In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 369 – 372

„Wörter des Tages“ - Tagesaktuelle wissensbasierte Analyse und Visualisierung von Zeitungen und Newsdiensten

Uwe Quasthoff[1], *Matthias Richter*[1], *Christian Wolff*[2]

[1]Universität Leipzig,
Institut für Informatik,
Augustusplatz 10/11, 04109 Leipzig
quasthoff@informatik.uni-leipzig.de,
matthias@vielfalt.de

[2]Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaften
93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Abstract
In this paper we describe a web service which presents media analysis results as both, structured lists of *Words of the Day*, as well as visualizations of relevant concepts in their semantic context and concept relevance timelines. Media analysis is based on a daily collection of online newspapers and news services. For each day, this collection is processed by a text mining suite.

1 Einführung

Dieser Beitrag versucht aufzuzeigen, wie Text Mining-Verfahren für die Analyse von Medienprodukten 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. Die „Wörter des Tages“ wurden im Rahmen des Projekts „Deutscher Wortschatz“ am Institut für Informatik der Universität Leipzig entwickelt (vgl. Quasthoff & Wolff 2000) und sind im World Wide Web unter http://www.wortschatz.uni-leipzig.de/wort-des-tages/ verfügbar.

2 Tagesaktuelle Analyse von Online-Informationsdiensten

Die Analyse von Online-Texten umfasst folgende Schritte:

- Quellenauswahl und –erfassung

- Quellenanalyse durch Text Mining
- Begriffsselektion (Kandidaten für „Wörter des Tages“)
- Kategorisierung und Überarbeitung von Kandidatenlisten

Die Quellenauswahl erfolgt unter Heranziehung technischer, quantitativer und qualitativer Merkmale. Dabei werden vor allem die Online-Sites überregionaler Medien (z. B. www.sueddeutsche.de, www.spiegel.de). Zum derzeitigen Stand der Quellengrundlage umfasst ein solches Tagescorpus ca. 20.000 Sätze. 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:

0. Bereitstellung eines um den Faktor 1000 umfangreicheren Referenzcorpus.
1. Textsegmentierung, insbesondere Satz- und Wortsegmentierung.
2. Indexierung der Beispieltexte und quantitative Erfassung der Einzelbegriffe.
3. Berechnung relevanter Satz- und Nachbarschaftskollokationen.
4. Speicherung der Ergebnisse in einer relationalen (Tages-)Datenbank.

Nach Durchführung der Text Mining-Analyse für das jeweilige Tagescorpus 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: Frequenz im aktuellen Tagescorpus, Frequenz im Referenzcorpus (Deutscher Wortschatz) und relative Übergewichtung eines aktuellen Begriffs.

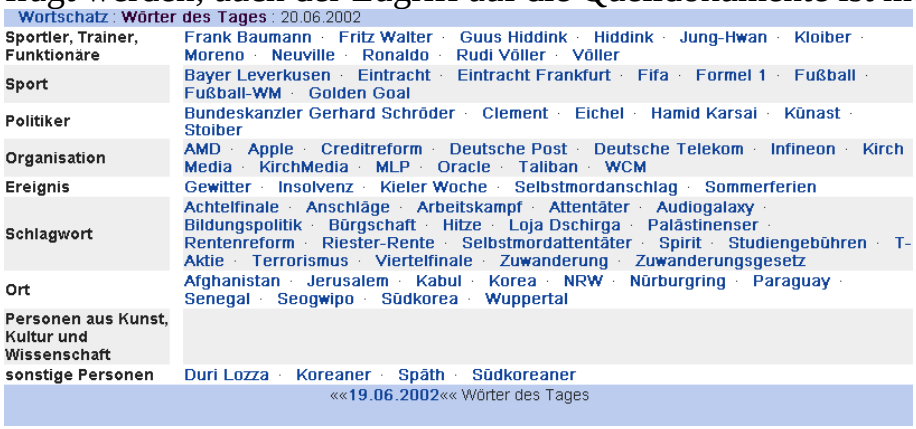
3 Ergebnisaufbereitung und Visualisierung

Die Ausgabe der Wörter des Tages lässt sich in folgende Bereiche gliedern:

- Ausgabe der Wörter des Tages (nach Kategorien geordnet)
- Ausgabe der Belegstellensammlung zu einem Wort des Tages
- Ausgabe eines Assoziationsgraphen, der Beziehungen eines Wortes des Tages zu anderen im Tagescorpus enthaltenen Konzepten darstellt
- Graphische Ausgabe des „Aktualitätsverlaufs“ von Worten 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 (Abb. 1). Bei der Analyse werden Belegstellen für die verschiedenen

„Wörter des Tages“ in einer Datenbank gesammelt und können online abgefragt werden; auch der Zugriff auf die Quelldokumente ist möglich.



Wortschatz : Wörter des Tages : 20.06.2002

Sportler, Trainer, Funktionäre	Frank Baumann · Fritz Walter · Guus Hiddink · Hiddink · Jung-Hwan · Kloiber · Moreno · Neuville · Ronaldo · Rudi Völler · Völler
Sport	Bayer Leverkusen · Eintracht · Eintracht Frankfurt · Fifa · Formel 1 · Fußball · Fußball-WM · Golden Goal
Politiker	Bundeskanzler Gerhard Schröder · Clement · Eichel · Hamid Karsai · Künast · Stoiber
Organisation	AMD · Apple · Creditreform · Deutsche Post · Deutsche Telekom · Infineon · Kirch Media · KirchMedia · MLP · Oracle · Taliban · WCM
Ereignis	Gewitter · Insolvenz · Kieler Woche · Selbstmordanschlag · Sommerferien
Schlagwort	Achtelfinale · Anschläge · Arbeitskampf · Attentäter · Audiogalaxy · Bildungspolitik · Bürgschaft · Hitze · Loja Dschirga · Palästinenser · Rentenreform · Riester-Rente · Selbstmordattentäter · Spirit · Studiengebühren · T-Aktie · Terrorismus · Viertelfinale · Zuwanderung · Zuwanderungsgesetz
Ort	Afghanistan · Jerusalem · Kabul · Korea · NRW · Nürburgring · Paraguay · Senegal · Seogwipo · Südkorea · Wuppertal
Personen aus Kunst, Kultur und Wissenschaft	
sonstige Personen	Duri Lozza · Koreaner · Späth · Südkoreaner

««19.06.2002«« Wörter des Tages

Abbildung 1: Webbasierte Ausgabe der *Wörter des Tages* vom 20. Juni 2002

Neben 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. Abb. 2 zeigt für das derzeit häufig als „Wort des Tages“ vertretene Konzept *T-Aktie* den tagesaktuellen Assoziationsgraphen vom 19. Juni 2002 sowie den Assoziationsgraphen aus dem Referenzcorpus. Dabei wird deutlich, dass die den Tagescorpora entnommenen Graphen jeweils aktuelle Relationen hervorheben (z. B. *Aktienoptionen*), während im Referenzgraphen eher grundsätzliche Beziehungen deutlich werden (*Volksaktie, Dax, Ausgabekurs, Ron Sommer*).

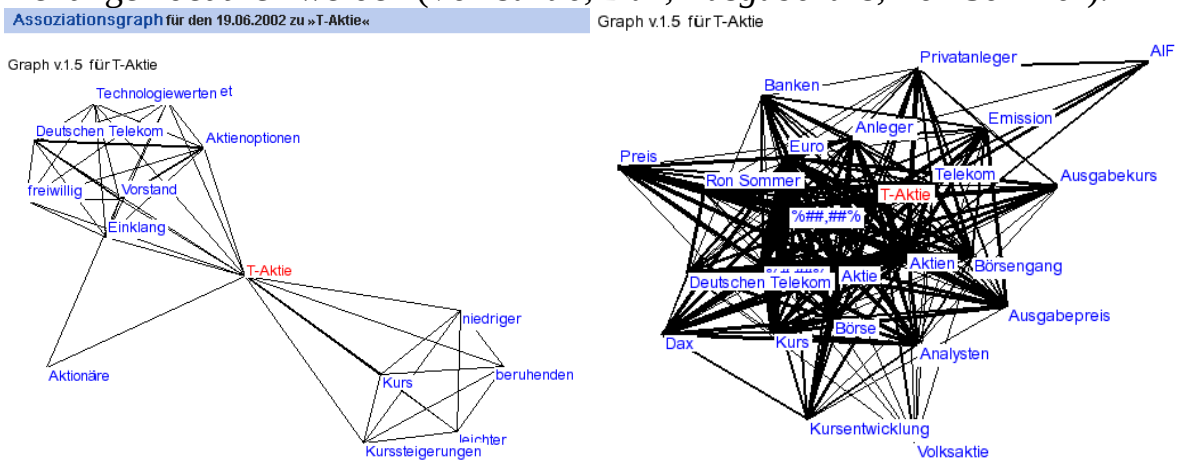


Abbildung 2: Graphen für *T-Aktie* aus Tagescorpus (links) bzw. Referenzcorpus (rechts)

Es liegt nahe, die tagesbezogenen Ergebnisse zu den Wörtern des Tages auch als längerfristige Entwicklungen zu untersuchen. Da die Analysecorpora für jeden Tag gesondert verfügbar sind, lassen sich die Auswahlmengen unmittelbar miteinander vergleichen. Die Darstellung erfolgt mit Hilfe eines Liniendiagramms, das sich für die Darstellung von Verlaufsentwicklungen sehr gut

eignet. Dabei werden als Graphen 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. In Abb. 3 wird der Aktualitätsverlauf zu *Michael Schumacher* dargestellt. Deutlich erkennbar ist dabei der Anstieg im 14-Tages-Rhythmus (Formel 1-Kalender).

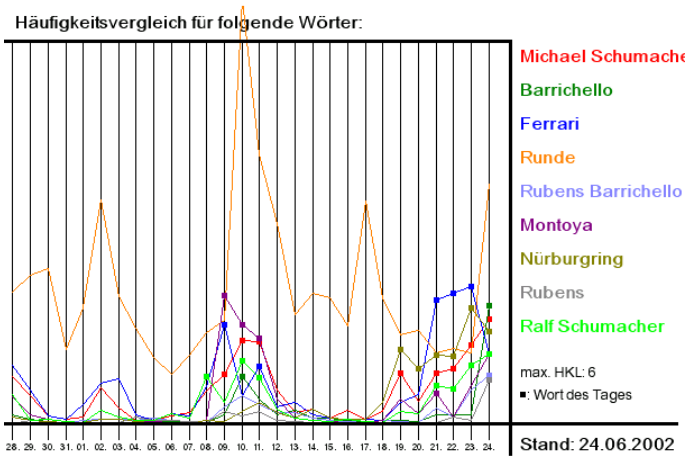


Abbildung 3: Aktualitätsverlauf zu *Michael Schumacher*

4 Fazit: Bewertung und Nutzung

Abschließend 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. Insofern ist auch der hier beschriebene Ansatz, durch quantitative Medienanalyse Aussagen über die zeitbezogene Relevanz zu treffen, nur eine Annäherung. Erstrebenswert ist die Entwicklung eines *media impact index*, der eine Bewertung von Ergebnissen, wie sie hier beschrieben sind, zulässt (vgl. dazu Posner 2001).

5 Literatur

Quasthoff, Uwe; Richter, Matthias; Wolff, Christian (2002). Wörter des Tages - medienanalytische Motivation, texttechnologische Konzeption und Realisierung als Web Service. Technical Paper, Universität Leipzig, Institut für Informatik, Juli 2002.

Quasthoff, Uwe; Wolff, Christian; "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.

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