Uwe Quasthoff, Matthias Richter, Christian Wolff

Medienanalyse und Visualisierung: Auswertung von Online-Pressetexten durch Text Mining

1 Einführung

Obwohl sich die Medienwissenschaft als inter- oder transdisziplinäres Arbeitsfeld versteht, haben methodische Ansätze aus der angewandten Informatik bisher kaum Eingang in das Methodeninventar der Medienwissenschaft gefunden (vgl. Rusch 2002b: 70f). Die so genannten „Neuen Medien" finden sich zwar als **Subjekt** medienwissenschaftlicher Betrachtungen wieder (vgl. Ludes 1998: 51ff, 129), es besteht aber ein Defizit hinsichtlich der Modernisierung geeigneter Methoden für die Medienanalyse, die bisher von qualitativen und an einzelnen Publikationen bzw. Medienereignissen orientierten Verfahren geprägt ist (vgl. Posner 2001, Andringa 2002, Schreier 2002).

Dieser Beitrag versucht aufzuzeigen, wie Text Mining-Verfahren für die inhaltliche Auswertung von Pressetexten genutzt werden können und so als „angewandte Medieninformatik" einen interdisziplinären Beitrag zur Medienanalyse leisten können. Es wird ein im World Wide Web verfügbarer Informationsdienst vorgestellt, der tagesaktuell überregionale Online-Medien auswertet und begriffsbasiert die jeweils als relevant erkannten Konzepte als „Wörter des Tages" präsentiert. Dabei kommen sowohl Darstellungen relevanter Begriffe, die einem einfachen Kategoriensystem zugeordnet sind zum Zuge, als auch Visualisierungen von aktuellen Begriffsassoziationen und Visualisierungen des Aktualitätsverlaufs einzelner Konzepte.[1]

2 Analyse von Online-Presseartikeln durch Text Mining

Der Ansatz, durch Auswertung von Online-Medien jeweils täglich „aktuelle" Begriffe bestimmen zu können, geht von folgenden Annahmen aus:

---

[1] Der hier beschriebene Ansatz wurde im Rahmen des Projekts „Deutscher Wortschatz" am Institut für Informatik der Universität Leipzig entwickelt und baut auf den dort gewonnenen Daten und Verfahren auf, vgl. Quasthoff & Wolff 2000, 2002. Die Wörter des Tages sind im World Wide Web unter http://www.wortschatz.uni-leipzig.de/wort-des-tages/ verfügbar.

- Die „Aktualität“ von Personen und Ereignissen lässt sich über einfache Parameter wie die Begriffsverwendungshäufigkeit erfassen, Medien lassen sich in diesem Sinn als „Spiegel der Wirklichkeit“ verstehen.[2]
- Hinreichend viele und repräsentative Onlinemedien lassen sich automatisch erfassen.
- Der täglich erfasste Textbestand hat einen Mindestumfang, der durch Text Mining-Verfahren analysiert werden kann.

Nachfolgend wird der Aufbau eines automatisierten Verfahrens zur Analyse von Online-Medien beschrieben. Dabei kann man grob folgende Schritte unterscheiden:

- Quellenauswahl und –erfassung
- Quellenanalyse durch Text Mining
- Begriffsselektion (Kandidaten für „Wörter des Tages“)
- Kategorisierung und Überarbeitung von Kandidatenlisten
- Aufbereitung und Präsentation der Ergebnisse in unterschiedlichen Visualisierungsformaten

2.1 Quellenauswahl

Die Quellenauswahl erfolgt unter Heranziehung technischer, quantitativer und qualitativer Merkmale. In technischer Hinsicht kommen nur Online-Angebote in Frage, deren Inhalte sich tagesaktuell durch an sie adaptierte *web spider* erfassen lassen, d. h. deren Archive ein einheitliches System der URL-Kodierung aufweisen. In quantitativer Hinsicht werden Quellen mit vergleichsweise großem (Text-)Umfang und – was damit in mittelbarem Zusammenhang stehen dürfte – auch entsprechender medialer Reichweite ausgewählt. Daneben kommen – analog zu den Auswahlkriterien für das Referenzkorpus des „Deutschen Wortschatz“ – qualitative Merkmale zum Zug: In die Quellenauswahl werden insbesondere solche Online-Dienste aufgenommen, die überregionalen Charakter haben und ein breites Themenspektrum aufweisen, insbesondere die Angebote der großen Tageszeitungen und Wochenzeitungen. Tabelle 1 zeigt am Beispiel eines aktuellen Tageskorpus die relative Ergiebigkeit der zehn hinsichtlich des derzeit verfügbaren Textumfangs wichtigsten Quellen.

[2] In einem weitergehenden Sinn lassen sich die hier beschriebenen Verfahren auch nutzen, um zu untersuchen, inwieweit Sachverhalte in den Medien unterschiedlich dargestellt werden und ihnen so auch eine die Wirklichkeit **prägende** Rolle zukommt (vgl. zur Untersuchung von Meinungsbildungsprozessen Gerhards, Neidhard & Rucht 1998).

Quelle	Anzahl der Sätze (20. Juni 2002)
http://www.sueddeutsche.de/	6586
http://www.berlinonline.de/	3102
http://www.welt.de/	2626
http://spiegel.de/	1540
http://www.ln-online.de/	1385
http://www.netzeitung.de/servlets/	1225
http://www.heute.t-online.de/	834
http://www.svz.de/	188
http://www.heise.de/	161

Tabelle 1: Quellenauswahl und –umfang.

Die Liste der erfassten Quellen wird laufend erweitert; hinsichtlich eines repräsentativen Medienquerschnitts weist sie zwar noch Mängel auf, da die Medienauswahl sich aus den o. a. Gründen nicht **ausschließlich** auf medienbezogene Parameter stützt; sie kann aber als erste Annäherung zur Veranschaulichung des gewählten Verfahrens genügen.

Für nicht nur qualitativ, sondern auch hinsichtlich der **quantitativen** Zusammensetzung zu bewertende Analyseergebnisse wäre eine normierte Quellenauswahl erstrebenswert, die sich beispielsweise nach den Online-Nutzungsdaten der „Interessengemeinschaft zur Feststellung der Verbreitung von Werbeträgern e. V." (IVW) richten könnte. Die auf der Website der IVW veröffentlichten Online-Daten (vgl. http://www.ivwonline.de/ausweisung/suchen.php) lassen die Definition eines repräsentativen „Medien-Warenkorbs" zu, der ungeachtet des dort eingeführten Kategoriensystems unterschiedlicher Online-Angebote, von der Orientierung an Printprodukten abstrahieren könnte.

2.2 Quellenanalyse und Text Mining

Die jeweils über Nacht gesammelten Quellen werden einer Text Mining-Analyse durch die im Projekt „Deutscher Wortschatz" entwickelten Werkzeuge unterzogen (vgl. Heyer, Quasthoff & Wolff 2000; Quasthoff & Wolff 2002). Diese besteht aus folgenden Schritten:

- Bereitstellung eines im Vergleich mit einer einzelnen Tagessammlung etwa um den Faktor 1000 umfangreicheren Referenzkorpus, das auch linguistische und semantische Angaben (z. B. umfangreiche Mehrwortbegriffslisten) enthält
- Textsegmentierung, insbesondere Satz- und Wortsegmentierung

- Indexierung der Beispieltexte und quantitative Erfassung der Einzelbegriffe
- Berechnung relevanter Satz- und Nachbarschaftskollokationen für die im Tageskorpus enthaltenen Begriffe[3]
- Speicherung der Analyseergebnisse in einer relationalen Datenbank

Für jeden Tag wird eine eigene Datenbank mit identischer Struktur angelegt, die für weiterführende Analysen zur Verfügung steht. Dies lässt insbesondere semiometrische Analysen zum Begriffswandel über einen längeren Zeitraum zu; zudem lassen sich daraus Datenbanken mit geänderter Zeitgranularität erstellen (z. B. „Wörter des Monats", „Wörter des Jahres" etc.).

2.2.1 Basisparameter eines Tageskorpus

Tabelle 2 zeigt exemplarisch für ein Tageskorpus (20. Juni 2002) den Umfang der Rohanalyseergebnisse:

Sätze	17645
laufende Wortformen	255680
Verschiedene Wortformen	41031
Satzkollokationen	90528
Nachbarschaftskollokationen	6581
Nachbarschaftskollokationen von Wörtern beginnend mit Großbuchstaben	546
Relative Korpusgröße im Vergleich mit dem Referenzkorpus	1 : 936,16

Tabelle 2: Basisparameter eines Tageskorpus.

Es lässt sich festhalten, dass trotz kleinerer Abweichungen, die durch den unterschiedlichen Publikationsumfang verschiedener Wochentage bedingt sind, über einen längeren Zeitraum das angestrebte Größenverhältnis zwischen Tages- und Referenzkorpus erreichen lässt.

[3] Grundlage der Auswertung sind dabei sowohl einzelne Wörter (Vollformen) als kleinste Analyseeinheiten, als auch Mehrwortbegriffe, soweit sie als solche durch Nachschlagen im Referenzkorpus erkennbar sind. Für den Mehrwortbegriff **Bundeskanzler Gerhard Schröder** wird eine Analyse sowohl für die Einzelwörter **Bundeskanzler**, **Gerhard**, **Schröder**, als auch für den Mehrwortbegriff selbst durchgeführt.

2.2.2 Selektion relevanter Begriffe

Nach Durchführung der Text Mining-Analyse für das jeweilige Tageskorpus muss eine „handhabbare" Menge von Wörtern des Tages selektiert werden, die sich für eine Präsentation im Rahmen eines Web Service eignet. Da sich die Eignung von Wörtern als Aktualitätsindikator der Medienanalyse nicht nach einem einzelnen Parameter richten kann, stehen für die Auswahl der „Wörter des Tages" die folgenden drei statistischen Basisparameter zur Verfügung:

Parameter		**Motivation**
n_1	Frequenz im aktuellen Tageskorpus	Ein im Tageskorpus hochfrequentes Wort deutet auf ein aktuelles Thema hin[4]
n_{ges}	Frequenz im Referenzkorpus (Deutscher Wortschatz)	Eine Mindestfrequenz im Referenzkorpus deutet darauf hin, dass es sich um ein bekanntes Konzept oder einen bekannten Eigennamen handelt (und nicht etwa um eine fehlerhafte Schreibung)
$1000\ n_1 / n_{ges}$	relative Übergewichtung eines aktuellen Begriffs	Ein Indikator, der deutlich macht, dass ein Begriff im Tageskorpus häufiger auftritt, als nach den Werten im Referenzkorpus zu erwarten gewesen wäre

Tabelle 3: Auswahlparameter für relevante Begriffe.

Durch Tests haben sich für diese Basisparameter folgende Schranken als sinnvoll erwiesen:[5]

- $1000\ n_1 / n_{ges} > 16$. Damit wird gewährleistet, dass ein Wort des Tages im Tageskorpus um wenigstens vier Häufigkeitsklassen (2^4) häufiger auftritt, als nach den Werten des Referenzkorpus zu erwarten wäre.
- Festlegung einer $n_{ges} > 20$. Damit ist sichergestellt, dass es sich um ein bekanntes Konzept handelt.
- Festsetzen einer Mindesthäufigkeit im aktuellen Korpus auf $n_1 > 8$. Bei Unterschreiten dieses Schwellwerts steigt die Anzahl der Kandidaten sehr stark an.

Die Parameter kommen jeweils *alternativ* zum Zuge, ein potentielles Wort des Tages muss also nur **einen** der Parameter erfüllen. Damit ist z. B. sichergestellt, dass ein im

[4] Stoppwörter werden vor der Auswertung eliminiert.

[5] Es ist offensichtlich, dass sich nur schwerlich eine exakte Begründung für die Höhe der Schwellwerte ableiten lässt; sie beruhen daher auf einer Reihe von Tests, bei denen jeweils eine analytische Bewertung der Ergebnismengen von Begriffen durchgeführt wurde.

Referenzkorpus nicht vorhandenes Konzept nicht von vornherein ausgeschlossen wäre. Zusätzlich zu diesen numerischen Faktoren werden nur Nomina (hilfsweise: Wörter, die mit einem Großbuchstaben beginnen) für die Auswahl als Wort des Tages herangezogen. Aufgrund der relativ wenigen flektierten Formen wird auf eine zusätzliche Grundformreduktion bisher verzichtet. Die nachfolgende Liste (Abbildung 1) zeigt die nach diesen Selektionskriterien erzeugte Liste der Kandidaten für die Wörter des Tages vom 24. Juni 2002:

Abschlussbericht, Abu, Annecy, Babcock, Barrichello, Bild am Sonntag, Bildungspolitik, Bin Laden, Bondy, Bulmahn, Bundespolitik, Butler, Coast, Coulthard, Daglfing, Djerba, Elfmeterschießen, Fifa, Foul, France Télécom, Fritz Walter, Fußball-WM, Gläubigerbanken, Golden Goal, Großaktionäre, Guus Hiddink, Gymnasien, Hartz, Hewitt, Hiddink, Ilhan, Insolvenz, Kahn, Koreaner, Kroetz, Lifestyle, Lukaschenko, Manfred Stolpe, Matthias Platzeck, Medienberichten, Mobilcom, MobilCom, Montoya, Müntefering, Naturwissenschaften, Nürburgring, NZZ, Oliver Kahn, Pisa, Plank, Platzeck, Platzecks, Ralf Schumacher, Rößler, Rubens, Rubens Barrichello, Rüge, Schill, Schmid, Schulsystem, Schuss, Senegal, Senegals, Sevilla, Skibbe, Stage, Stolpe, Stolpes, Strüver, Südkorea, Südkoreaner, Südkoreas, Talkline, Tokyo, Trienekens, Völler, Walküre, Wienand, Wittenberge, WM-Halbfinale, Zeppelin, Zuwanderungsgesetz

Abb. 1: Sortierte Liste der Kandidaten für den 24. Juni 2002.

2.3 Kategorisierung der Wörter des Tages

Wendet man ausschließlich statistische Analyseverfahren bei Text Mining-Analysen an, so erhält man Datenmengen, die in der Regel durch den Menschen gut interpretierbar sind, die sich aber offensichtlich durch Anwendung wissensbasierter Filter weiter aufbereiten lassen (vgl. dazu ausführlich Heyer et al. 2001, Heyer, Quasthoff, Wolff 2002). Dies zeigt auch die voranstehende Liste von Kandidaten für den 24. Juni 2002. Um die Präsentation besser strukturieren zu können, werden die Wörter des Tages nach einem einfachen Kategorienschema klassifiziert, dessen Granularität in etwa der Ressortaufteilung einer Tageszeitung entspricht. Derzeit enthält das Kategorienschema die in Abb. 2 in Spalte 2 dargestellten Einträge und folgt gleichzeitig einer Einteilung in Eigennamen (d. h. Personennamen) und Konzept. Zusätzlich nimmt eine unspezifische Klasse „Schlagwort“ alle als relevant erachteten Wörter des Tages auf, die sich nicht einer der bereits vorhandenen Klassen zuordnen lassen.

Die Klassifikation baut auf bereits vorhandenem Wissen aus dem Referenzkorpus auf. Die dort verfügbaren umfangreichen Listen von Eigennamen und Mehrwortbegriffen gehen in die Text Mining-Analyse mit ein.[6] Zusätzlich existiert eine webbasierte Editierungsschnittstelle, mit deren Hilfe sich u. a.

- neu aufzunehmende Mehrwortbegriffe überprüfen lassen und
- die Kandidaten für die Wörter des Tages dem Kategoriensystem zugeordnet werden können

Abb. 2 zeigt die Schnittstelle für die Kategorisierung von Kandidaten, in der zu jedem Begriff jeweils zusätzlich eine Belegstelle angezeigt wird, um bei seltenen Begriffen die Einordnung zu erleichtern.

Begriff	Kategorie	gew. Signifikanz	#heute	#wortschatz	gew. Anzahl	Beispiel
Quartal	–bitte auswählen–	3.42	101	13030	6.6460	Der Schweizer Finanzkonzern Credit Suisse Group hat im dritten Quartal einen Nettoverlust von 2,1 Milliarden Franken (etwas mehr als 1,4 Milliarden Euro) erwirtschaftet.
Analysten	NICHT ANZEIGEN	3.13	57	6650	7.3491	Der Verlust war zugleich höher als von Analysten erwartet.
Rente	Schlagwort	2.27	41	5211	6.7460	Menschen, die jetzt in Rente sind, dürften nur begrenzt belastet werden, sagte Rürup.
Resolution	Schlagwort	2.73	39	2609	12.8166	Die USA und Großbritannien haben mehrmals erklärt, dass sie zu einem Angriff auf Irak im Falle einer Nichterfüllung der Resolution bereit sind.
Schnabel	Person	2.95	36	678	45.5257	Im Comroad-Prozess könnte auch die Rolle der Wirtschaftsprüfergesellschaft KPMG beleuchtet werden, nachdem Schnabel kein Geständnis abgelegt hat.
Ffm	NICHT ANZEIGEN	12.82	35	330	90.9364	Grillmaier, Homburg, 21,16; Prof. Jürgen Grimm, Münster, 10,75; Werner Grundmann, Hofheim, 12,41; Marlies Gunkel, Ffm 50,-; Manfred Günther, Neu-Isenburg, 10,86 Euro.
Gesamtjahr	NICHT ANZEIGEN	2.99	32	1930	14.2160	Trotz dieser Entwicklungen geht EADS davon aus, dass die Ebit-Prognose von 1,4 Milliarden Euro im Gesamtjahr erreicht wird.
Vorstandschef	NICHT ANZEIGEN	3.45	32	3632	7.5542	Als Übergangslösung wurde Helmut Sihler zum Vorstandschef benannt, doch die externe Suche nach einem Nachfolger Sommers blieb offenbar ohne Erfolg.
Betriebsergebnis	Schlagwort	3.66	27	2384	9.7105	Der Baukonzern Hochtief hat in den ersten neun Monaten ein Betriebsergebnis von 164 Millionen Euro erzielt.
Reporter	NICHT ANZEIGEN	1.52	23	3069	6.4256	Als jedoch Linenger von einem Reporter gefragt wurde, ob die UFO-Kontroverse seines Erachtens einen wahren Kern beinhalte, nahm der Small talk eine unerwartete Wendung.
Bundesanstalt	Organisation	6.07	22	1837	10.2683	Die arbeitsmarktpolitischen Mittel der Bundesanstalt müssten zielgerichteter eingesetzt werden.
Konzerte	NICHT ANZEIGEN	2.38	21	2860	6.2956	In Museen und Konzerte kämen pro Jahr mehr als doppelt so viele Besucher wie zu allen Spielen der Bundesliga.
Geständnis	–bitte auswählen–	3.49	20	2059	8.3283	Im Comroad-Prozess könnte auch die Rolle der Wirtschaftsprüfergesellschaft KPMG beleuchtet werden, nachdem Schnabel kein Geständnis abgelegt hat.
Vorjahreszeitraum	–bitte auswählen–	2.82	19	2372	6.8679	Aufgrund eines Rückgangs der Airbus-Auslieferungen und des niedrigeren Dollarkurses lag der Umsatz zwischen Januar um September den Angaben zufolge nur bei etwa 20 Milliarden Euro nach 20,7 Milliarden im entsprechenden Vorjahreszeitraum.
Jackson	Künstler	3.47	18	1275	12.1045	Als der wegen Steuerhinterziehung verurteilte Konzertmanager im März 1998 seine Haftstrafe antrat, reiste Jackson an, um sich persönlich von ihm zu verabschieden.
Babys	NICHT ANZEIGEN	2.28	17	1207	12.0761	"Babys werden fast immer richtig gesichert", sagt Gabriele Scheulen, Projektleiterin Kinder und Jugend bei der Deutschen Verkehrswacht (DVW) in Meckenheim bei Bonn.
Hewitt	Politiker	3.93	17	76	191.7868	Hewitt unterlag in der roten Gruppe dem Spanier Carlos Moya mit 4:6 und 5:7. Dabei zog Moya die Partie aufgrund einer Unachtsamkeit noch in die Länge.
Ausblick	–bitte auswählen–	2.71	17	1298	11.2294	Nach deutlichen Rückgängen bei Gewinn und Ergebnis gab der Konzern wenige Wochen vor Jahresende einen pessimistischen Ausblick für das Gesamtjahr.
Route	–bitte auswählen–	2.13	17	1867	7.8071	Das 243 Meter lange Schiff hatte für eine griechische Reederei die Route von Riga (Lettland) zur britischen Kolonie Gibraltar befahren.

Abb. 2: Kategorisierung von Kandidaten für die **Wörter des Tages** (15.11.2002).

[6] Dabei wird sowohl bestehendes Wissen aus dem Referenzkorpus eingearbeitet als auch mit Hilfe eines **named entity recognizers** laufend der Datenbestand erweitert (Erfassungsrate: Ca. 1.500 Eigennamen pro Tag, vgl. Quasthoff, Biemann & Wolff 2002).

2.4 Erkennen von Mehrwortgruppen

Um Ambiguitäten besser auflösen zu können und die Beschreibungsgenauigkeit der Wörter des Tages verbessern zu können, ist es wünschenswert, auch Mehrwortgruppen zu erkennen und als Wörter des Tages einbinden zu können. Hierzu werden die bei der Text Mining-Analyse jedes Tageskorpus generierten signifikanten Nachbarschaftskollokationen auf Nomina eingeschränkt und als Vorschlagsliste präsentiert. Diese Vorschlagsliste muss allerdings intellektuell nachbearbeitet werden, da nicht alle generierten Mehrwortgruppen als eigenes Konzept sinnvoll erscheinen (vgl. Abb. 3: Gruppen wie **El Baradei**, **Nenad I** oder **Masters Cup** erscheinen akzeptabel, während **Ehefrau Ingrid** kaum ein adäquates Konzept darstellen dürfte). Durch Iteration können auch umfangreichere Mehrwortgruppen entstehen, da bei erneuter Analyse Zweiwortgruppen mit einem weiteren Term eine Dreiwortgruppe bilden können und dieser Prozess erneut (spätestens am nächsten Tag) iteriert wird. Neben der Selektion von Mehrwortgruppen erfolgt auch eine vollautomatische Extraktion von Eigennamen (Vor- und Nachnamen) mit Hilfe eines **named entity extractors** (s. o. Fn. 6).

Mehrwortbegriffe

"+": künftig als Mehrwortbegriff verwenden
"-": nicht wieder vorschlagen

Begriff	signifikanz	+	-
Jet Li	55	☐	☑
Asset Allocation	44	☐	☐
Masters Cup	41	☑	☐
Göttliche Kerle	38	☐	☑
Oude Kamphuis	38	☑	☐
Adnan Rizvic	37	☐	☐
Tschechen Jiri Novak	37	☐	☐
Ehefrau Ingrid	36	☐	☑
Le Defi	36	☐	☐
Spanier Carlos Moya	36	☐	☐
Nenad I	34	☐	☐
Casa Grande	31	☐	☐
Linkin Park	31	☐	☐
El Baradei	29	☐	☐
Internationalen Atomenergiebehörde	29	☐	☐
Stuttgarter Kevin Kuranyi	29	☐	☐

Abb. 3: Auswahl geeigneter Mehrwortgruppen (15.11.2002).

Mit der Kategorisierung und der Identifikation von Mehrwortgruppen ist die Analysephase für die Wörter des Tages beendet. Die Analyseergebnisse stehen in einer relationalen Datenbank für die Präsentation und Visualisierung bereit.

3 Ergebnisaufbereitung und Visualisierung

Die Ausgabe der Wörter des Tages erfolgt als **web service** und lässt sich in folgende Bereiche gliedern:

- Textuelle Ausgabe der Wörter des Tages (nach Kategorien geordnet)
- Ausgabe der Belegstellensammlung zu einem Wort des Tages einschließlich Querverweisen zu den Quelldokumenten
- Ausgabe eines Kollokationsgrafen, der wesentliche inhaltliche Beziehungen eines Wortes des Tages zu anderen im Tageskorpus enthaltenen Konzepten darstellt
- Grafische Ausgabe des „Aktualitätsverlaufs“ von Worten des Tages im Verhältnis zu verwandten Konzepten

3.1 Darstellung der „Wörter des Tages“

Die Startseite des **web service** gibt eine Übersicht zu den aktuellen Wörtern des Tages, die nach den ihnen zugeordneten Kategorien sortiert ausgegeben werden. Über diese Übersichtsseite ist die Navigation zu vorangegangenen (oder nachfolgenden) Tagescorpora ebenso möglich wie der Aufruf der Detailergebnisse (Belegstellen, Visualisierungen) über die mit den einzelnen Wörtern des Tages verbundenen Hypertextlinks (Abb. 4).

Wortschatz : **Wörter des Tages** : 15.11.2002

Sportler, Trainer, Funktionäre	Agassi · Andre Agassi · Ballack · Bobic · Carlos Moya · Effenberg · Federer · Hitzfeld · Jiri Novak · Jäggi · Völler
Sport	Bayer Leverkusen · Champions League · DTB · FC Kaiserslautern · Länderspiel · Schalke · VfL Bochum · Zwischenrunde
Politiker	Berlusconi · Bundesfinanzminister Hans Eichel · Bundeskanzler Gerhard Schröder · Bush · Clement · Eichel · Faltlhauser · Jiang · Jiang Zemin · Platzeck · Saddam · Saddam Hussein · Steinbrück
Organisation	BASF · Cinerenta · Deutsche Telekom · E.ON · Epcos · France Telecom · France Télécom · Hochtief · Linde · MobilCom · Mobilcom · Neuen Markt · Rot-Grün · T-Online · Versicherer
Ereignis	Insolvenz · Regierungserklärung
Schlagwort	Enduring Freedom · Fonds · Massenvernichtungswaffen · Missbrauch · Rekordverlust · Sicherungsverwahrung · Steuerschätzung · T-Aktie · Tanker · UN-Resolution · Zwischenlager · Öffentlichen Dienst
Ort	Bagdad · Dannenberg · Gorleben · Irak · Manchester · NRW · Nablus · Russland · Schanghai · Wendland
Personen aus Kunst, Kultur und Wissenschaft	Rolling Stones
sonstige Personen	Atomkraftgegner · Augstein · Australier · Bin Laden · Bodo Schnabel · Gates · Harry Potter · Hartz · Helmut Sihler · Inspektoren · Kai-Uwe Ricke · Leiharbeiter · Panke · Ron Sommer · Rürup · Schmid · Sexualstraftäter · Sihler · Treuhänder

««**14.11.2002**«« Wörter des Tages

Abb. 4: Webbasierte Ausgabe der Wörter des Tages vom 15. 11. 2002.

3.2 Vernetzung mit Belegstellen

Um eine weitergehende Analyse zu ermöglichen, werden die Belegstellen für die verschiedenen „Wörter des Tages“ in einer Datenbank gesammelt und können online abgefragt werden. Die Auswahl der Belegstellen richtet sich dabei nach der Reihenfolge der Quellenanalyse und ist nicht grundsätzlich beschränkt.

Die Belegstellen werden – auch aus urheberrechtlichen Gründen – satzbasiert mit Hervorhebung des jeweiligen Wortes des Tages als Liste ausgegeben. Über das Wort des Tages lässt sich das Quelldokument unmittelbar abrufen. Damit ist gewährleistet, dass ausgehend von der quantitativen Analyse der täglichen Medienproduktion auch eine qualitative Feinanalyse im Sinne einer traditionellen medienanalytischen Vorgehensweise möglich ist. Die Abbildungen 5 und 6 zeigen die Sammlung von Belegstellen zu **Rot-Grün** als Wort des Tages vom 15. November 2002 sowie eine ausgewählte Online-Quelle (Website der **Süddeutschen Zeitung**).

Wortschatz: Wörter des Tages: Belegstellen für »Rot-Grün« am 15.11.2002

1. Einig ist sich **Rot-Grün** mit der Opposition zudem darin, dass künftig DNA-Analysen bei Exhibitionisten vorgenommen werden.
2. Im Bundestag sagte Staatssekretär Karl Diller (SPD), **Rot-Grün** werde im nächsten Jahr die niedrigste Neuverschuldung seit der Wiedervereinigung vorweisen können.
3. In Regierungskreisen hieß es zu den Vorwürfen, **Rot-Grün** werde die jetzt angestrebte Neuverschuldung von 18 Milliarden Euro nicht überschreiten.
4. SZ: **Rot-Grün** und die CDU gründen Kommissionen und versuchen somit, bei dem heiklen Thema Zeit zu gewinnen.
5. Was **Rot-Grün** heute im Parlament beschließt, ist nur Stückwerk ohne Konzept.
6. **Rot-Grün** belastet jetzt vorrangig nur Arbeitnehmer und Arbeitgeber.
7. SZ: **Rot-Grün** reduziert in den nächsten Jahren im Zuge der Riester-Rente durchaus das Rentenniveau.
8. **Rot-Grün** wies die Vorwürfe zurück.
9. Auch die von **Rot-Grün** behauptete Lenkungswirkung der Ökosteuer gebe es nicht.
10. Erst vor zwei Jahren hat **Rot-Grün** den Standort Deutschland durch steuerliche Vergünstigungen für Holdings wieder attraktiv gemacht.
11. Das Problem von **Rot-Grün** ist nicht der Einfluss der Gewerkschaften, sondern die Glorifizierung des Automanagers und seiner Ideen als Rettung für den Arbeitsmarkt.
12. Nach Ansicht von Unionsfraktionsvize Wolfgang Schäuble (CDU) betreibt **Rot-Grün** eine "unverantwortliche" Außen- und Verteidigungspolitik.
13. Schäuble forderte eine klare Antwort, ob **Rot-Grün** die möglichen "ernsten Konsequenzen" der UN-Resolution zu Irak mittragen werde.
14. Der CDU-Abgeordnete Karl Lamers forderte **Rot-Grün** auf: "Machen Sie Schluss mit dem deutschen Sonderweg."
15. CDU und CSU würden **Rot-Grün** weder im Bundestag noch im Bundesrat die Hand zu Steuererhöhungen reichen.
16. Und man darf annehmen, dass Keitel - hinter verschlossener Tür, versteht sich - ob der Steuerpläne von **Rot-Grün** bisweilen Flüche von schwejkscher Derbheit ausstößt.
17. **Rot-Grün** hob die ursprünglich im Gesetzentwurf fixierte Grenze von 50 auf 52 Jahre an.
18. Nach dem alten Dienstmädchenprivileg, das **Rot-Grün** zum 1. Januar 2002 ersatzlos gestrichen hatte, konnten sie noch das siebenfache - 18.000 D-Mark - absetzen.
19. Die Spitzen von **Rot-Grün** hatte sich im Laufe der Woche und nach der Bundestagsanhörung auf etwa ein halbes Dutzend inhaltliche Änderungen und Ergänzungen gegenüber dem Ursprungsentwurf verständigt.
20. Dazu hat **Rot-Grün** ein "stark vereinfachtes und unbürokratische Verfahren" vorgesehen ohne Versicherungsnummer und ohne Sozialversicherungsausweis.
21. Unionsfraktionsvize Wolfgang Schäuble (CDU) sagte, **Rot-Grün** betreibe eine unverantwortliche Außen- und Verteidigungspolitik.
22. Dazu hat **Rot-Grün** ein "stark vereinfachtes und unbürokratische Verfahren" vorgesehen - als Anreiz, solche Beschäftigungsverhältnisse zu legalisieren.
23. Auch hatten die Spitzen von **Rot-Grün** vereinbart, von 2003 an etwa 150 Millionen Euro der 1,4 Milliarden in ein Altbausanierungsprogramm fließen zu lassen.
24. Im Bereich Rente einigten sich **Rot-Grün** auf eine Erhöhung des Beitragssatzes von 19,1 auf 19,5 Prozent.
25. Die Spitzen von **Rot-Grün** hatten sich im Laufe der Woche und nach der Bundestagsanhörung auf etwa ein halbes Dutzend inhaltliche Änderungen und Ergänzungen gegenüber dem Ursprungsentwurf verständigt.
26. Der CDU-Abgeordnete Michael Meister sagte, mit den Neuregelungen erweise sich wieder einmal, dass **Rot-Grün** vor allem für höhere Steuern stehen.
27. Er sprach von einer "sozialistischen Umverteilungspolitik", mit der **Rot-Grün** das Land "in den Abgrund" reiße.

Abb. 5: Belegstellen zu **Rot-Grün** als Wort des Tages vom 15. 11. 2002.

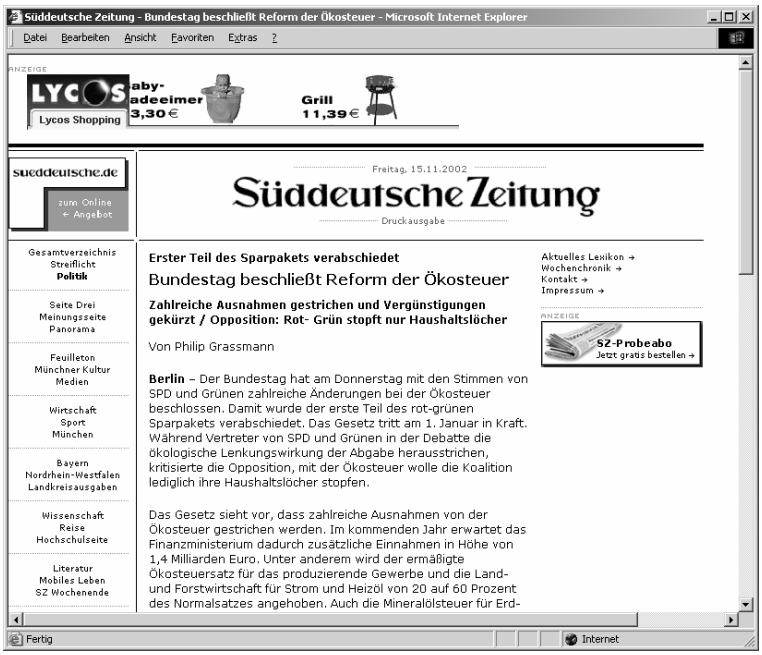

Abb. 6: Quelldokument zu Belegstelle 9 in Abb. 5 (**Rot-Grün**).

3.3 Visualisierung begrifflicher Zusammenhänge

Neben der der textbasierten Präsentation einer jeweils aktuellen Auswahl von „Wörtern des Tages“ lassen die zugrundeliegenden Text Mining-Verfahren auch eine **Visualisierung begrifflicher Zusammenhänge** bzw. die Darstellung des **zeitlichen Verlaufs der Aktualität** von Wörtern des Tages zu.

3.3.1 Assoziationsgrafen

Zu den Ergebnissen der Text Mining-Analyse gehört die satzbasierte Berechnung signifikanter Kollokationen des jeweiligen Ausgangsbegriffs. Obwohl die für einen Tag anfallende Textmenge als Analysekorpus vergleichsweise gering ist, lassen sich für in diesem Korpus relativ häufig auftretenden Begriffe Kollokationsmengen berechnen, in denen Konzepte enthalten sind, die zusammen mit dem Ausgangsbegriff signifikant häufig auftreten. Solche Kollokationsmengen können unter Anwendung

eines **simulated annealing**-Verfahrens als interaktiver Graf visualisiert werden (vgl. Davidson & Harel 1996, Schmidt 1999).

Abbildung 7 zeigt für den derzeit häufig als **Wort des Tages** vertretenen Namen **Eichel** (Bundesfinanzminister Hans Eichel) zwei tagesaktuelle Assoziationsgrafen (15. November 2002 und 9. September 2002) sowie den Assoziationsgrafen aus dem Referenzkorpus. Dabei wird deutlich, dass die den Tagescorpora entnommenen Grafen jeweils aktuelle Relationen hervorheben (im September 2002 (Abb. 7a): **EU-Partner**, **Zusage**, **Wachstumsraten;** im November 2002 (Abb. 7b): **Haushaltsmisere**, **Wahlbetrug**;), während im Referenzgrafen (Abb. 7c) eher grundsätzliche Beziehungen deutlich werden (**Finanzminister**, Namen von Ministerkollegen, Bezüge aus der Vergangenheit (**Landeregierung Hessens**)).

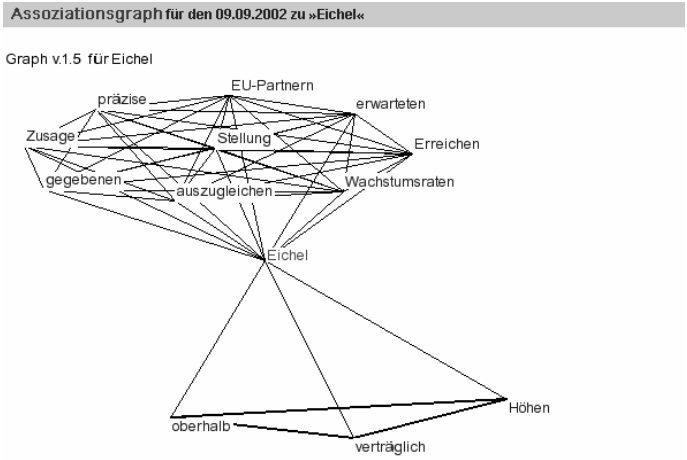

Abb. 7a: Assoziationsgrafen für **Eichel** aus dem Tageskorpus (9.9.2002).

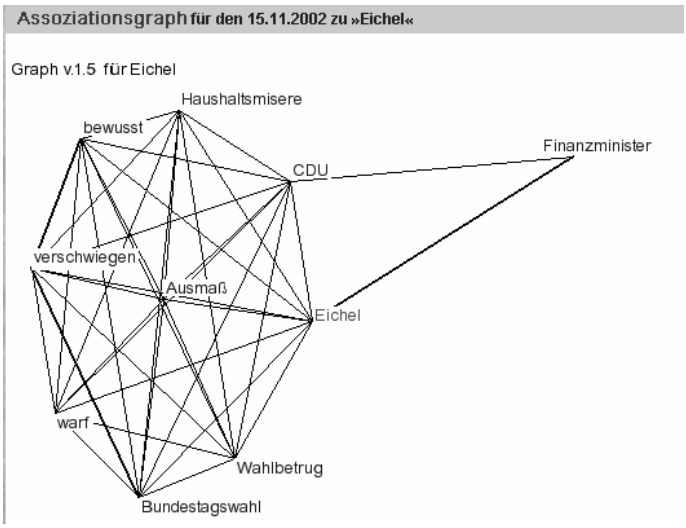

Abb. 7b: Assoziationsgrafen für **Eichel** aus dem Tageskorpus (15.11.2002).

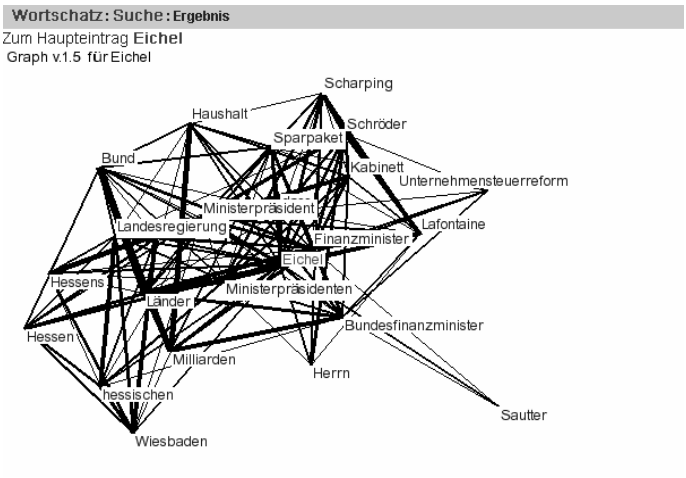

Abb. 7c: Assoziationsgrafen für **Eichel** aus dem Referenzkorpus.

3.3.2 Aktualitätsverläufe von Wörtern des Tages

Es liegt nahe, die jeweils tagesbezogenen Ergebnisse zu den Wörtern des Tages auch als längerfristige Entwicklungen bzw. Trends zu untersuchen. Da die Analysecorpora für jeden Tag gesondert verfügbar sind, lassen sich die Auswahlmengen unmittelbar miteinander vergleichen, wobei jeweils vorausgesetzt ist, dass der Analyse dieselben Quellen sowie eine vergleichbare Textmenge zugrunde liegen.

Ein solcher Aktualitätsverlauf ist allerdings nur für Begriffe sinnvoll, die nicht nur lokal innerhalb eines kurzen Zeitraums in die Wörter des Tages aufgenommen werden, sondern die mittel- und langfristig über eine gewisse Mindestaktualität verfügen.

Die Darstellung erfolgt mit Hilfe eines Liniendiagramms, das sich als Diagrammformat für die Darstellung von Verlaufsentwicklungen in besonderer Weise eignet (vgl. Wolff 1996:117ff). Dabei werden als Grafen jeweils die relativen Aktualitätswerte des ausgesuchten Begriffs sowie seiner jeweils stärksten Kollokationen angezeigt. Zusätzlich werden diejenigen Tage markiert, an denen der jeweilige Begriff unter den Wörtern des Tages war.

Zwei Beispiele sollen dies illustrieren: In Abbildung 8 wird der Aktualitätsverlauf zu **Michael Schumacher** dargestellt. Deutlich erkennbar ist dabei der Anstieg im 14-Tages-Rhythmus, der dem Verlauf der Formel 1-Rennplanung entspricht. Auch im zweiten Beispiel, **Bush**, (Abb. 9) sieht man eine durch die Verabschiedung der UN-Resolution zur Waffenkontrolle im Irak bedingte Aktualitätsspitze.

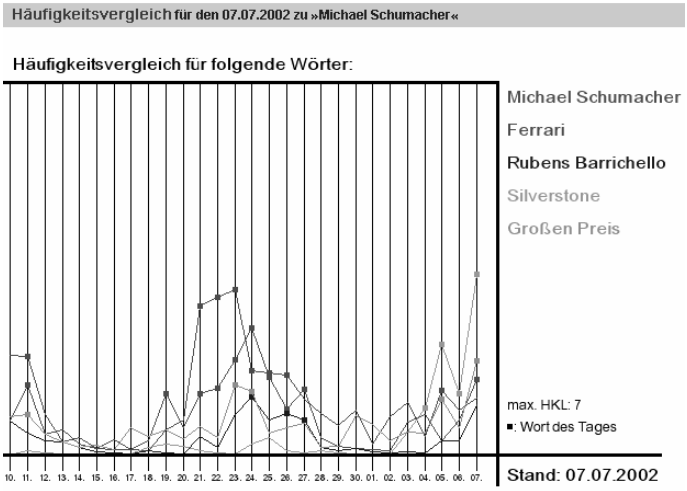

Abb. 8: Aktualitätsverlauf zu **Michael Schumacher** (7.7.2002).

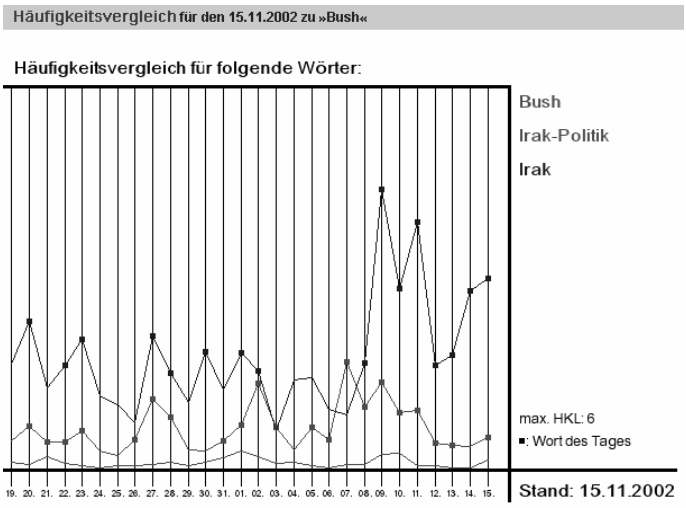

Abb. 9: Aktualitätsverlauf zu **Bush** (15.11.2002).

4 Fazit: Bewertung und Nutzung

In den Kandidatenlisten für die **Wörter des Tages** (vgl. oben Abbildung 2) lassen sich offensichtliche Fehler finden wie

- Vollformen desselben Konzepts (Senegal, Senegals, Stolpe, Stolpes),
- autoreferentielle Bezüge zu den Medien selbst (Medienberichte, NZZ) oder
- Dopplungen von Mehrwortbegriffen und deren Teilworten (Rubens, Rubens Barrichello).

Jenseits solcher einfach zu bereinigender Schwächen stellt sich die Frage nach einer Bewertung der jeweils getroffenen Auswahl von Ereignissen und Personen. Selbst bei exakter Definition von Randparametern der Analyse wie Medienauswahl, Zeit, Ort, gesellschaftlich-politisches Bezugssystem oder Zusammensetzung und Anzahl der Rezipienten ist die Definition **objektivierbarer Kriterien** für die Relevanzmessung von Personen und Ereignissen problematisch.[7]

Insofern ist auch der hier beschriebene Ansatz, durch quantitative Medienanalyse Aussagen über die zeitbezogene Relevanz zu treffen, nur eine Annäherung. Erstre-

[7] Zur Vielfalt die Medien als „multiplexe Systeme“ prägender Systeme vgl. Rusch 2002a: 80ff, insb. Abb. 2.

benswert ist die Entwicklung eines **media impact index**, der eine Bewertung von Ergebnissen, wie sie hier beschrieben sind, zulässt (vgl. dazu Posner 2001).

Es liegt allerdings auf der Hand, dass sich dieser Ansatz in vielfältiger Weise weiterentwickeln und verallgemeinern lässt:

- Durch Berücksichtigung von Gewichtungsfaktoren hinsichtlich der **tatsächlichen Nutzung** der Medien lässt sich das Verfahren nutzen, um nicht nur relevante Ereignisse und Personen zu ermitteln, sondern auch deren unterschiedliche Gewichtung in den Medien zu bestimmen und damit einen Beitrag zur Analyse der Konstitution von Öffentlichkeit durch (Online-)Medien zu leisten.
- Im Kern enthält der voranstehende Ansatz bereits ein vergleichendes Verfahren, da zeitbezogene Quellensammlungen in Bezug zu einem großen Referenzkorpus gesetzt werden. Diese Analysemethode lässt sich auf Untersuchungen übertragen, die unterschiedliche **Typen** von Online-Medien miteinander vergleichen oder die Medienberichterstattung in unterschiedlichen Ländern analysieren.
- Praktische Anwendungen sind hinsichtlich der Trendforschung und Medienwirkungsforschung denkbar, in denen im Sinne der empirischen Medienanalyse für Personen oder Unternehmen nicht nur deren Aktualitätsverläufe, sondern über die Auswertung von signifikanten semantischen Beziehungen auch Inhalt und Struktur der öffentlichen Wahrnehmung dieser Konzepte untersucht werden.

Literatur

Charlton, Michael; Schneider, Silvia (edd.) (1997). Rezeptionsforschung. Theorien und Untersuchungen zum Umgang mit Massenmedien. Opladen: Westdeutscher Verlag.

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

Gerhards, Jürgen; Neidhardt, Friedhelm; Rucht, Dieter (1998). Zwischen Palaver und Diskurs. Strukturen öffentlicher Meinungsbildung am Beispiel der deutschen Diskussion zur Abtreibung. Opladen: Westdeutscher Verlag.

Großmann, Brit (1999). Medienrezeption. Bestehende Ansätze und eine konstruktivistische Alternative. Opladen: Westdeutscher Verlag.

Heyer, Gerhard; Läuter, Martin; Quasthoff, Uwe; Wolff, Christian (2001). „Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse". In: Lobin, H. (ed.) (2001). Sprach- und Texttechnologie in digitalen Medien. Proc. GLDV-Jahrestagung 2001, Universität Gießen, 71-83.

Heyer, Gerhard; Quasthoff, Uwe; Wolff, Christian (2000) "Aiding Web Searches by Statistical Classification Tools. " In: Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz (2000), 163-177.

Heyer, Gerhard; Quasthoff, Uwe; Wolff, Christian (2002). "Knowledge Extraction from Text: Using Filters on Collocation Sets." In: Proc. LREC-2002. Third International Conference on Language Resources and Evaluation. Las Palmas, May 2002, Vol. III, 241-246.

Ludes, Peter (1998). Einführung in die Medienwissenschaft. Berlin: Erich Schmidt.

Merten, Klaus (1999). Einführung in die Kommunikationswissenschaft. Bd. 1: Grundlagen der Kommunikationswissenschaft. Münster: Lit.

Posner, Richard A. (2001). Public Intellectuals. Cambridge/MA.: Harvard University Press.

Quasthoff, Uwe; Biemann, Christian; Wolff, Christian (2002). „Named Entity Learning and Verification: Expectation Maximization in Large Corpora". In: Roth, Dan; van den Bosch, Antal (edd.) (2002). Proc. 6th Conf. on Natural Language Learning 2002 (CoNLL-2002). New Brunswick/NJ: The Association for Computational Linguistics, 8-14 [= Coling 2002 post-conference workshop].

Quasthoff, Uwe; Wolff, Christian (2000) "An Infrastructure for Corpus-Based Monolingual Dictionaries". In: Proc. LREC-2000. Second International Conference on Language Resources and Evaluation. Athens, May / June 2000, Vol. I, 241-246.

Quasthoff, Uwe; Wolff, Christian (2002). „The Poisson Collocation Measure and its Applications". In: Proc. Second International Workshop on Computational Approaches to Collocations, Wien, Juli 2002 [erscheint].

Rusch, Gebhard (2002b). „Medienwissenschaft als transdisziplinäres Forschungs-, Lehr- und Lernprogramm". In: Rusch (2002a), 69-82.

Rusch, Gebhard (ed.) (2002a). Einführung in die Medienwissenschaft. Konzeptionen, Theorien, Methoden, Anwendungen. Opladen: Westdeutscher Verlag.

Schmidt, Fabian (1999). Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung. Diplomarbeit, Universität Leipzig, Institut für

Informatik, Abt. Automatische Sprachverarbeitung, April 1999 [online verfügbar unter: http://dol.uni-leipzig.de/pub/1999-18].

Szyszka, Peter (ed.) (1999). Öffentlichkeit. Diskurs zu einem Schlüsselbegriff der Organisationskommunikation. Opladen: Westdeutscher Verlag.

Wolff, Christian (1996). Graphisches Faktenretrieval mit Liniendiagrammen. Gestaltung und Evaluierung eines experimentellen Rechercheverfahrens auf der Grundlage kognitiver Theorien der Graphenwahrnehmung. Konstanz: UVK Informationswissenschaft [= Schriften zur Informationswissenschaft, Bd. 24, 1996].