktion implementiert werden, mit deren Hilfe die zeitliche Entwicklung eines Themas in Form von Balkendiagrammen veranschaulicht werden kann. Dazu muss vom Benutzer lediglich der gewünschte Zeitraum und die zeitliche Auflösung (Monate, Jahre) ausgewählt werden. TMT führt dann automatisch eine Reihe von Anfragen aus, in der die eigentliche Suchanfrage mit den Abschnitten des gewünschten Zeitraums kombiniert und die Anzahl der Treffer protokolliert und als Tabelle gespeichert wird. Das Ergebnis kann per Mausklick entweder als Excel- oder Latex-Datei gespeichert werden. Abbildung 2 zeigt ein Beispielergebnis für den Zeitraum 1991-2006 mit jährlicher Auflösung.
Lucene implementiert neben einem Standard-Volltextindex auch das sogenannte Vektormodell [Salton1971] des Information Retrievals. Dieses Modell folgt der Annahme, dass Dokumente durch die Häufigkeiten der enthaltenen Terme charakterisiert werden. Im Gegensatz zur Booleschen Volltextsuche, bei der nur geprüft wird, ob ein Dokument ein Suchwort enthält oder nicht, wird beim Vektormodell zusätzlich die Wichtigkeit oder Relevanz des Wortes im jeweiligen Dokument berücksichtigt. Die Treffermenge kann dadurch nach Relevanz sortiert werden. Dazu speichert Lucene zu jedem Term die Auftretenshäufigkeit im jeweiligen

Dokument („Termfrequenz“ TF) sowie seine Häufigkeit in der gesamten Dokumentensammlung („Dokumentenfrequenz“ DF). Aus diesen Angaben kann das bekannte TF-IDF-Maß zur Bestimmung der Relevanz eines Terms berechnet werden. Ein Term ist dabei umso wichtiger, je häufiger er im jeweiligen Dokument vorkommt und je seltener er insgesamt in der Dokumentensammlung auftaucht. Auf dieser Grundlage haben wir ein Relevanzfeedback [Rocchio1971] und eine Suche nach inhaltlich ähnlichen Dokumenten implementiert.
Beim Relevanzfeedback kann der Benutzer die zu einem Suchergebnis angezeigten Trefferdokumente durch Anklicken als relevant oder nicht relevant bewerten, und die Anfrage dann per Mausklick wiederholen. Die Suchanfrage wird automatisch um die relevantesten Terme aus den vom Nutzer als relevant bewerteten Dokumenten erweitert. Terme aus den als irrelevant bewerteten Dokumenten werden aus der Suchanfrage entfernt. Die Terme der automatisch erzeugten Suchanfrage können ausgegeben werden.
Per Mausklick kann auch nach inhaltlich ähnlichen Dokumenten zu einem gegebenen Dokument gesucht werden (siehe Abbildung 1). Dabei wird das Ausgangsdokument durch einen Vektor seiner relevantesten Terme repräsentiert, die mit ihrem TF-IDF-Wert gewichtet sind. Dieser Vektor kann als Suchanfrage an den Lucene-Index geschickt werden, der dann die ähnlichsten Dokumente als Treffer ausgibt. Dieses Verfahren arbeitet für den einsprachigen Fall bereits sehr zufriedenstellend. Für den crosslingualen Fall, also z.B. die Suche nach englischen Dokumenten zu einem gegebenen deutschen Dokument, muss der Vektor der relevantesten Terme erst in die Zielsprache übersetzt werden. Unser erster Ansatz, einfach alle im Wörterbuch aufgeführten Übersetzungsmöglichkeiten in den Zielvektor aufzunehmen, führte zu inakzeptablen Ergebnissen, was sicher auch an der großen inhaltlichen Homogenität des Korpus liegen mag. Hier muss noch ein geeignetes Verfahren zur Disambiguierung der Übersetzungsmöglichkeiten gefunden werden. Wir planen, dafür die Kookkurrenz- bzw. distributionellen Ähnlichkeitswerte zwischen Wörtern, wie sie von DISCO geliefert werden, einzusetzen. Außerdem haben wir bereits vielversprechende erste Experimente zur automatischen Extraktion neuer Wort-Übersetzungen aus den multilingualen Korpora durchgeführt. Dabei werden die signifikanten Kookkurrenten eines Ausgangsworts der Quellsprache mit dem vorhandenen Wörterbuch in die Zielsprache übersetzt und dann mit den Kookkurrenzprofilen aller Wörter der Zielsprache verglichen. Das am Ähnlichsten verwendete Wort der Zielsprache wird dann als Übersetzung des Ausgangswortes vorgeschlagen [Rapp1999].
Ein weiteres in TMT implementiertes Suchverfahren bildet die automatische Kategorisierung von Dokumenten in ein vom Benutzer vorgegebenes Kategorienmodell. Zuerst muss vom Benutzer ein Kategorienmodell (eine Hierarchie in Form eines Baums) erstellt werden. Jede Kategorie wird durch eine Anzahl manuell ausgewählter Dokumente (sogenannter Prototypen) definiert. Eine Kategorie kann bereits durch ein einzelnes Dokument definiert werden. Die automatische Einordnung neuer Dokumente in das Kategorienmodell erfolgt über die zuvor beschriebene Ähnlichkeitssuche. Das neue Dokument wird mit allen prototypischen Dokumenten im Kategorienmodell auf Ähnlichkeit verglichen und in die Kategorie mit den ähnlichsten Dokumenten eingeordnet. Dazu wird das sogenannte *k-nearest-neighbour*-Verfahren [Sebastiani2002] eingesetzt (kNN-Verfahren). Es arbeitet folgendermaßen: das neue Dokument x wird mit allen prototypischen Dokumenten im Kategorienmodell auf Ähnlichkeit verglichen. Die k ähnlichsten Dokumente werden ausgewählt (z.B. k = 20, in Abhängigkeit von der Anzahl der Kategorien). Jedes dieser k Dokumente erhält ein "Stimmrecht", dessen Wert gleich seinem Ähnlichkeitsgrad zum Dokument x geteilt durch seinen Rangplatz in der Ergebnisliste ist. Die Stimmen von Dokumenten, die aus der gleichen Kategorie stammen, werden addiert. Diejenige Kategorie gewinnt, die die meisten Stimmen erhält. Für den Fall k = 1 gewinnt die Kategorie, in der sich das zu x ähnlichste Dokument befindet. Das kNN-Verfahren zeichnet sich durch Robustheit, Geschwindigkeit und eine gute Skalierbarkeit hinsichtlich der Kategorienanzahl aus. Ein großer Vorteil unseres Kategorisierungsverfahrens besteht darin, dass kein eigener

Trainingsschritt erforderlich ist. Dadurch können prototypische Dokumente nach Belieben zu einer Kategorie hinzugefügt oder daraus entfernt werden, und die Auswirkungen werden sofort sichtbar. Prototypen können auch zwischen Kategorien verschoben werden. Dem Benutzer kann beim Aufbau eines Kategorienmodells Hilfestellung gegeben werden. Wenn z.B. ein neues Dokument als Prototyp einer Kategorie hinzugefügt werden soll, kann sofort angezeigt werden, wenn die Ähnlichkeit zu einem prototypischen Dokument einer anderen Kategorie größer ist als zu den übrigen Dokumenten derselben Kategorie. Durch dieses unmittelbare Feedback können qualitativ wesentlich bessere Kategorienmodelle aufgebaut werden.

3.1.2 DISCO

Mit dem Distributionsanalyse-System DISCO [Kolb2008] lassen sich zu einem Suchwort die signifikanten Kookkurrenzen und die distributionell ähnlichen Wörter anzeigen (Abbildung 5). Zudem können auf Grundlage der distributionellen Ähnlichkeit Wortcluster berechnet und graphisch dargestellt werden (Abbildung 3). Die Kookkurrenzen vermitteln einen ersten Eindruck, in welchen Zusammenhängen das Suchwort im Korpus verwendet wird. Auf Basis der Kookkurrenzen berechnet DISCO die Wörter, die im Korpus eine ähnliche Distribution

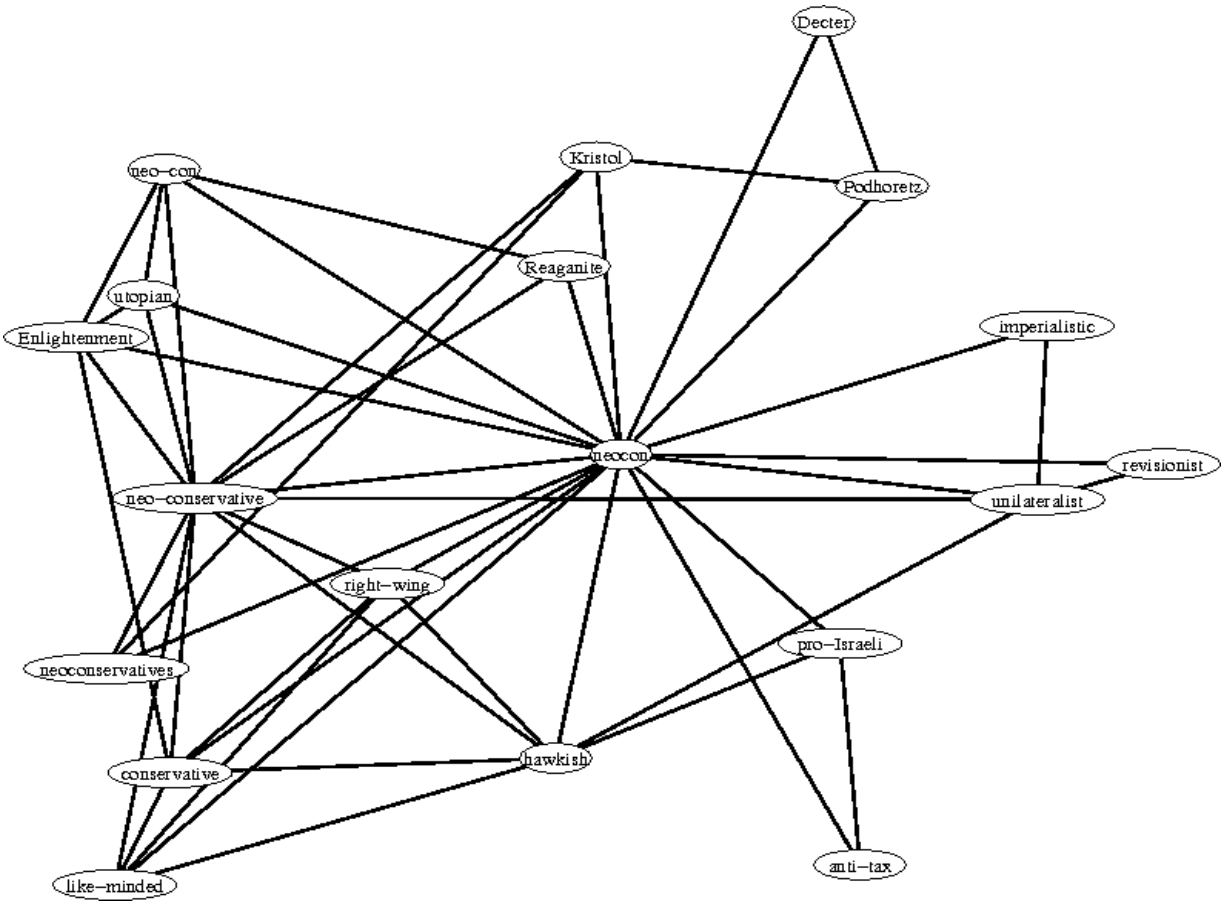


Abbildung 3: Automatisch erzeugter Graph zu 'neocon'

aufweisen. Besonders im Falle abstrakter Nomen erhält man hier semantisch ähnliche Wörter zum Ausgangswort, teilweise das ganze semantische Spektrum. Für das vorliegende, thematisch sehr spezifische Korpus eignete sich DISCO zur schnellen Identifizierung von Wortfeldern von bestimmten Abstrakta (z.B. “Intervention”, “Massenmord” etc.), und zwar in einem Umfang und Tempo, das über WordSmith nie herzustellen gewesen wäre.

3.1.3 Themenmaschine

Um die Tokenebene von DISCO zu verlassen ist geplant, im Korpus automatisch *Themen* in Form relevanter Phrasen zu identifizieren. Diese sollen mit Hilfe linguistischer Analysen zu einer Taxonomie aus Ober- und Unterthemen verknüpft werden, die graphisch als Baum dargestellt wird und navigierbar ist. Die Dokumente, in denen die Themen gefunden wurden, bilden die Blätter des Baums. Zur Unterstützung der Themenvernetzung sollen mit Hilfe syntaktischer Muster aus dem Korpus Wortkandidaten für bestimmte semantische Relationen wie Hyponymie, Meronymie etc. extrahiert werden [Hearst1992]. Außerdem können die von DISCO gefundenen ähnlichen Wörter zur weiteren Vernetzung „assoziierter" Themen genutzt werden.
Das Verfahren zur Themenextraktion arbeitet wie folgt. Durch einen auf dem vorhergehenden PoS-Tagging aufbauenden Analyseschritt werden zunächst Nominalphrasen erkannt. Zum Beispiel werden aus dem Satz

Außenminister Joschka Fischer fordert militärische Intervention

die Phrasen *Außenminister Joschka Fischer* und *militärische Intervention* extrahiert. Im anschließenden Normalisierungsschritt werden verschiedenartige Phrasen auf eine gemeinsame Form gebracht. Dadurch können inhaltliche Übereinstimmungen zwischen unterschiedlich formulierten Textabschnitten erkannt werden. Beispielsweise würden die drei Phrasen

militärische Intervention, Intervention des Militärs, Militärintervention

alle zu *Intervention Militär* normalisiert. Dies erfordert (zumindest im Deutschen) den zusätzlichen Einsatz eines morphologischen Analyseschrittes zur Zerlegung von Komposita in ihre Konstituenten. Als nächstes werden Phrasen und Teilphrasen zusammengefasst. Hierbei würde eine Phrase wie *Bundesaußenminister Fischer* mit der oben aufgeführten Phrase *Außenminister Joschka Fischer* identifiziert werden. Durch diese Art der Zusammenfassung ähnlich formulierter Ausdrücke über den ganzen Text hinweg wird erreicht, dass sich die anschließende Relevanzberechnung deutlich verbessert. In diesem folgenden Arbeitsschritt werden mittels statistischer Auswertung aus Phrasen Themen. Nur diejenigen Phrasen, die einen bestimmten Relevanzwert erreichen, gelangen in den Themenindex. Die Relevanz einer Phrase ist umso größer, je öfter sie oder ihre Teilphrasen im aktuellen Dokument vorgekommen sind, und je niedriger, je öfter die Phrase insgesamt in der indexierten Dokumentensammlung vorgekommen ist. Das Modul Themensuche ermöglicht, dass die Dokumente thematisch erfasst werden. Die Informationen werden dadurch thematisch verknüpft, wie nachfolgende Darstellung veranschaulicht.

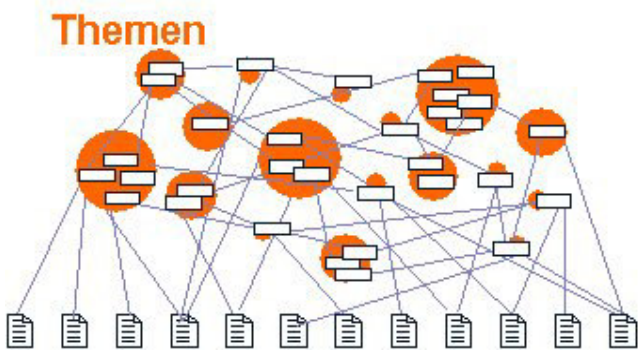


Abbildung 4: Verknüpfung von Themen und Dokumenten

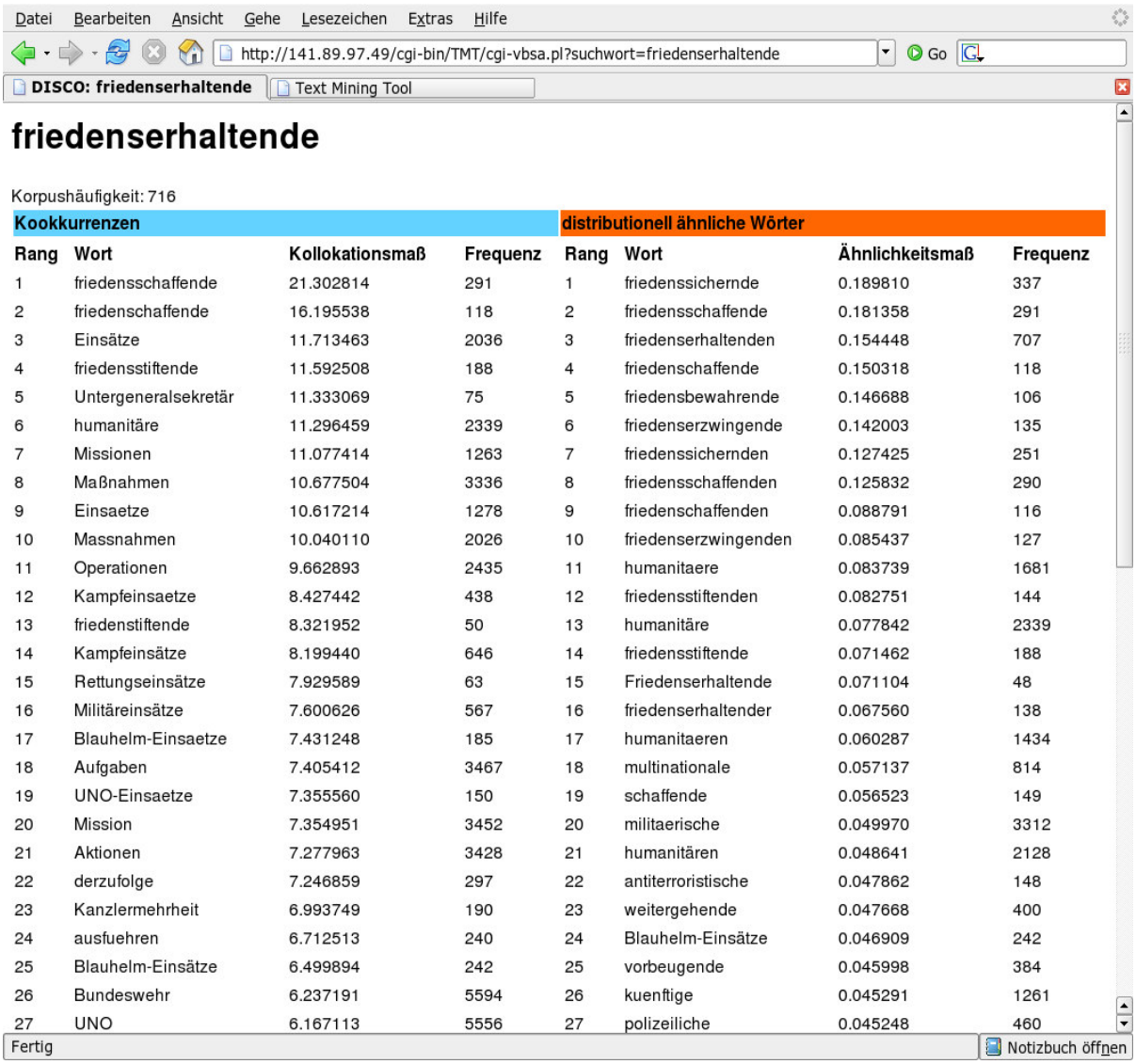
Dieser Prozess vollzieht sich kontinuierlich. Jedes eingehende Dokument wird automatisch nach seinen Themen analysiert und die Verknüpfungen zu den bereits vorhandenen Themen hergestellt. Neue Themen werden erkannt und aufgenommen.

Wir hoffen, dass die Navigation im Themenbaum die mühsame Inspektion von Konkordanzen zur Validierung der jeweils aktualisierten Bedeutung ersetzen können wird.

3.2 Beispielhafte Anwendungsfälle

3.2.1 Erstellung komplexer Suchanfragen mittels semantischer Felder

Eine zu lösende Aufgabe ist die Erstellung von Suchanfragen, mit denen aus dem Korpus diejenigen Zeitungsartikel abgerufen werden können, die sich mit bestimmten, interessierenden Fragestellungen oder Themen befassen. Hierbei kommt es auf eine gleichzeitige Maximierung von Präzision und Vollständigkeit („Recall“) der Suchergebnisse an, damit anschließende quantitative Aussagen auf einer gesicherten Grundlage ruhen. Am Beispiel des Konzepts 'friedenserhaltende Intervention' soll gezeigt werden, wie DISCO bei der Formulierung einer komplexen Suchanfrage helfen kann.
Als erste Suchanfrage dient der Name des Konzepts selbst, also *friedenserhaltende Intervention*.



friedenserhaltende

Korpushäufigkeit: 716

Kookkurrenzen				distributionell ähnliche Wörter			
Rang	Wort	Kollokationsmaß	Frequenz	Rang	Wort	Ähnlichkeitsmaß	Frequenz
1	friedensschaffende	21.302814	291	1	friedenssichernde	0.189810	337
2	friedenschaffende	16.195538	118	2	friedensschaffende	0.181358	291
3	Einsätze	11.713463	2036	3	friedenserhaltenden	0.154448	707
4	friedensstiftende	11.592508	188	4	friedenschaffende	0.150318	118
5	Untergeneralsekretär	11.333069	75	5	friedensbewahrende	0.146688	106
6	humanitäre	11.296459	2339	6	friedenserzwingende	0.142003	135
7	Missionen	11.077414	1263	7	friedenssichernden	0.127425	251
8	Maßnahmen	10.677504	3336	8	friedensschaffenden	0.125832	290
9	Einsaetze	10.617214	1278	9	friedenschaffenden	0.088791	116
10	Massnahmen	10.040110	2026	10	friedenserzwingenden	0.085437	127
11	Operationen	9.662893	2435	11	humanitaere	0.083739	1681
12	Kampfeinsaetze	8.427442	438	12	friedensstiftenden	0.082751	144
13	friedenstiftende	8.321952	50	13	humanitäre	0.077842	2339
14	Kampfeinsätze	8.199440	646	14	friedensstiftende	0.071462	188
15	Rettungseinsätze	7.929589	63	15	Friedenserhaltende	0.071104	48
16	Militäreinsätze	7.600626	567	16	friedenserhaltender	0.067560	138
17	Blauhelm-Einsaetze	7.431248	185	17	humanitaeren	0.060287	1434
18	Aufgaben	7.405412	3467	18	multinationale	0.057137	814
19	UNO-Einsaetze	7.355560	150	19	schaffende	0.056523	149
20	Mission	7.354951	3452	20	militaerische	0.049970	3312
21	Aktionen	7.277963	3428	21	humanitären	0.048641	2128
22	derzufolge	7.246859	297	22	antiterroristische	0.047862	148
23	Kanzlermehrheit	6.993749	190	23	weitergehende	0.047668	400
24	ausfuehren	6.712513	240	24	Blauhelm-Einsätze	0.046909	242
25	Blauhelm-Einsätze	6.499894	242	25	vorbeugende	0.045998	384
26	Bundeswehr	6.237191	5594	26	kuenftige	0.045291	1261
27	UNO	6.167113	5556	27	polizeiliche	0.045248	460

Abbildung 5: Ausgabe von DISCO zum Suchwort "friedenserhaltende"

Um die Präzision hoch zu halten, wird nach der exakten Phrase gesucht, indem (wie bei Google) die Suchwörter in doppelte Anführungszeichen gesetzt werden. Textstellen mit *friedenserhaltenden Mission* oder *friedenserhaltende militärische Mission* werden also nicht

gefunden. Die erste Suchanfrage, „*friedenserhaltende Mission*“ ergibt 14 Treffer auf dem deutschsprachigen Datenbestand. Um die Vollständigkeit der Treffermenge zu erhöhen, wird in DISCO jetzt nach ähnlich gebrauchten Wörtern zum Term *friedenserhaltende* gesucht. Das Ergebnis der DISCO-Anfrage ist in Abbildung 5 dargestellt. Die distributionell ähnlichen Wörter wie *friedenssichernde*, *friedensschaffende*, *humanitäre* usw. können in die Anfrage aufgenommen werden, entsprechend wird mit dem zweiten Suchterm verfahren. Zum Suchwort *Intervention* liefert DISCO die folgenden Wörter unter den ersten 30 distributionell ähnlichsten Wörtern:

> *Eingreifen Militäraktion Invasion Militärintervention Einmischung Einsatz Militärschlag Angriff Mission Vorgehen Operation Gewaltanwendung Einmarsch Engagement ...*

Damit lässt sich durch eine Disjunktion (per booleschem Operator OR) aller möglichen Kombinationen der Suchterme eine komplexe Suchanfrage mit stark erhöhtem Recall der Treffermenge formulieren:

> *„friedenserhaltende Intervention“ OR „friedenssichernde Intervention“ OR „friedensschaffende Intervention“ OR ... OR „humanitäre Mission“*

3.2.2 Schreibvarianten von Namen

Ein häufiges Problem, das die Vollständigkeit von Treffermengen negativ beeinflusst, ist die große Varianz bei der Schreibung insbesondere fremdsprachiger Namen. So finden wir in unserem englischsprachigen Teilkorpus zur Anfrage *Arafat* 12.610 Trefferdokumente. Um die Präzision sicherzustellen, soll aber nach dem vollständigen Namen gesucht werden. Die DISCO-Anfrage mit dem Suchwort *Arafat* liefert den richtigen Vornamen, inklusive mehrerer im Korpus verwendeter Schreibvarianten:

> Kookkurrenzen: *Fatah exiling Yasser PLO YASSIR Palestinian Barak Qureia Kanafani Yasir Yassar Peres ...*
> distributionell ähnliche: *Yasser Yasir Abbas Yassir PLO Dahlan Fatah Netanyahu Peres Rabin Erekat Qureia Mazen Barak ...*

Während die Suche nach „*Yasser Arafat*“ lediglich zu 6.607 Trefferdokumenten führt, ergibt die komplexe Suchanfrage

> *„Yasser Arafat“ OR „Yasir Arafat“ OR „Yassir Arafat“ OR „Yassar Arafat“ OR (Arafat AND (PLO OR Palestine OR Palestinian))*

12.444 Treffer und ist somit hinsichtlich Präzision und Vollständigkeit optimal.
Zudem kann man mittels DISCO schnell erkennen, dass es für die Schreibung des Nachnamens *Arafat* keine gängigen Varianten gibt – unter den 50 distributionell ähnlichsten Wörtern zu *Arafat* findet sich nämlich kein ähnlich geschriebener Kandidat. Eine Suchanfrage nach der am plausibelsten erscheinenden Variante *Arrafat* ergibt in der Tat nur einen einzigen Treffer (in ca. 300.000 Dokumenten).
Eine weitergehende Automatisierung des Aufdeckens von Namensvarianten wäre durch eine Filterung der Ähnlichkeitslisten mit einem String-Ähnlichkeitsmaß wie z.B. dem Levenshtein-Editierabstand in einfacher Weise realisierbar. Solche Äquivalenzklassen würden dann den ersten Schritt zu einem automatischen Ontologieaufbau [Cimiano et al. 2006] oder zur Erweiterung einer vorhandenen Ontologie bilden. Die durch die Äquivalenzklassen identifizierten Konzepte würden mit Hilfe der in Abschnitt 3.1.3 beschriebenen Verfahren in eine Taxonomie oder Ontologie eingeordnet werden.

4. Zwischenbilanz und Ausblick

Für die Erfassung von semantischen Feldern zu bestimmten inhaltlichen Konzepten (institutionelle Akteure, Europa), die aus politikwissenschaftlicher Sicht interessierten, eignete sich die gegenwärtige Version von DISCO nur bedingt, weil lediglich ähnlich verwendete und kookkurrierende Wörter (zu "Deutschland" bspw. "Frankreich", zu "Europa" "Asien") erfasst werden, nicht aber semantische Untergruppen wie "Bundesregierung", "Auswärtiges Amt" etc. Eine weitere Einschränkung von DISCO ist, dass es lediglich tokenbasiert arbeitet und daher keine Mehrwortausdrücke gesucht werden können. Auch ist mit DISCO nicht unmittelbar festzustellen, ob durch ein Wort im Korpus stets der gerade gesuchte Akteur benannt wird. Beispielsweise bezieht sich ein Vorkommen von "Bundeskanzler" nicht immer auf Deutschland als außenpolitischen Akteur. Um sicherzustellen, dass sich bestimmte Vorkommen einzelner Wörter regelmäßig auf das gesuchte semantische Feld bzw. den dadurch bezeichneten Akteur beziehen, ist eine Untersuchung der einzelnen Belegstellen notwendig, was derzeit nur durch eine manuelle Konkordanzanalyse geleistet werden kann.

Im Text Mining wie bei der Inhaltsanalyse besteht eine grundsätzliche Kluft zwischen der konzeptuellen Ebene einerseits und der textlich-lexikalischen Ebene andererseits. Existiert eine manuell erstellte Ontologie aus interessierenden Konzepten, d.h. abstrakten semantischen Einheiten, müssen diese in konkret formulierten sprachlichen Ausdrücken wiedergefunden werden. Texte müssen also erst einmal mit Konzepten annotiert werden. Dabei treten die schon aus dem Information Retrieval bekannten Probleme der Lesartenambiguität und Paraphrase auf, d.h. ein Korpus mit einer vorhandenen Ontologie zu annotieren ist keineswegs trivial. Wir erwarten, mit einem korpusgetriebenen Ontologieaufbau mittels DISCO dieses Problem entschärfen zu können.

Ein alternativer Ansatz beseht darin, textbasiert vorzugehen und aus relevanten Phrasen im Text einen "Themenbaum" aufzubauen. Hier wird versucht, von der konkreten Formulierung im Text zu einer semantischen, möglichst abstrakten Darstellung zu gelangen. Die Schwierigkeit ist dabei, einen hinreichenden Abstraktionsgrad zu erreichen, um gleichbedeutende, aber unterschiedlich ausgedrückte Inhalte überhaupt aufeinander beziehen zu können. Außerdem ist zu erwarten, dass es sich bei den automatisch extrahierten Themen nicht unbedingt um Konzepte der Art handelt, die einer intellektuell erstellten Ontologie entsprechen. Wir hoffen, durch die kombinierte Verwendung der verschiedenen Ansätze die Kluft zwischen konzeptueller und textueller Ebene überbrücken zu können.

Literatur

P. Cimiano, J. Völker und R. Studer (2006): Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. *Information, Wissenschaft und Praxis*, 57(6-7), S. 315-320.

M. Hearst (1992): Automatic acquisition of hyponyms from large text corpora. *COLING 1992*, Nantes, Frankreich, S. 539—545.

P. Kolb (2008): DISCO: A Multilingual Database of Distributionally Similar Words. *Tagungsband der 9. Konferenz zur Verarbeitung natürlicher Sprache – KONVENS 2008*, Berlin.

R. Rapp (1999): Automatic Identification of Word Translations from Unrelated English and German Corpora. *Proceedings of ACL*, College Park, Maryland, S. 519–526

J.J. Rocchio (1971): Relevance Feedback in Information Retrieval. In G. Salton (Hrsg.): *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall.

G. Salton (1971): *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall.

H. Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop,* S. 47—50.

F. Sebastiani (2002): Machine learning in automated text categorization. *ACM Computing Surveys*, 34, S. 1—47.