

Nutzen und Benutzen von Text Mining für die Medienanalyse

Von der Fakultät für Mathematik und
Informatik der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades
DOCTOR philosophiae
(Dr. phil.)

im Fachgebiet
Informatik

vorgelegt

von Mattias Richter, M.A.

geboren am 27. November 1976 in Hof a.d. Saale

Die Annahme der Dissertation haben
empfohlen:

1. Prof. Dr. Gerhard Heyer, Universität Leipzig
2. Prof. Dr. Maximilian Eibl, Technische
Universität Chemnitz
3. Prof. Dr. Christian Wolff, Universität
Regensburg

Die Verleihung des akademischen Grades
erfolgt auf Beschluss des Rates der Fakultät für
Mathematik und Informatik vom 15.11.2010 mit
dem Gesamtprädikat *cum laude*.

*We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.*

— Donald E. Knuth

DANKSAGUNGEN

Diese Dissertation wäre mir nicht möglich gewesen ohne die Medienstiftung der Sparkasse Leipzig, welche dem Vorhaben mit einem Stipendium den finanziellen Rückhalt gab und durch ihre Förderung an der Schnittstelle von Medien und Informatik interessante Kontakte ermöglichte.

Der Abteilung Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig gebührt Dank dafür, dass ich, ein nicht-Diplominformatiker, als gleichwertiger Kollege angenommen wurde, dass mir viel Freiheit zur Forschung gewährt wurde und ich, wo es nötig war, den entsprechenden Rückhalt gefunden habe.

Doch am wenigsten hätte ich das Projekt ohne den Rückhalt meiner Familie vollendet, besonders nicht ohne Dich Ina, die Du deswegen so viel mit mir durchmachen musstest.

INHALTSVERZEICHNIS

Abbildungsverzeichnis v

Tabellenverzeichnis viii

1	EINLEITUNG	1
1.1	Sinn einer wissenschaftlichen Arbeit zu Beginn des 21. Jahrhunderts	1
1.2	Verortung der Arbeit in der Ordnung der Wissenschaften	1
1.3	Vor dem Text	2
1.4	Beitrag zu Forschung und Praxis	3
1.5	Anlage und Aufbau der Arbeit	4
2	GRUNDLAGEN	5
2.1	Textdaten	5
2.1.1	Zeichen	5
2.1.2	Verweise	6
2.1.3	Encoding	6
2.1.4	Umwandlung	7
2.2	Untersuchungsobjekte	7
2.2.1	Begriffe	7
2.2.2	Verteilung	8
2.2.3	Kookkurrenzen	12
2.3	Exkurs: Ein Verteilungsexperiment	12
2.3.1	Setup	12
2.3.2	Einfluss der Samplegröße	14
2.3.3	Einfluss der Korpusgröße	14
2.3.4	Wiederauftauchen von Types und Kookkurrenzen	14
2.4	Zeit	18
2.4.1	Definition	18
2.4.2	Betrachtungsarten	18
2.4.3	Zeitreihenanalyse	18
2.5	Wahrheit und Information	19
3	ZUGÄNGE ZU TEXT	21
3.1	Inhaltsanalyse	21
3.1.1	Geschichte	21
3.1.2	Vorgehen	22
3.1.3	Kritik	23
3.1.4	Mit Computer	23

3.1.5	Medienresonanzanalyse	24
3.1.6	Exkurs: Automatische Analyse von Meinungen und Einstellungen	25
3.1.7	Ein anderer Zugang zu Text durch Text Mining	26
3.2	Beispiele	27
3.2.1	Nachrichtensuchmaschinen	27
3.2.2	Nachrichtenzusammenfassungen	28
3.2.3	Nachrichtenüberblicke	29
4	DIE WÖRTER DES TAGES	34
4.1	Einordnung und Ursprung	34
4.1.1	Projekt Deutscher Wortschatz	34
4.1.2	Idee zu „Wörtern des Tages“	37
4.1.3	Verwandte Ansätze und Arbeiten	38
4.2	Archivierung	39
4.2.1	Zur Funktion von Archiven	40
4.2.2	Rechtliche Rahmenbedingungen	40
4.3	Implementierung	44
4.3.1	Daten und Datenacquire	45
4.3.2	Vorverarbeitung	50
4.3.3	Linguistische Aufbereitung	54
4.3.4	Tägliche Verarbeitung	58
4.3.5	Präsentation	65
4.3.6	Evaluation	70
4.4	Weiterentwicklungen und Perspektiven	71
4.4.1	Anwendungen	71
4.4.2	Mashup	74
4.4.3	Medien- und Trendanalyse	78
5	SCHLUSS	84
A	WEITERE BEISPIELE AUS DER ANWENDUNG	85
A.1	Wirtschaft	85
A.2	Papst: Tod und Neuwahl	87
A.3	Weltsicherheitsrat	93
B	LISTINGS	94
C	DATENBANKSCHEMA	110
D	WISSENSCHAFTLICHER WERDEGANG	112
E	PUBLIKATIONEN	113
	LITERATURVERZEICHNIS	114

ABBILDUNGSVERZEICHNIS

Abbildung 1	Zipf-Verteilung: 1,5 Mio. Sätze SPIEGEL und 2 Mio. Sätze DIE ZEIT	9
Abbildung 2	Zipf-Verteilung: Je 100 000 Sätze Deutsch, Französisch und Türkisch	10
Abbildung 3	Wiederauftreten: Types, Nachbarn, Kookkurrenzen	15
Abbildung 4	Wiederauftreten der top n Types (S=1, K=40)	15
Abbildung 5	Wiederauftreten der top n Types (S=125, K=1) vs. (S=1, K=40)	
Abbildung 6	Wiederauftreten von Kookkurrenzen der Top n Types (S=1, K=40)	16
Abbildung 7	Wiederauftreten von Kookkurrenzen der Top n Types (S=125, K=1)	17
Abbildung 8	Anzahl der Veröffentlichungen pro Jahr in der sentiment-analysis-Bibliographie von Andrea Esuli.	26
Abbildung 9	News in Essence	28
Abbildung 10	Newsblaster: Kategorie <i>world</i> am 6.12.2008	29
Abbildung 11	EMM: Clusteransicht	30
Abbildung 12	EMM: Schreibungen (inklusive Fehlschreibungen), Titel und Bild zur Named Entity „Angela Merkel“	30
Abbildung 13	JRC Newsbrief: Top-10 Nachrichtenübersicht	30
Abbildung 14	JRC Newsbrief: Themen, Artikel und Personen (Deutschland, 6.12.2008)	
Abbildung 15	Textmap: Netzwerk für „Angela Merkel“	31
Abbildung 16	Textmap: „Angela Merkel“ – Verlauf der Verteilung der Berichte auf Oberkategorien	
Abbildung 17	Textmap: Juxtapositions für „Angela Merkel“ in unterschiedlichen Zeitintervallen	32
Abbildung 18	Textmap: Gegenüberstellung des zeitlichen Verlaufs von Frequenz und Sentiment	33
Abbildung 19	Das Wortschatz-Portal (Stand: 13. Dezember 2007)	35
Abbildung 20	Grober Überblick: Ablaufplan Wörter des Tages	45
Abbildung 21	Beispielartikel von Financial Times Deutschland (22. Januar 2007).	51

- Abbildung 22 Spinnennetzdiagramm für den Begriff „Subventions-Heuschrecke“ am 17. Januar 2008 mit Kategorien von Spiegel Online. Die fette schwarze Linie ist die Nulllinie. 62
- Abbildung 23 Kookkurrenz-Cluster der Wörter des Tages vom 12. Februar 2003. 63
- Abbildung 24 Übersichtsseiten 66
- Abbildung 25 Detailansicht für *Ord i Dag* „fugleinfluensa“ (norw. *Vogelgrippe*) am 15. März 2006. 68
- Abbildung 25 Detailansicht für *Ord i dag* „fugleinfluensa“ (dt. *Vogelgrippe*) am 15. März 2006. 69
- Abbildung 26 Pressebarometer aus Wörtern des Tages (Stand: 6. Juni 2005) 72
- Abbildung 27 Schröder vs. Stoiber (Wahl am: 22. September 2002) 72
- Abbildung 28 Schröder vs. Merkel (Wahl am: 18. September 2005) 72
- Abbildung 29 Treemap-Darstellungen für News. 75
- Abbildung 30 Architektur des Mashups von Google Maps und Wörtern des Tages 77
- Abbildung 31 Mashup aus Google Maps und Wörtern des Tages: Anzeige im Browser. 78
- Abbildung 32 Nachrichtenkarte von romso.de für den 22. November 2006, 14:00 Uhr. 79
- Abbildung 33 Kookkurrenzgraphen zum Begriff »Handgepäck« in der deutschen Presse für den 11. und 12. August 2006. 80
- Abbildung 34 Frequenzverlauf des Begriffs »Vogelgrippe« in der deutschen Presse vom 1. Januar 2003 – 1. August 2006. 80
- Abbildung 35 Kookkurrenzgraphen für *Vogelgrippe* 81
- Abbildung 36 Kookkurrenzgraphen zum Begriff »Handgepäck« für den 11. August 2006 in der norwegischen und deutschen Presse. 82
- Abbildung 37 Annotierter Verlaufsgraph zum Begriff *hybrid* 86
- Abbildung 38 Annotierter Verlaufsgraph zum Begriff *Unternehmertum* 86
- Abbildung 39 Annotierter Verlaufsgraph zum Begriff *Innovationsoffensive* 86
- Abbildung 40 *Papst*: Frequenzverlauf 88
- Abbildung 41 *Papst* mit Kookkurrenzen vom 28. März 2005: Die Stimme versagt beim Ostersegen 88

- Abbildung 42 *Papst* mit Kookkurrenzen vom 4. April 2005:
Die Kirche, Polen, die ganze Welt blickt nach
Rom 89
- Abbildung 43 *Papst* mit Kookkurrenzen vom 5. April 2005:
Papst Johannes Paul II. ist gestorben 89
- Abbildung 44 *Papst* mit Kookkurrenzen vom 8. April 2005:
Testamentsverkündung und Abschied 90
- Abbildung 45 *Papst* mit Kookkurrenzen vom 18. April 2005:
Das Konklave steht bevor 90
- Abbildung 46 *Papst* mit Kookkurrenzen vom 21. April 2005:
Der Deutsche Joseph Ratzinger wurde zum
Papst gewählt und trägt nun den Namen
Benedikt XVI. 91
- Abbildung 47 *Papst* mit Kookkurrenzen vom 25. April 2005:
Der erste Segen auf dem Petersplatz durch
Joseph Ratzinger 91
- Abbildung 48 *Joseph Ratzinger* mit Kookkurrenzen vom 21.
April 2005: Ratzinger stand vorher an der
Spitze der Glaubenskongregation der ka-
tholischen Kirche 92
- Abbildung 49 Kookkurrenzgraph für *Joseph Ratzinger* am
21. April 2005 92
- Abbildung 50 Auflösung des Begriffes *Wahl*: Nicht nur
ein Papst wurde im März/April 2005 ge-
wählt. Der Grauwert kodiert die Stärke der
Kookkurrenz. 93
- Abbildung 51 Wechsel des Fokus vom Iran zum Irak im
Weltsicherheitsrat 93

TABELLENVERZEICHNIS

Tabelle 1	Durchschnittliche Anzahl Tokens in Abhängigkeit von Korpus- und Samplegröße ☞ im Mittel kaum ein Unterschied	13
Tabelle 2	Durchschnittliche Anzahl Types in Abhängigkeit von Korpus- und Samplegröße ☞ mehr Zufall, mehr Types	13
Tabelle 3	Durchschnittliche Anzahl Satzkookkurrenzen in Abhängigkeit von Korpus- und Samplegröße ☞ mehr Zufall, weniger Kookkurrenzen	13
Tabelle 4	Durchschnittliche Anzahl Nachbarschaftskookkurrenzen in Abhängigkeit von Korpus- und Samplegröße ☞ Kein so starker Einfluss auf Nachbarschaftskookkurrenzen	13
Tabelle 5	Seitenabrufe und Besucherzahlen im Lauf der Zeit	37
Tabelle 6	Prozentuale Textabdeckung der top n types	52
Tabelle 7	Reduktionsregeln für die Vollformen des norwegischen Wortes „stå“ (dt. <i>stehen</i>).	57

EINLEITUNG

*The beginning of knowledge
is the discovery of something
we do not understand.*

— Frank Herbert

1.1 SINN EINER WISSENSCHAFTLICHEN ARBEIT ZU BEGINN DES 21. JAHRHUNDERTS

Aktuelle Leitbegriffe wie Informationsüberflutung, postmoderne Aufmerksamkeitsökonomie sowie allgegenwärtige Techniken der Komplexitätsreduktion verlangen einer wissenschaftlichen Arbeit heute in besonderem Maße ab, knappen Überblick des Wesentlichen mit klarem Einblick in das Entscheidende zu verbinden. Dies bedeutet gerade nicht, auf breite und tiefe systematische Recherche und Literaturwürdigung zu verzichten, sondern fordert vielmehr, auf deren Grundlage das Relevante und Essentielle vom bloß Vorhandenen zu (unter-)scheiden, ja: zu befreien. Bleiben dabei am Ende wenige stringente Gedankengänge übrig, so mag man dies vielleicht wertvoller einschätzen als ein zielloses Referat zahlreicher, auch irrelevanter, Quellen, das doch prinzipbedingt nicht abgeschlossen oder vollständig sein kann.

1.2 VERORTUNG DER ARBEIT IN DER ORDNUNG DER WISSENSCHAFTEN

Die vorliegende Arbeit beschäftigt sich damit, wie bestimmte Techniken aus dem Bereich der Sprache verarbeitenden und analysierenden Informatik, der Textlinguistik und -statistik, dem Text Mining und der Informationsvisualisierung auf Problemstellungen der Medienanalyse treffen. Insbesondere geht es dabei um die Analyse der zeitlich segmentierten Text-Genres *Online News* und *Weblogs* – wobei nicht etwa konkrete Einzelfragen, sondern vielmehr abstrakte Methoden des Handlings und der Erschließung digitaler Textdatenmassen, sowie prinzipielle Herangehensweisen an konkrete Fragen bearbeitet werden. Die tatsächliche Anwendung der hier vorgestellten und exemplarisch vorgeführten Technologie kann in der Praxis dann Einsatzgebiete

te der visuellen Textmedienanalyse wie Forschung, Marketing, Öffentlichkeitsarbeit, Kommunikationscontrolling etc. bedienen.

Nicht alle Disziplinen, welche im Folgenden berührt werden, können in dem Umfang abgehandelt werden, wie der Kern der praktisch informatischen Anforderungen, Ideen, Umsetzungen und Anwendungen; zu diesen Bereichen gilt es methodische Schnittstellen zu beschreiben, Einflüsse zu skizzieren und kommunikative Anschlüsse zu provozieren, soweit dies die Kompetenz eines einzelnen Autors eben zu bündeln vermag. Interdisziplinarität ist dabei eine Art des ambitionierten Sich-zwischen-die-Stühle-setzens – am richtigen Platz reicht es vielleicht für eine eigene Nische, die von allen Richtungen geachtet wird; am falschen Platz taugt es intellektuell wie materiell zum Verhungern – in der Hoffnung auf Ersteres.

1.3 VOR DEM TEXT

Vor dem Text und im Text stecken unter anderem Schriftlichkeit, Schriftzeichen, Wörter, Struktur, Sinn, Ausdruck, Meinung, Wissen, Wahrheit, Fakten, Fiktion, Gedanken, Wille, Kommunikation, Selbstbezüge und Fremdbezüge, Anschlüsse, Fortführungen und Abschlüsse. Das Lesen und Schreiben, das Gebrauchen und Verfassen von Texten, ist eine spezifische Form der Kommunikation und eine Eigenheit unserer Kultur. Türcke (2005) gibt hierüber einen Überblick von Vor-Schriftlichkeit bis Nach-Textlichkeit.

Zugänge zu Texten gibt es viele, z. B. psychologisch, soziologisch, kommunikations-, sprach- oder literaturwissenschaftlich motivierte, vgl. dazu etwa Titscher u. a. (1998). Diese sind an sich nicht richtig oder falsch, sondern unterscheiden sich voneinander in den Unterscheidungen und Selektionen, welche sie treffen, also in den Erkenntnisgegenständen, die sie interessieren. *Text Mining* (siehe Heyer u. a., 2006) betrachtet beliebigen digital(isiert) vorliegenden Text als unstrukturierten Speicher von Wissen und zielt darauf ab, daraus verborgene Schätze in Form von sprachlichen Sinn und Struktur stiftenden Regeln, Entitäten und Relationen usw. mittels Domänen- und Sprachwissen, Statistik und Algorithmen zu extrahieren, zu kondensieren, zu systematisieren und somit verfügbar und anwendbar zu machen. In besonderem Maße eignet sich das sozialwissenschaftliche Verfahren der Inhaltsanalyse dazu durch Text Mining Unterstützung zu finden, weil jene sich auf die manifesten Inhalte fossilierteter Kommunikation Merten (2004) bezieht, auf welche dieses zugreift.

Diesem analytischen Instrumentarium gesellt sich als generel-

ler Rahmen der Zugang über das Auge, die Interaktion mit dem Benutzer, die Anwendung zur Generierung von Mehrwert hinzu zu den Rohdaten von Verfahren. Techniken der *Information Visualization* und Überschneidungen mit den *Visual Analytics* treten hier auf den Plan. Das Ziel ist es, anhand belegter Daten mit nachvollziehbaren Verfahren visuell unterstützte Darstellungen und Erklärungen anzufertigen, bzw. Bausteine dazu zu liefern. In einem englischsprachigen Begriff gefasst bedeutet dies gewissermaßen auch *Visual Media Analytics* zu betreiben, also die trockenen Analysedaten publikumsgerecht und zielführend zum Leben zu erwecken (vgl. Cipollone (2006)).

1.4 BEITRAG ZU FORSCHUNG UND PRAXIS

Der Beitrag dieser Arbeit zur Forschung ist wie folgt: Einerseits werden bestehende Ergebnisse aus so unterschiedlichen Richtungen wie etwa der empirischen Medienforschung und dem Text Mining zusammengetragen. Es geht dabei um Inhaltsanalyse, von Hand, mit Unterstützung durch Computer, oder völlig automatisch, speziell auch im Hinblick auf die Faktoren wie Zeit, Entwicklung und Veränderung. Die Verdichtung und Zusammenstellung liefert nicht nur einen Überblick aus ungewohnter Perspektive, in diesem Prozess geschieht auch die Synthese von etwas Neuem.

Die Grundthese bleibt dabei immer eine einschließende: So wenig es möglich scheint, dass in Zukunft der Computer Analysen völlig ohne menschliche Interpretation betreiben kann und wird, so wenig werden menschliche Interpretatoren noch ohne die jeweils bestmögliche Unterstützung des Rechners in der Lage sein, komplexe Themen zeitnah umfassend und ohne allzu große subjektive Einflüsse zu bearbeiten – und so wenig werden es sich substantiell wertvolle Analysen noch leisten können, völlig auf derartige Hilfen und Instrumente der Qualitätssicherung zu verzichten.

Daraus ergeben sich unmittelbar Anforderungen: Es ist zu klären, wo die Stärken und Schwächen von menschlichen Analysten und von Computerverfahren liegen. Darauf aufbauend gilt es eine optimale Synthese aus beider Seiten Stärken und unter Minimierung der jeweiligen Schwächen zu erzielen. Praktisches Ziel ist letztlich die Reduktion von Komplexität und die Ermöglichung eines Ausganges aus dem Zustand des systembedingten „overnewsed but uninformed“¹-Seins.

¹ So lautet der Titel der Diplomarbeit von Stefan Bräutigam in welcher er dem Phänomen Berichterstattung mit unterschiedlichen Infografiken ästhetisch gelungen und vielfach preisgekrönt zu Leibe rückt.

1.5 ANLAGE UND AUFBAU DER ARBEIT

Diese Arbeit ist um Kürze und Verständlichkeit nicht nur für Informatiker bemüht. Als Publikum wünscht sie sich sogar zuerst interessierte Sozialwissenschaftler und Praktiker. Gleichwohl sind die informatischen Hintergründe nicht zu verleugnen und sollen auch nicht derart in den Hintergrund gerückt worden sein, dass sie nicht mehr zu erkennen sind. Dennoch ist z. B. jeglicher Code und das algorithmische Detail ganz bewusst und absichtlich in den Anhang gerückt.

Zunächst gilt es eine gemeinsame Grundlage für die Sozialwissenschaft und die Computerwissenschaft zu schaffen, und eventuell gegenseitig vorhandenes Unverständnis aufzulösen (in den Kapiteln 2 und 3). Daraufhin werden einige Verfahren vorgestellt, die im Rahmen eines Integrationsprojektes angewandt wurden (in Kapitel 4). Im Anhang findet sich neben weiteren Beispielen für die Anwendung auch die eine oder andere Zeile Quellcode.

Nothing is more usual than for philosophers to encroach on the province of grammarians, and to engage in disputes of words, while they imagine they are handling controversies of the deepest importance and concern.

— David Hume

Dieses einführende Kapitel beschäftigt sich mit grundsätzlichen Dingen wie den verwendeten Begriffen, der Struktur des untersuchten Materials, den Grundlagen für die angewendeten Statistiken usw.

2.1 TEXTDATEN

Textdaten werden hier nicht als Einheiten im Sinne der Literatur verstanden, sondern abstrakt als Mengen von Folgen von Zeichen. Inhaltliche oder formale Durchbrechungen der Linearität von literarischen Texten bleiben insofern irrelevant, als dass jeweils Einheiten betrachtet werden, die an sich linear vorliegen, zum Beispiel verbleibende Resttext an den Knoten eines Hypertextgraphen. Grundvoraussetzung für die automatische Verarbeitung von Textdaten ist, dass die Gleichheit definierter Teilfolgen der Textdaten überprüft werden kann.

2.1.1 Zeichen

Um mit Textdaten sinnvoll arbeiten zu können, müssen Zeichen interpretiert, unterschieden und auf unter Umständen mehreren Stufen zusammengefasst werden. Zeichen sind die kleinsten eigenständigen Teile von Schriftsystemen und verweisen auf konkrete oder abstrakte Bedeutungen. Außer den geläufigen Alphabetschriften wie Lateinisch, Kyrillisch, Griechisch, etc. gibt es noch weitere Arten von Schriftsystemen. Die verbreitetsten davon sind Silbenschriften (z.B. Hangul, Katakana, ...), Segmentalschriften (z.B. Hebräisch, Arabisch, ...) und Ideogrammschriften (z.B. Hanzi, Kanji, ...). Neben den Schriftzeichen sind in Textdaten auch Ziffern, Sonderzeichen, wie Satzzeichen, Rechenzeichen und Währungssymbole, sowie nicht dargestellte Zeichen, wie Leerzeichen, Tabulatoren und Zeilenende, zu finden.

2.1.2 Verweise

Beschränkt man die Betrachtung der einfachen Verständlichkeit halber auf Alphabetschriften, hat jedes dieser Zeichen zweierlei Referenzen: Beschreibungen wie *kleiner lateinischer Buchstabe a* und Darstellungen in Form einer so genannter Glyphen wie *a*, *ā* oder *á*. Grundsätzlich gibt es hierdurch mehrere Probleme: Es gibt Zeichen, welche mehrere Beschreibungen zulassen: so ist *å* sowohl als das Zeichen *å* als auch als Kombination der Zeichen *a* und *°* korrekt beschrieben. Es gibt identisch aussehende Glyphen, welche unterschiedliche Zeichen darstellen: Die Glyphen *P* kann nicht nur für ein lateinisches sondern auch für ein anderes kyrillisches oder ein nochmals anderes griechisches Zeichen stehen. Und schließlich gibt es Glyphen, welche mehrere Zeichen auf einmal darstellen, die so genannten Ligaturen, wie z. B. *æ* (*a + e*) oder *fi* (*f + i*).

2.1.3 Encoding

Zudem gibt es für die digitale Notation der Zuordnung von Zeichenbeschreibungen zahlreiche nicht immer kompatible Ansätze wie zum Beispiel ASCII, EBCDIC, DOS Codepages, ISO-Kodierungen und Unicode. Diesen so genannten *Encodings* gemeinsam ist, dass bestimmte Zeichenbeschreibungen einen bestimmten Zahlenwert zugewiesen bekommen und Zeichensätze anhand dieser Zahlenwerte die erwünschten Glyphen an ein Darstellungsgerät ausgeben können. Ohne Kenntnis des zugehörigen Encodings ist grundsätzlich kein Text sinnvoll lesbar, denn die interne Repräsentation ist nichts als eine Folge von Bytes! Dies würde deutlicher, wenn sich die einzelnen Encodings auch im Bereich der lateinischen Standardschrift mehr unterschieden; da dies aber nicht der Fall ist und deswegen die Angabe und Überprüfung des Encodings oft widersprüchlich ist, also sträflich vernachlässigt wird, und bisweilen sogar fehlerhafterweise Mischungen aus unterschiedlichen Encodings in einer Texteinheit Verwendung finden, ohne dass dies zu unmittelbar auftretenden fatalen Fehlern führt, treten bei der Verarbeitung von Textdaten oftmals erst im Nachhinein sichtbare Mängel und Fehler auf, deren grundsätzliche Ursache zwar klar ist, deren genauer Ursprung aber dennoch zeitaufwändig zu finden sein kann.

2.1.4 *Umwandlung*

Im Folgenden soll davon ausgegangen werden, dass bei Textdaten immer das Encoding bekannt ist und dass die Textdaten vollständig korrekt danach kodiert sind. Dann können mit Hilfe freier Tools wie *iconv* oder GNU *recode* Textdaten kompatibler Encodings ineinander umgewandelt werden. Insofern Vielsprachigkeit von Nöten ist, bietet es sich an als Zielencoding eine Form von Unicode zu wählen, da alle gängigen und auch die untereinander inkompatiblen Encodings durch Transformation in Unicode nebeneinander dargestellt werden können.

Unicode löst zwar nicht alle Probleme prinzipiell, denn weiterhin gibt es verschiedene Wege zusammengesetzte Zeichen und Zeichenzusammensetzungen darzustellen, es ist aber auf Grundlage der Unicoderepräsentation prinzipiell möglich sowohl eine für die Verarbeitung benötigte definierte Auflösung der Kompositionen als auch das Original verlustfrei nebeneinander zu speichern. Die Auflösung von Ligaturen¹ durch ihre äquivalente Darstellung als Einzelzeichen und die Zusammenfassung von Akzenten und Zeichen zu akzentierten Zeichen ist grundsätzlich eine durch Zeitaufwand und Geduld lösbare Aufgabe.

2.2 UNTERSUCHUNGSOBJEKTE

2.2.1 *Begriffe*

Beim Umgang mit Daten natürlicher Sprache ist es hilfreich einige Begriffe zu definieren, um ähnlich Aussehendes mit unterschiedlicher Bedeutung auseinanderhalten zu können. Im Bezug auf die aus Zeichen aufgebauten größeren Einheiten sollen die folgenden unterschieden werden:

- *Wörter* als Begriffe, wie sie in einem Wörterbuch lexikalisiert werden,
- *Types* als Zusammenfassung dessen, was in einem realen Text als separierbare, vergleichbare und unterscheidbare Zeichenketten gefunden werden kann und schließlich
- *Tokens* als die individuellen Vorkommen von Types, welche dem einem oder dem anderen Type zugeordnet werden können (z. B. *gehe* (Token) → *gehe* (Type) → *gehen* (Wort)).

Lemmatisierung bedeutet quasi, Tokens durch bestimmte über entsprechende Types definierte Wörter zu ersetzen. Im Allgemei-

¹ Nicht jedoch das Setzen! Vergleiche z. B. Auflage vs. Abflug

nen wird dies für eine gute Idee gehalten, etwa um die Datenbasis nicht unnötig aufzufächern und um substantieller Aussagen über relativ seltene Wörter zu erhalten. Gleichzeitig ist Lemmatisierung aber auch eine Fehlerquelle, einmal weil sie nicht immer eindeutig ist oder eindeutige Ergebnisse produziert, dann, weil dabei unumkehrbar Information verloren geht, die später vielleicht doch noch interessant sein könnte. Im Grunde wäre es also am besten, wenn das System sowohl die lemmatisierten wie auch die unlemmatisierten Daten zur Verfügung und verarbeitet hätte.

Ein *Daten-Satz* sei nichts als eine Folge von Tokens, die gegebenenfalls bereits durch Whitespace getrennt sind. Es wird dabei explizit nicht verlangt, dass linguistische Definitionen von Sätzen oder grammatische Regeln erfüllt sein müssen.

Für die Analyse können auch *Wortgruppen* interessant sein, dies sind Abfolgen von Tokens, typischerweise kleiner als Daten-Sätze, welche zusätzlich zu den enthaltenen Tokens jeweils noch ein weiteres Mal so betrachtet werden, als wären sie ein einzelnes Token. So könnte etwa aus dem Satz: *Es regnet in New York.* ein Daten-Satz mit den folgenden Tokens entstehen: SATZ-ANFANG – Es – regnet – in – New – York – New York – . – SATZ-ENDE

2.2.2 Verteilung

2.2.2.1 Zipfsches Gesetz

Die Verteilung von Types über ein Korpus kann durch ein Potenzgesetz angenähert werden; dessen einfachste Fassung wird nach dem Entdecker dieses und einiger anderer Umstände, George Kingsley Zipf, auch als Zipfsches Gesetz (vgl. Zipf (1935)) bezeichnet. Darin wird beschrieben, dass für in der Natur beobachtete Rangordnungen das Produkt aus dem Rang und der gemessenen absoluten Größe (etwa der Typefrequenz) konstant ist. Tests mit sehr großen Mengen von Daten, zum Beispiel auch im Zusammenhang mit Richter (2004), haben ergeben, dass dies eine gute Abschätzung für eine obere Grenze darstellt und vor allem für die mittelfrequenten Types eine besonders hohe Übereinstimmung herrscht.

Besonders für kleinere Korpora schätzt die Zipf-Formel die Frequenz der häufigsten Types falsch und die Frequenz und Anzahl² der seltenen Types zu hoch. Da ein Korpus einer leben-

² Gemeint ist: Frequenz bedeute die Anzahl Tokens, die für einen Type sprechen und Anzahl die Zahl unterschiedlicher Types.

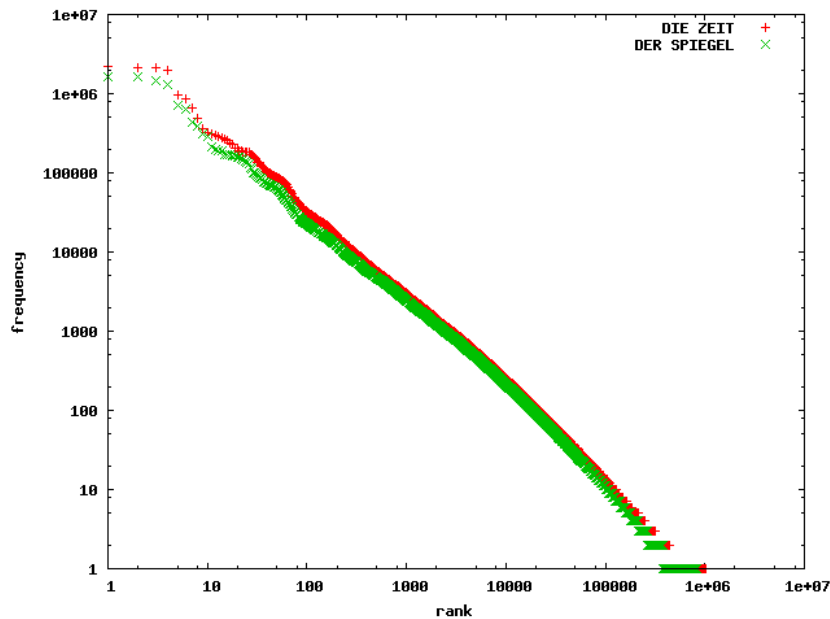


Abbildung 1: Zipf-Verteilung: 1,5 Mio. Sätze SPIEGEL und 2 Mio. Sätze DIE ZEIT

digen Sprache nur eine endliche Stichprobe darstellt und somit per Definition unvollständig ist, könnte dieser Umstand auch wegzu erklären versucht werden. Es gibt aber z. B. mit der Zipf-Mandelbrot-Formel (siehe auch Kap. 4.1 in Heyer u. a. (2006)) auch für den tatsächlich angetroffenen Sachverhalt treffendere, allerdings auch kompliziertere, Abschätzungen. Um bestimmte Grundgedanken zu motivieren genügt die einfache Version von Zipf aber völlig.

Abbildung 1 und Abbildung 2 zeigen zwei Beispiele für die Verteilung von Types in unterschiedlich großen Korpora (basierend auf 1,5 Millionen Sätzen aus DER SPIEGEL, und ca. 2 Millionen Sätzen aus DIE ZEIT) bzw. annähernd gleich großen Korpora unterschiedlicher Sprachen (je 100 000 Sätze Deutsch, Französisch und Türkisch). Die Erstellung solcher Korpora wird in Unterabschnitt 4.3.2 noch näher beschrieben werden. Die Abbildungen zeigen, dass, wie erwartet, das Ergebnis von einer Geraden etwas abweicht: Besonders die Frequenzen der häufigsten Types und die Anzahl der seltensten Types wird falsch abgeschätzt, umso mehr, je kleiner das betrachtete Korpus ist.

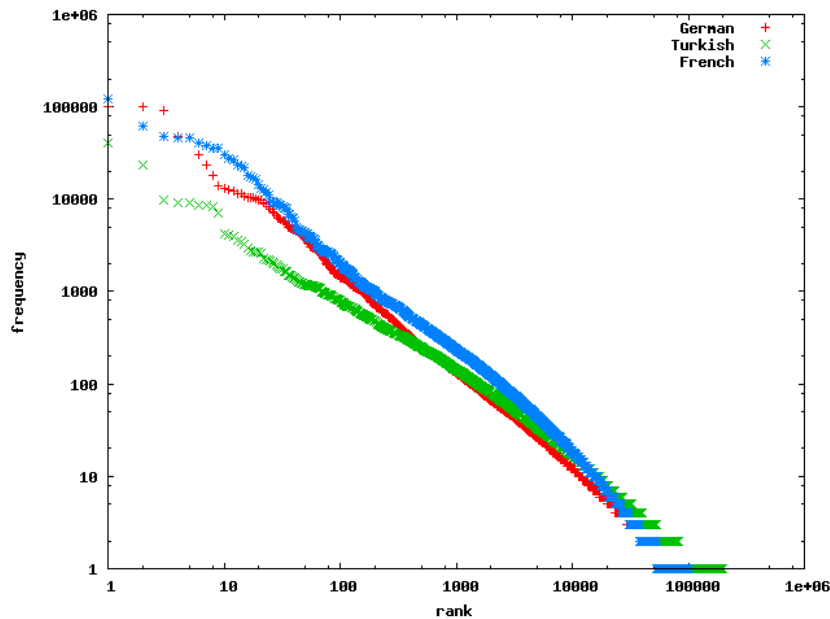


Abbildung 2: Zipf-Verteilung: Je 100 000 Sätze Deutsch, Französisch und Türkisch

2.2.2.2 Stoppwörter und Textabdeckung

In retrieval-orientierten Informationssystemen, vor allem in Suchmaschinen und Content Management Systemen ist es immer noch üblich bestimmte sehr häufige oder inhaltsleere Types von der Indexierung auszunehmen, z. B. um Platz zu sparen oder um Abfragen performanter beantworten zu können. Die Annahme, die dabei unterstellt wird ist, dass nach diesen so genannten Stoppwörtern ohnehin niemand sinnvolle Suchanfragen stellt. Diese Annahme trifft zwar oftmals zu, ist in ihrer Allgemeinheit aber natürlich falsch, was etwa bei der Suche nach einer der berühmtesten Stoppwort-Kombination der Literatur: *to be or not to be* leicht auffällt. Tatsächlich unterstützen zumindest die gängigen Internetsuchmaschinen, wie Google, Yahoo oder Live, inzwischen längst Abfragen nach Stoppwörtern.

Stoppwörter sind im Prinzip einzelsprachabhängig, wobei es natürlich zu Überlagerungen (*in* mit ungefähr gleicher Bedeutung z. B. im Deutschen und Englischen) und Zweideutigkeiten (*die* als Artikel im Deutschen und als möglicherweise doch interessantes Verb im Englischen) kommen kann. Grundsätzlich gibt es zwei Ansätze zum Erhalten von Stoppwortlisten, einmal das Nutzen von linguistischem Wissen über eine Sprache, z.B. durch das Heranziehen von Listen geschlossener Wortklassen wie Artikel, Pronomen, Präpositionen, Hilfsverben etc. und einmal das

rein frequenzbasierte Vorgehen. Während erstere Methode die Liste unnötig mit selten benutztem Sprachinventar aufzublähen neigt, eröffnet letzteres eine kollektionsspezifische Quelle von Problemen. Ist etwa für ein Literaturkorpus *Liebe* ein zentraler Begriff, um den sich vieles oder gar alles dreht, oder schlicht ein Stoppwort? Fällt eine gefühlsmäßige Entscheidung hier durch die eigene Sprachkompetenz auch leicht, so lässt sie sich für Material in einer unbekannten Sprachen nicht treffen – die rationale Begründung kann also nicht rein frequenzbasiert auftreten. Aber auch bei einer Mischform, also einer Auswahl der aufgrund von Anzahl und Wortart als irrelevant angenommenen Types, steht einer Ersparnis von einigen Prozent Speicherplatz ein Informationsverlust gegenüber, der je nach Anwendungsfall unproblematisch oder eben sehr störend sein kann; Tabelle 6 auf S. 52 aus Quasthoff u. a. (2006) zeigt für 15 Sprachen exemplarisch Daten zur Textabdeckung durch die Top n Types. Demnach genügen wenige Top-Wörter um etliche Prozent des Materials abzudecken, mehr aber auch nicht: Größenordnungsmäßig bleibt ein Gewinn aus.

2.2.2.3 Gleichmäßigkeit und Burstiness

Die Verteilungsplots der Zipfkurve zeigen, wie die Types über das gesamte Korpus verteilt sind – nicht jedoch, wie sie im Korpus verteilt sind. Grundsätzlich könnte die Verteilung variieren zwischen blockweise und homogen; ersteres widerspricht unmittelbar dem gewohnten Umgang mit Sprache: es gibt keine (sinnvollen längeren) Texte, die ihre Tokens nach Type geordnet enthalten. Dass auch letzteres nicht der Fall ist, bedarf etwas Motivation: Insofern Texte Themen haben und Themen zur Beschreibung einen spezifischen Wortschatz bevorzugt benutzen, kann die Verteilung von Types nicht zufällig und homogen sein. Wie De Roeck u. a. (2005) gezeigt haben, gilt die Annahme einer homogenen Verteilung noch nicht einmal für die häufigsten Types.

Kleinberg (2002) nutzt das plötzliche Ansteigen der Wiederhol frequenz bestimmter Types aus, um Themenwechsel in Textströmen zu erkennen. Die Grundannahme dabei ist auch in alltäglichen Gesprächen leicht nachzuvollziehen: Sobald man ein neues Thema anschneidet, werden eben eine eher geringe Zahl von spezifischen Begriffen plötzlich und nur, solange das Thema auf der Tagesordnung steht, häufiger zum Einsatz kommen.

2.2.3 Kookkurrenzen

Der Begriff Kookkurrenz bezeichne mit Heyer u. a. (2006) das statistisch auffällige gemeinsame Auftreten von Objekten in Bezug auf eine Vielzahl untersuchter Einheiten. Typische Objekte seien Types, typische Untersuchungseinheiten Sätze oder unmittelbare Wortnachbarn. In Abweichung von Heyer u. a. (2006) (Kapitel 4.7) wurden für diese Arbeit Kookkurrenzen nicht mit dem Poisson-Maß (Quasthoff und Wolff, 2002) sondern mit dem log-likelihood-Maß (Dunning, 1993) in der Implementierung von Bordag (2007) berechnet.

Werden Kookkurrenzen nach ihrer Signifikanz geordnet, ergibt sich wieder eine Verteilung, die innerhalb der Grenzen des Maßes mit einem Potenzgesetz beschrieben werden kann.

2.3 EXKURS: EIN VERTEILUNGSEXPERIMENT

Weil es im Rahmen dieser Arbeit mit den Wörtern des Tages später um zeitlich zerlegte Korpora gehen wird, sollte die Frage nicht ungestellt und -beantwortet bleiben, welchen Einfluß verschiedene Formen der Zerlegung auf die Messergebnisse haben. Dazu wurde eine Serie von Experimenten unternommen deren Rahmen im Anschluß erklärt sowie deren Resultate beschrieben werden.

2.3.1 Setup

Ausgangsbasis der Experimente ist ein 40 Millionen Sätze umfassendes Textkorpus des Deutschen. Die Sätze darin sind durchnummeriert, sprich in den Texten aufeinander folgende Sätze folgen auch im Korpus aufeinander. Außerdem sind die Sätze der Entstehung nach sortiert; sie stammen aus dem Zeitraum 1995–2005. Nun werden zufällig jeweils 40 Stichproben á 100 000 Sätze gezogen, wobei zwei Parameter variiert werden: Erstens die Anzahl Sätze S , die unmittelbar aufeinander folgend gezogen werden (1, 5, 25, 125) und zweitens der Bereich K aus dem die Sätze gezogen werden können (1 Mio., 4 Mio., 10 Mio., 40 Mio. Sätze). Damit werden unterschiedliche Auswahlmechanismen (einzelne Sätze, Absätze, Kapitel, Texte) sowie unterschiedliche Korpusgrößen simuliert.

S / K	1 Mio.	4 Mio.	10 Mio.	40 Mio.
1	2020883	2021002	2020852	2021959
5	2021460	2021786	2020905	2020908
25	2022077	2021242	2020676	2021914
125	2019810	2021365	2017396	2019008

Tabelle 1: Durchschnittliche Anzahl Tokens in Abhängigkeit von Korpus- und Samplegröße ☞ im Mittel kaum ein Unterschied

S / K	1 Mio.	4 Mio.	10 Mio.	40 Mio.
1	154921	162324	162576	166351
5	155180	157563	159686	163161
25	149873	153437	157507	158694
125	147634	151286	153014	156025

Tabelle 2: Durchschnittliche Anzahl Types in Abhängigkeit von Korpus- und Samplegröße ☞ mehr Zufall, mehr Types

S / K	1 Mio.	4 Mio.	10 Mio.	40 Mio.
1	438460	413683	405442	395641
5	446044	426394	417282	409559
25	461865	443872	435792	430336
125	472383	456571	448435	443922

Tabelle 3: Durchschnittliche Anzahl Satzkoookkurrenzen in Abhängigkeit von Korpus- und Samplegröße ☞ mehr Zufall, weniger Kookkurrenzen

S / K	1 Mio.	4 Mio.	10 Mio.	40 Mio.
1	169308	167249	167000	166408
5	164870	168984	168360	168027
25	171898	170945	170386	170447
125	172564	171944	171502	171460

Tabelle 4: Durchschnittliche Anzahl Nachbarschaftskookkurrenzen in Abhängigkeit von Korpus- und Samplegröße ☞ Kein so starker Einfluss auf Nachbarschaftskookkurrenzen

2.3.2 *Einfluss der Samplegröße*

In Tabelle 1 ist bei den Tokens keine signifikante Abweichung zu erkennen. Bei den Types in Tabelle 2 dagegen ist ein Trend zu erahnen, dass je größer das zusammenhängende Sample wird, es umso weniger Types gibt. Signifikant ist der Unterschied jeweils nur zwischen der obersten und untersten Spalte. Genau umgekehrt zu den Types ist das Ergebnis für beide Arten von Kookkurrenzen in Tabelle 3 und Tabelle 4: Je größer die einzelnen zusammenhängenden Samples, desto mehr Kookkurrenzen gibt es, wobei die Satzkookkurrenzen stärker betroffen sind. Signifikant sind wiederum nur die Unterschiede zwischen den Einträgen in der obersten und untersten Spalte. Die Ergebnisse stehen im Einklang mit der im vorigen Abschnitt erwähnten Beobachtung von Thematisierung, also der Steigerung der Wiederauftretenswahrscheinlichkeit thematisch relevanter (sonst eher seltener) Wörter nach deren ersten Auftreten in einem Text.

2.3.3 *Einfluss der Korpusgröße*

In Tabelle 1 ist bei den Tokens wiederum keine signifikante Abweichung zu erkennen. Für Types steigt in Tabelle 2 die Anzahl mit der Größe des Korpus, aus dem die Stichprobe gezogen wird, wobei signifikant nur der Unterschied zwischen äußerster linker und äußerster rechter Spalte ist. Der Einfluss auf die Satzkookkurrenzen ist auch bei diesem Parameter umgekehrt zum Einfluss auf die Types, sprich mit größerer Variation gibt es weniger Satzkookkurrenzen, gleichfalls nur mit signifikanten Werten in den Extremlagen der Tabelle 3. Auf die Nachbarschafts-Kookkurrenzen ist in Tabelle 4 kein signifikanter Einfluss feststellbar, wenngleich eine Ahnung besteht, dass bei wachsendem Parameterunterschied sich ein solcher analog zu den Satzkookkurrenzen noch einstellen würde. Auch diese Ergebnisse sprechen nicht gegen die Theorie der Thematisierung.

2.3.4 *Wiederauftauchen von Types und Kookkurrenzen*

Neben der schieren Größe ist ein weiterer Faktor für die geplante Form der Analyse noch entscheidender; um Konstanz und Wandel beschreiben zu können, bedarf es einer bekannten Erwartung an die Objekte, ob und wie verändert sie in mehreren Messungen beobachtet werden können. Grundsätzlich zeigt Abbildung 3, dass sich sowohl bei den Types als auch bei den Kookkurrenzen eine Verteilung ähnlich einem Potenzgesetz er-

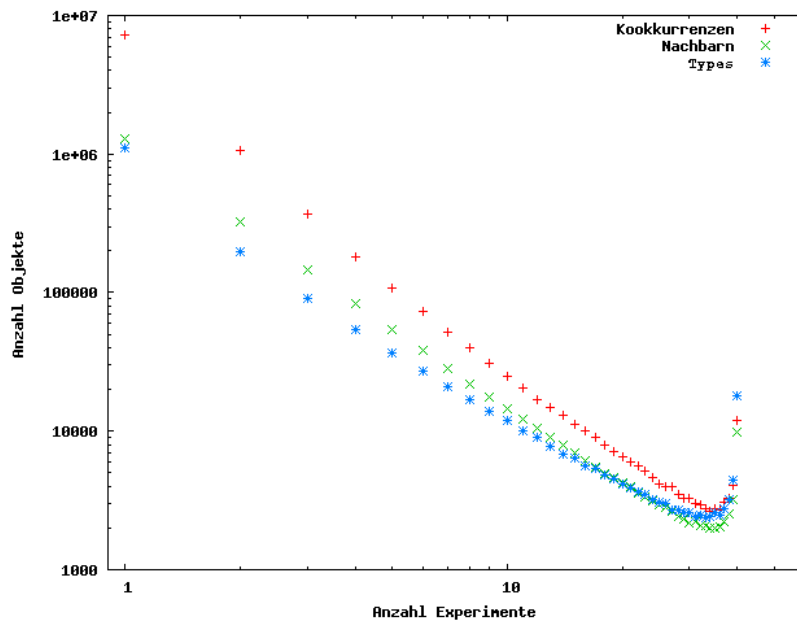
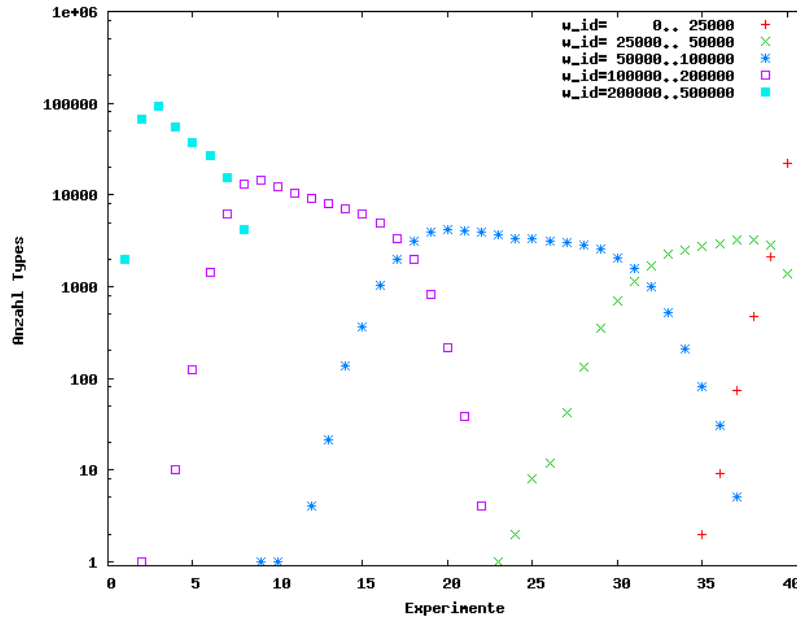


Abbildung 3: Wiederauftreten: Types, Nachbarn, Kookkurrenzen

Abbildung 4: Wiederauftreten der top n Types ($S=1$, $K=40$)

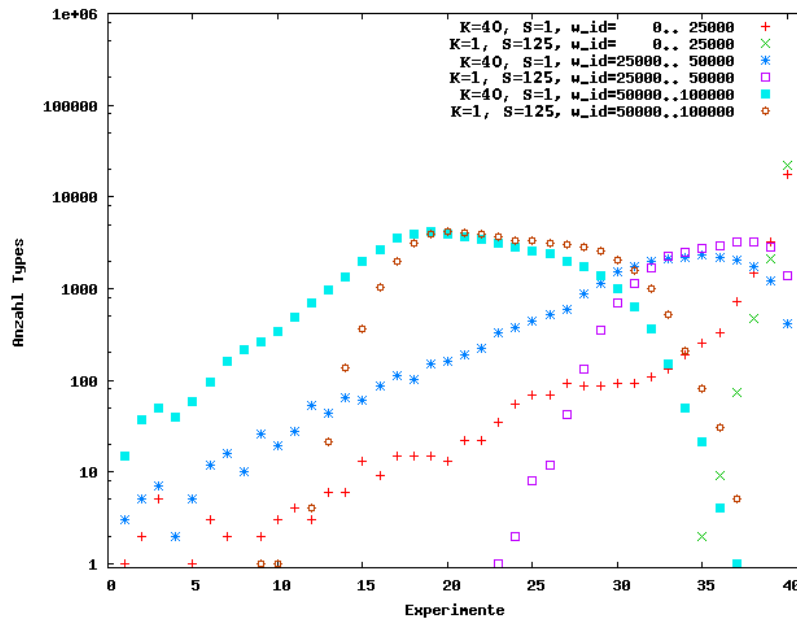


Abbildung 5: Wiederauftreten der top n Types ($S=125, K=1$) vs. ($S=1, K=40$)

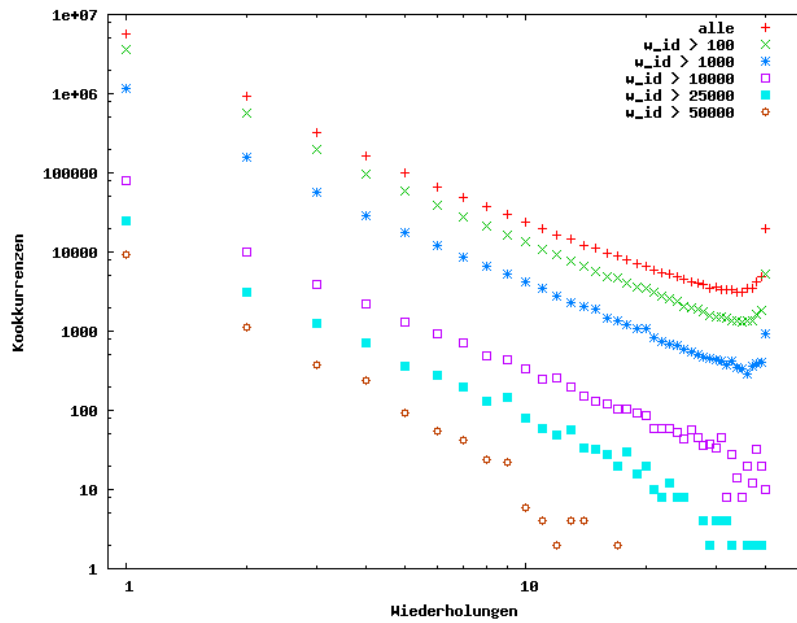


Abbildung 6: Wiederauftreten von Kookkurrenzen der Top n Types ($S=1, K=40$)

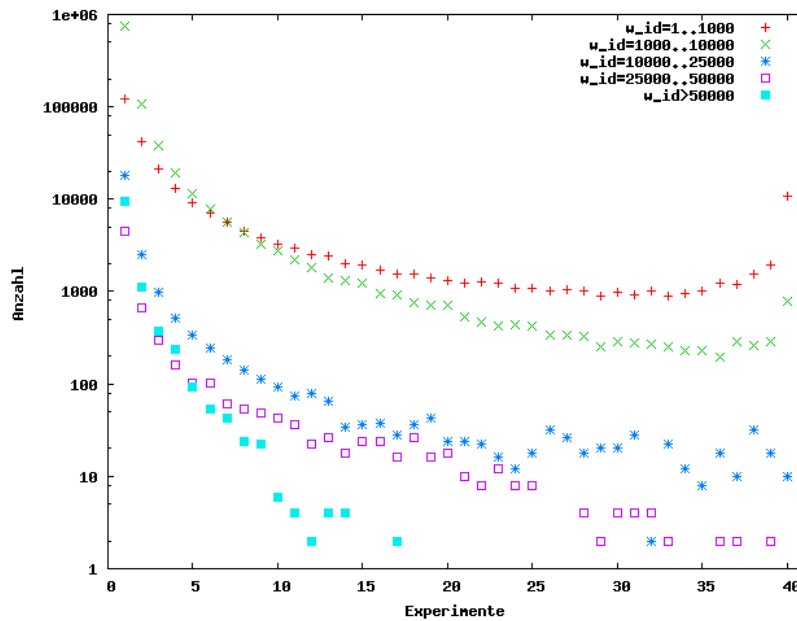


Abbildung 7: Wiederauftreten von Kookkurrenzen der Top n Types ($S=125$, $K=1$)

gibt, wobei sich am rechten Rand des Spektrums ein der Größe der Korpora im Experiment entsprechender Teil herauskristallisiert, der nicht diesem Gesetz folgt, sondern quasi immer vorhanden ist. Die Vermutung liegt nahe, dass dies der thematisch unspezifische Anteil des Untersuchungsgegenstands ist.

Natürlich hat dieser Teil eine gewisse Mindestfrequenz, um überhaupt in allen (oder zumindest sehr vielen) Experimenten gemessen werden zu können. Wie genau die Verteilung sich in der möglichst viel Zufall garantierenden Auswahl mit $K=40$ und $S=1$ dem absoluten Häufigkeitsranking nach aufteilt, zeigt Abbildung 4 und wie groß hierauf der verzerrende Einfluss zusammenhängender Auswahlen aus einem zeitlich limitierten Subkorpora mit $K=1$ und $S=125$ ist Abbildung 5.

Bei den Satzkookkurrenzen (und nur um die wird es im Folgenden gehen) stellt sich ein ähnliches Bild ein wie Abbildung 6 ($K=40$, $S=1$) und Abbildung 7 ($K=1$, $S=125$) zeigen. Der maximale Rang des Types mit dem höheren der beiden Ränge der Kookkurrenz, ist bei den Kookkurrenzen, welche tatsächlich in jedem Subkorpora enthalten sind, deutlich kleiner als bei den Types. Dies verwundert nicht weiter, da ja für eine Kookkurrenz eine höhere Mindestfrequenz als 1 erforderlich ist.

Zusammenfassend bleibt festzuhalten, dass das regelmäßige Beobachten von Wörtern auch bei kleineren Korpora als im Experiment kein Problem darstellen sollte, dass dies bei Kookkur-

renzen aber keineswegs der Fall ist und somit das Erneute auftreten von Kookkurrenzen in einem anderen Subkorpus das vergleichsweise stärker zu bewertende Zeichen ist.

2.4 ZEIT

2.4.1 *Definition*

Die Zeitforschung ist in zahlreichen Wissenschaften ein sehr komplexes Forschungsgebiet; einen Überblick darüber gibt zum Beispiel Lenz (2005). Im Rahmen dieser Arbeit wird nur eine sehr einfache Vorstellung von Zeit verwendet: Zeit sei eine Möglichkeit zur diskreten Anordnung von Texten in einer logischen Reihenfolge, die über das Datum gegeben ist. Der Verlauf der in den Texten behandelten Themen wird anhand von zusammen verwendeten Types nachgezeichnet.

2.4.2 *Betrachtungsarten*

Zwei Arten von zeitlichen Daten tauchen beim Umgang mit zeitlich segmentiertem Text auf. Einmal Zeitpunkte und -spannen, die sich anhand der Segmentierung ergeben, dann zeitangeben- de Ausdrücke (*temporal expressions*), welche im Text selbst vorkommen. Erstere können leicht durch einen Zeitstempel verwaltet werden, letztere sind wesentlich komplexer. Um die Vielfalt von Zeitangaben und -referenzierungen in Texten darstellen und auszeichnen zu können, wurde z.B. mit der TIMEML (Pustejovsky u. a., 2003) eine XML-basierte Auszeichnungssprache entworfen. Gegeben ein Zeitstempel für den Text lässt sich zu einer Referenz wie „am kommenden Freitag“ ein konkretes Datum ermitteln. Für die Analyse zeitlich segmentierter Frequenz- und Kookkurrenzdaten wird aber allein auf die Zeitdaten aus der Segmentierung, gewissermaßen als Messwerte, zurückgegriffen.

2.4.3 *Zeitreihenanalyse*

Solche Daten unterscheiden sich von den Daten klassischer Analysen von Zeitreihen (Wetter, Aktienkurse, ...) in der Erwartung an die Kontinuität der gemessenen Werte. Typische Zeitreihenanalyse (vgl. etwa Hamilton (1994)) zielt darauf ab, zyklische Komponenten und Trends zu extrahieren; die Erwartung an den Wert an der Stelle $n+1$ errechnet sich aus dem Wert an der Stelle n , zyklischen Anteilen, einem Trendanteil und Rauschen. Äußere Ereignisse haben dabei Einflüsse auf Trends, globale Verände-

rungen betreffen möglicherweise die zyklischen Anteile. Unbeeinflusst ist der Wert als konstant zu erwarten.

Bei den zeitlich segmentierten Frequenz- und Kookkurrenzdaten verhält es sich teilweise anders: Die Erwartung von zumindest mittelfristiger Konstanz wird gehalten, solange es sich um Wörter handelt, welche auf einen relativ stabilen Zustand verweisen (z.B. »Bundeshauptstadt – Berlin«). Auch insofern Wörter sich auf Ereignisse und Themen zurückführen lassen, was bei den Inhalt tragenden Wörtern vorausgesetzt werden kann, ist ihnen zwar die Tendenz von Diskursen sich zu etablieren und insofern ein gewisser Drang präsent zu bleiben zu eigen, dem entgegen wirkt aber stets ein Neuigkeitsdruck, also die Anforderung neuen Themen und Aspekten Platz zu machen und also – da ja, wie der Volksmund weiß, nichts so alt ist wie die Zeitung von gestern – im Lauf der Zeit sich zu verändern (z.B. »Europameister – Griechenland«, »Europameister – Spanien«) oder ganz zu verschwinden. Zeitreihenanalyse aus dem Lehrbuch ist also eher nicht geeignet, mit Wortfrequenzdaten sinnvoll umzugehen.

2.5 WAHRHEIT UND INFORMATION

Wir können nicht entscheiden, ob das, was wir Wahrheit nennen, wahrhaft Wahrheit ist, oder ob es uns nur so scheint.

— Heinrich von Kleist

Außerhalb entscheidbarer logischer Systeme ist Wahrheit nicht mehr objektiv feststellbar. In der Lebenswelt stellt dies freilich kein Problem dar: Solange Gründe, wie Indizien oder Vertrauen Annahmen, Positionen und Aussagen genügend gegen Widersprüche absichern, werden diese subjektiv wie auch intersubjektiv nicht nur als wahrscheinlich sondern oftmals auch als wahr angenommen. Dies trifft insbesondere dann zu, wenn es sich um "[...] Informationen, die im Modus von Nachrichten und Berichterstattung angeboten werden, [...]" (Luhmann, 2004, S.25) handelt. Die empirische Wissenschaft kann, Sir Karl R. Popper folgend (Popper, 1995), eine Theorie nur irgendwann einmal falsifizieren, niemals aber verifizieren – und auch in der Sicht des Konstruktivismus (Kepplinger, 1989) bleiben Erklärungen nur bis zu ihrer Falsifizierung viabel und haltbar.

Bei der Arbeit mit Daten, etwa in der Wissensrepräsentation oder eben beim Anfertigen von Inhaltsanalysen, gilt es dies gegebenenfalls zu berücksichtigen: „Auch als wahr identifizierte Aussagen haben demnach eine stets nur vorläufige Gültigkeit; sie können jederzeit durch neues Wissen erneut strittig werden.“ (Haller, 1993, S. 145) Daten können unvollständig, widersprüch-

lich oder schlicht falsch sein und die möglichen Gründe hierfür sind vielfältig, wie zum Beispiel Mangel an Wissen, Fehler und Irrtum oder Fälschung. Im Folgenden bleibt die Unterstellung und Beleuchtung gut- und böswilliger Absichten oder Motive ausgeblendet. Bei Texten über Vorgänge in der Welt, sind eine Reihe von Fehlerquellen involviert, begonnen dabei, dass eigentlich gar keine objektive Realität vorausgesetzt werden kann, und einschließlich all der Verknüpfungen, Schichtungen, Veränderungen und Selektionen (Luhmann, 2004), welche im Rahmen der Produktions-, Redaktions- und Distributionsprozesse erfolgen.

Während eine absolute Erkenntnis der Wahrheit von Sachverhalten auch dem Menschen nicht möglich ist, kann vielleicht die Maschine zumindest plausible Erklärungen erkennen und annehmen, sowie nicht plausible Erklärungen verwerfen. Die Kriterien nach denen solche Anpassungen vorgenommen werden, müssen dabei expliziert werden, um den Geboten der Transparenz und Nachvollziehbarkeit Genüge zu leisten. Dazu kann entweder auf die Daten selbst oder auf Metadaten verwiesen und bei diesen Datum, Quelle, Autor etc., bei jenen Form, Struktur sowie Inhalt bewertet werden.

The purpose of computing is insight not numbers

— Richard Hamming

Wie bereits eingangs erwähnt, gibt es eine Vielzahl an Theorien vom Zugang zu Text. Titscher u. a. (1998) besprechen z.B. explizit 15 Arten. Ob dies tatsächlich eine vollständige Auflistung ist, sei an dieser Stelle nicht weiter diskutiert, vielmehr sind die grundlegenden Unterschiede in Sachen Erkenntnisinteresse und Forschungsgegenstand der unterschiedlichen Herangehensweisen interessant. Für eine computergestützte Methode sind zweifelsohne solche Verfahren besser geeignet, welche sich unmittelbar auf den Text beziehen und ihre Regeln explizieren. Wenig geeignet scheinen dagegen Verfahren, welche der Anwendung komplexer Gedankengebäude in einem nicht zwingend regelgeleitetem Vorgehen näher stehen.

Die Beschränkung der betrachteten Textzugangsverfahren auf die häufigsten, das sind die im weiteren Sinne der Inhaltsanalyse verwandten, ist gängige Praxis. Auch im Folgenden soll dies so gehandhabt werden. Als scheinbarer Gegensatz wird dazu das Text Mining auf zwei Arten vorgestellt, einmal als Hilfsmittel bei der Inhaltsanalyse, dann als eigenständige Herangehensweise.

3.1 INHALTSANALYSE

3.1.1 *Geschichte*

Die Bildung des englischen Begriffs *content analysis* wird dem Forscherteam um Berelson und Lazarfeld am Office of Radio Research bzw. später im Bureau of Applied Social Research in den 1940er Jahren zugeschrieben, ebenso wie eine erste Systematisierung in Berelson (1952). Ihre Arbeiten gründeten freilich bereits auf einem reichen Fundus an Vorläufern und Verwandten, welche z.B. Krippendorff (2004) bis in das 16. Jahrhundert zurück verfolgt; für Merten (2004) beginnt eine erster Entwicklungsschritt sogar bereits mit der Geschichte der Menschheit. Früh (1998) ist primär die „...Entwicklung der Inhaltsanalyse zu einer wichtigen und eigenständigen wissenschaftlichen Methode ...“ (S. 11) statt einer historischen Darstellung wichtig.

Einen Anfang dafür setzt er kurz nach Beginn des 20. Jahrhunderts an, programmatisch in Deutschland etwa mit Max Webers Geschäftsbericht auf der Tagung der Deutschen Gesellschaft für Soziologie im Jahre 1910, und einem Höhepunkt um die eingangs bereits erwähnten Entwicklungen zur Zeit des Zweiten Weltkriegs. Dies sehen auch Rössler (2005) und Krippendorff (2004) vergleichbar. Diefenbach (2001); Züll und Mohler (1992) beschäftigen sich im besonderen mit den Vorläufern und der Geschichte der computerunterstützten Inhaltsanalyse, ebenso Züll und Alexa (2001), doch gehen letztere mehr auf aktuelle Entwicklungen ein. Einen *[h]istorische[n] Überblick über Trends der inhaltsanalytischen Forschung* bis ca. 1980 gibt einleitend Kuttner (1981) und unterscheidet dabei mit Krippendorff folgende wesentlichen Entwicklungen:

1. Vorläufer der Inhaltsanalyse
2. Quantitative Zeitungsanalyse
3. Massenkommunikationsforschung
4. Forschung für Krieg und Propaganda
5. Interdisziplinäre Erweiterung
6. Automatische Textverarbeitung

Aus heutiger und hiesiger Sicht, vgl. z.B. Früh (1998); Lissmann (2001); Mayring (2000), wäre explizit die Entwicklung einer Qualitativen Inhaltsanalyse hervorzuheben.

3.1.2 Vorgehen

Über den exakten Ablauf von Inhaltsanalyse findet sich viele Seiten in den im vorigen Unterabschnitt aufgeführten Büchern. Gemeinsam haben diese Ansätze mindestens folgendes:

- Forschungsfragen und -hypothesen werden entwickelt.
- Diese werden anschließend operationalisiert und in der Regel mit Hilfe von Wörter erklärenden Diktionären und einem Kategoriensystem in einem Kodierbuch explizit gemacht.
- Die zu analysierenden Medien, in der Regel Texte, werden mit Hilfe des Kodierbuchs von Menschen oder teilautomatisch in Mengen von Codes aus dem Kategoriensystem überführt.

- Die resultierenden Codes werden statistisch analysiert und auf dieser Grundlage Aussagen über die Hypothesen gemacht.

3.1.3 Kritik

Auch wenn die Inhaltsanalyse die am weitesten verbreitete Methode der Informationsgewinnung aus Texten ist, so ist sie bei weitem nicht perfekt. Klingemann (1984) kritisierte bereits Mitte der 80er Jahre die Inhaltsanalyse dafür, „... daß sie keinen gesicherten Bestand an Meßinstrumenten hervorgebracht hat, der in standardisierter und valider Weise soziale Realität beschreibt und damit konkurrierende theoretische Erklärungsansätze empirisch überprüfen konnte“. (S. 9) Immer neue, spezielle und ad hoc entwickelte Klassifikationssysteme führen zu einer disparaten Methodenlandschaft und erschweren damit Replikation, Verbesserung der Instrumente und eine Kumulation von Wissen. Züll und Alexa (2001) legen nahe, dass sich an dem Grundproblem nicht viel geändert hat. Butscher (2006) bringt das Unbehagen der Abhängigkeit von der individuellen Kompetenz der beteiligten Kodierer auf den Punkt: „Problematisch an der Codierung bleibt der durch die Sprachkompetenz des Forschers bedingte Einfluss auf den Originärtext: Eine Codierung von Textstellen setzt voraus, Aussagen der Autoren zu verstehen, diese gegebenenfalls zu interpretieren und auf ein Schema abzubilden. Verschiedene Ausdrucksformen oder sprachliche Differenzierungsmöglichkeiten von Autoren werden durch eine Codierung »geglättet« und fortan nicht weiter berücksichtigt. Dieses Problem betrifft alle inhaltsanalytischen Techniken: Es zeigt sich insbesondere bei zunächst wertfreien Begriffen, denen aber je nach Kontext entweder eine positive oder negative Bedeutung zukommt.“ (S. 42)

3.1.4 Mit Computer

Der Einsatz von Computern in der Inhaltsanalyse hat eine lange Tradition, die mindestens bis zum *General Inquirer* in den 1960er Jahren zurückreicht. Grundsätzlich läßt sich dabei eine Intensitätsskala aufspannen von

- der Verwendung des Rechners als simpler Datenspeicher und für statistische Berechnungen
- über Stichwörter und Kontext suchende Verfahren,
- sowie technologisch verstärkte Varianten derselben

- hin zu teil- und vollautomatischen Verfahren,

mithin also von der Replikation einer Inhaltsanalyse ohne Computer am Rechner, hin zu einem um alle technologischen Möglichkeiten aufgewerteten Verfahren. Bestimmte Möglichkeiten des Computereinsatzes können dabei freilich auch mit klassischen Herangehensweisen brechen. Dies ist notwendig, um neue Wege zu erproben und dann sinnvoll, wenn über Evaluationen nachvollziehbarer Mehrwert gewonnen werden kann.

Maurer und Reinemann (2006) erörtern die Freitextrecherche, den diktionsbasierten Ansatz und den Co-Occurrence-Ansatz. Methoden wie die semantische Netzwerkanalyse oder das Text Mining werden zwar wahrgenommen, jedoch angesichts eines bisher nur seltenen Einsatzes in der Praxis nicht weiter besprochen. Die Autoren sehen insbesondere in Deutschland noch erhebliches Wachstumspotential bei der computerunterstützten Inhaltsanalyse und gehen davon aus, dass durch CUI tatsächlich eine Schulung von Codern entfallen kann. Dass statt dessen erheblicher Aufwand für das Trainieren automatischer Verfahren anfällt, blenden die Autoren freilich aus.

3.1.5 Medienresonanzanalyse

Unter *Medienresonanzanalyse* versteht man in der Theorie so etwas wie „... ein computergestütztes, empirisches Instrument zur Beobachtung der veröffentlichten Meinung im Print-, Hörfunk und TV-Bereich ...“ (Femers und Klewes, 1997), auch wenn sich der Umfang in der Praxis oft nur auf Print- und inzwischen natürlich auch Onlinemedien erstreckt. Dabei werden im Verständnis des deutschen PR-Wirtschaftsverbands, der GPRA, Fragen beantwortet, die einer modifizierten LASSWELL-Formel entsprechen: *Wer sagt was wo in welcher Form mit welcher Meinungstendenz über ein(e) bestimmte(s) Unternehmen / Person / Thema?* Das Wesen der Resonanzanalyse ist es dabei, dass nicht bei der Analyse des Inhalts stehengeblieben wird, sondern ein stetiger Rückschluß auf die Realität stattfindet. Somit sollte Medienresonanzanalyse im Grunde ein permanenter Prozess an zentraler Stelle in Organisationen sein – und keine lästige Pflichtaufgabe, die Praktikanten einmal im Jahr mit ein bisschen in Excel und Powerpoint Herumklicken erledigt¹.

Was Medienresonanzanalyse in der Realität bedeutet, hängt stark von demjenigen ab, der behauptet sie durchzuführen. Klaus Merten hat in Merten und Wienand (2004) und Merten u. a. (2005)

¹ Was sie dabei wenigstens beachten müssen, steht in dem Praktikerbuch von Besson (2005).

als einer der wenigen umfangreiche Anforderungen an und Analysen von Medienresonanzanalysen erstellt. Diese hier zu replizieren scheint unnötig. Entscheidend ist, dass im Rahmen von Medienresonanzanalysen erhebliche Verdichtungen und Reduktion von Komplexität (von vielen Texten zu wenigen Kennzahlen) in derartigem Maß stattfinden, dass die Ergebnisse für den Laien inhaltlich auch nicht nur im geringsten nachvollziehbar sind. Die Durchführung von Medienresonanzanalysen erfordert somit maximale Wahrnehmung von Verantwortung und dazu gehört umfassende Nachvollziehbarkeit wenn schon nicht des Inhalts, so doch der Methoden.

Ziel des praktischen Teils dieser Arbeit ist es nicht, Medienresonanzanalyse zu betreiben, sondern die Durchführung von Medienresonanzanalysen zu erleichtern und in ihrem Instrumentarium zu erweitern, indem bestimmte, gut algorithmisierbare Zugangsarten zu Texten und Visualisierungsmöglichkeiten für Zwischenergebnisse beschrieben werden. Diese können freilich nicht nur der Spezialanwendung Resonanzanalyse zu Gute kommen, sondern allen eHumanities. Gerade das Einhergehen von Analyse mit korrekt eingesetzter Visualisierung kann entscheidend zur Verständlichkeit und dem Verstehen der relevanten Ergebnisse einer Analyse beitragen, oder, wie Edward R. Tufte es zu Beginn seines Buches *Visual Explanations* formuliert: „When principles of design replicate principles of thought, the act of arranging information becomes an act of insight.“ Entscheidend ist also jeweils Passungen zwischen Vorhandenem, Dargestelltem und Wahrgenommenem zu befördern. – Wie alle anderen Hilfsmittel lässt sich natürlich auch mit Hilfe *geeigneter* (aber faktisch mit zu unterstellender Betrugsabsicht verwendeten) Visualisierungen die schlechtere Sache zur besseren machen, siehe zum Beispiel im Medien DAX 30, dargestellt und kritisiert in Merten u. a. (2005).

3.1.6 Exkurs: Automatische Analyse von Meinungen und Einstellungen

Ein zentraler Punkt in der Medienresonanzanalyse ist die Identifikation von Meinungen und Meinungsträgern sowie die Klassifikation und Quantifizierung der Meinungstendenz. Dass dies in der Praxis keineswegs einfach ist, haben einige eigene Experimente außerhalb des Rahmens dieser Arbeit gezeigt, die unter anderem mit Hilfe einer manuell erstellten Liste positiver und negativer Wörter und den Daten aus Unter-Unterabschnitt 4.3.1.3 durchgeführt wurden ohne nennenswert gute Resultate zu erbringen. Für besser definierte Anwendungen, wie im Falle von

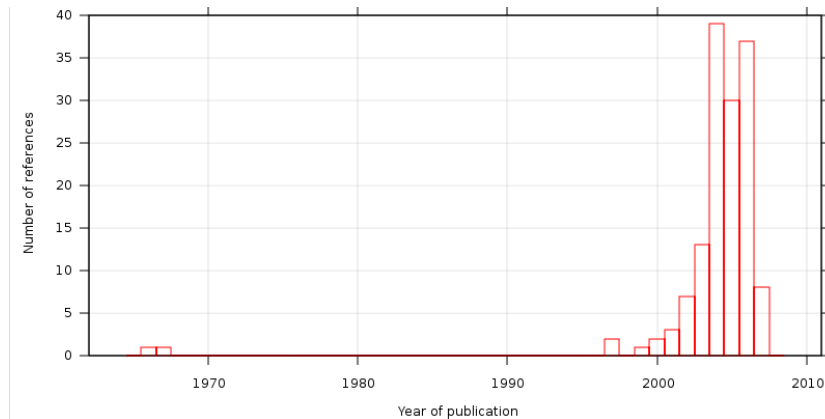


Abbildung 8: Anzahl der Veröffentlichungen pro Jahr in der sentiment-analysis-Bibliographie von Andrea Esuli.

Teichert u. a. (2008) bei der Einschätzung von Assoziationen von Konsumenten, sind solche Ansätze durchaus vielversprechend. Es bedarf aber, um erfolgreich *Opinion Mining* oder *Sentiment Analysis*² betreiben zu können, erheblicher Vorarbeiten und Investitionen in eine Operationalisierung von Meinungen, Einstellungen und Bewertungen, um auch für das Deutsche zu guten Ergebnissen zu kommen. Esuli und Sebastiani (2007) haben z. B. für das Englische mit SentiWordNet eine Resource geschaffen für die es im Deutschen noch keine Entsprechung gibt³. In der Dissertation von Esuli (2008) und online gibt es eine umfangreiche Bibliographie des Themas unter der URL <http://patty.isti.cnr.it/~esuli/research/sentiment/Sentiment.bib>. Die zeitliche Auswertung der Bibliographie (Stand Mitte 2007) in Abbildung 8 zeigt, dass Meinungsanalyse ein aktuell⁴ heiß diskutiertes Thema ist. Der Fokus der vorliegenden Arbeit liegt jedoch auf dem Faktor Zeit, nicht auf der automatischen Analyse von Meinungen.

3.1.7 Ein anderer Zugang zu Text durch Text Mining

Die Inhaltsanalyse zielt typischerweise zunächst auf konkrete Einzeltexte aus welchen nach einer Codiervorschrift Daten ermittelt und statistisch weiterverarbeitet werden. Diese Einzeltexte werden per repräsentativer Stichprobe aus einer wohldefinierten

² Der Deutsche Begriff Sentimentanalyse wird nur im Kontext von Stimmungen an der Börse benutzt

³ Zwar gibt es eine Reihe von Diktionären zu diversen bereits durchgeführten Inhaltsanalysen, diese sind aber nicht systematisch zusammengeführt worden oder gar frei zur Benutzung online verfügbar.

⁴ Das laufende Jahr war noch nicht vollständig erfasst.

Grundgesamtheit ausgewählt. Text Mining (vgl. z. B. Heyer u. a. (2006)) dagegen geht anders mit Texten um; es versteht Text als Rohstoff aus dem durch die Integration linguistischer und statistischer Verfahren direkt Wissen gewonnen werden kann. Damit die daran beteiligten statistischen Verfahren sinnvoll funktionieren, sollte die untersuchte Textmenge so groß wie möglich sein. Dass diese maximale Größe nicht unbedingt gleich der Grundgesamtheit allen Materials sein sollte, sondern nach offengelegten Kriterien balanciert sein sollte, wird in der Praxis (leider) bisweilen gerne vernachlässigt. Dass eine Mindestgröße nötig ist, führte nicht zuletzt bei Landmann und Züll (2004) in ihrem Praxistest zu dem Urteil, dass ein rein statistischer Ansatz zur Inhaltsanalyse mit Kookkurrenzen eher ergänzenden statt ein Diktionär ersetzenden Charakter hätte. Natürlich können Kookkurrenzen dazu verwendet werden, die Erstellung von Diktionären und Kategoriensystemen zu erleichtern, wie das im Bereich Ontologieaufbau bereits im Einsatz ist (vgl. zur Übersicht Biemann (2005)). Die linguistische Komponente eines Text Mining Verfahrens würde auch wesentliche Schwächen bisher verfügbarer Software wie die Problematik der Erstellung einer Stoppwortliste und Lemmatisierung per Suchen und Ersetzen adressieren können (so diese Schritte nötig sind, vgl. die Argumentation im vorangehenden Kapitel). Die wesentliche Stärke des Text Mining Ansatzes liegt aber der Meinung des Autors nach in der explorativen Analyse: in einer Perspektive, die nicht einzeltextspezifisch sondern korpusbasiert Information und Wissensbausteine den Analysierenden vor Augen führt und in die Hand gibt.

3.2 BEISPIELE

3.2.1 *Nachrichtensuchmaschinen*

Eine nicht mehr aktuelle oder vollständige, aber immerhin umfangreiche Übersicht zu Nachrichtensuchmaschinen liefert die Seite <http://searchenginewatch.com/2156261>⁵. Die Auftritte der ersten Generation von Nachrichtensuchmaschinen wie z.B. Paperball oder -boy sind Ende 2008 längst aus dem Netz verschwunden bzw. haben sich in ihrem Featureumfang erweitert. Zumindest Benutzerinteraktion, RSS-Alerts, Clusterings von Nachrichten etc. sind seit dem Markteintritt der großen Suchanbieter wie mit Google News oder Yahoo News nötig, um auch nur einen Funken Aussicht auf Erfolg zu haben.

⁵ Zugriff am 13. September 2007

The screenshot displays the News in Essence interface. On the left, under 'User Clusters', a list of news items is shown with their titles, article counts, and summary counts. On the right, a detailed view of a cluster titled 'FOX.com FOX Poll: Bush Approval Rating Hits New Low' is shown, including a 2% summary and a full text excerpt. Below this, a 'Cluster Documents' section lists six documents with checkboxes and 'Use As Seed' links. At the bottom, there are controls for 'Redraw', 'Reset', 'Compression' (set to 10%), and 'Summarize'.

User Clusters (Archive)

- **'Lynne Spears: You're still my brave little girl'**
2 articles, 3 summaries: 09/09, 8:20 AM
- **'Canada's best source for news continuously updated from The Globe and Mail'**
24 articles, 3 summaries: 09/07, 3:56 PM
- **'AsiaPacific 'Hundreds buried' by China quake'**
4 articles, 7 summaries: 09/12, 10:33 AM
- **'Canada's best source for news continuously updated from The Globe and Mail'**
10 articles, 3 summaries: 09/08, 5:21 PM
- **'Email Services guardian.co.uk'**
2 articles, 3 summaries: 04/26, 4:17 AM
- **'UK UK Politics Cameron pledge on EU treaty poll'**
8 articles, 7 summaries: 01/20, 5:34 PM
- **'Your views: Kenya smoking ban, Public smoking ban'**
5 articles, 3 summaries: 11/22/07, 2:04 AM
- **'Howard's power slips as fewer find a way to advance in Australia Special reports Guardian Unlimited'**
17 articles, 3 summaries: 11/21/07, 7:17 AM
- **'Nassau jail officer wins second award from county'**
10 articles, 3 summaries: 11/03/07, 4:05 AM
- **'Demand for White House's Abramoff Data'**
4 articles, 3 summaries: 11/02/07, 2:11 AM

Cluster Documents

1. CEO confidence declines in second quarter Jul. 11. 2007 (Cached) [money.cnn.com] (Use As Seed)
2. China's economy grows fastest since 1994 Jul. 11. 2007 (Cached) [money.cnn.com] (Use As Seed)
3. Xinhua English (Cached) [news.xinhuanet.com] (Use As Seed)
4. FOX.com FOX Poll: Third Party President Good for Country Polls Gallup Poll Opinion Polls (Cached) [www.foxnews.com] (Use As Seed)
5. FOX.com FOX Poll: Bush Approval Rating Hits New Low Polls Gallup Poll Opinion Polls (Cached) [www.foxnews.com] (Use As Seed)
6. FOX.com Mortgage Demand Climbs in Spite of Rising Interest Rates Business And Money Business Financial (Cached) [www.foxnews.com] (Use As Seed)

Redraw Reset Compression: 10% Summarize

Abbildung 9: News in Essence

3.2.2 Nachrichtenzusammenfassungen

3.2.2.1 News In Essence

Von 2003 – 2007 bereitete NewsInEssence (Radev u. a., 2005) tag-
gesaktuell Nachrichtencluster auf und erstellte Summaries. Seit-
dem steht nur noch das Archiv unter der Adresse <http://lada.ssi.umich.edu:8080/clair/nie1/nie.cgi> im Netz. Die Darstel-
lung (Abbildung 9) der Cluster ist eher spartanisch ausgefallen
und erinnert an ein Google News ohne Bilder. Die Zusammen-
fassungen werden durch die Extraktion einzelner besonders pas-
sender Sätze aus Dokumenten aus einem Cluster erzeugt und
sind in der Länge konfigurierbar.

3.2.2.2 California Newsblaster

Der California Newsblaster der Columbia University (McKeown
u. a., 2002) fasst seit 2002 täglich Artikel aus etwa zwei Dutzend
Quellen zusammen, ermittelt Cluster und erstellt Zusammenfas-
sungen dazu. Die Darstellung ist mehr benutzerorientiert als bei
News in Essence, dafür fehlt jegliche Konfigurationsmöglichkeit.

World

Mumbai attacks: U.S. calls for India-Pakistan cooperation in investigating Mumbai attacks
(World, 31 articles)



Indian police arrested two men accused of providing mobile phone cards to the gunmen in the Mumbai attacks, the first known arrests in the probe since the siege ended, police said Saturday. India suspects two senior leaders of a banned Pakistani militant group orchestrated the deadly Mumbai attacks, officials said Thursday, as Pakistan's president vowed strong action against any elements in his country involved in the siege. India picked up intelligence in recent months that terrorists were plotting attacks against Mumbai targets, an official said Tuesday, as the government demanded that Islamabad hand over suspected terrorists believed living in Pakistan. LONDON The attack on India's financial capital bears all the trademarks of al-Qaida simultaneous assaults meant to kill scores of Westerners in iconic buildings but clues so far point to homegrown Indian terrorists, global intelligence officials said Thursday. Pakistani airstrikes and a suspected suicide attack left 34 dead near the Afghan border on Wednesday, security forces said, as the U.S. urged broader action against militants after the Mumbai terror attacks. With corpses still being pulled from a once-besieged hotel, India's top security official resigned Sunday as the government struggled under growing accusations of security failures following terror attacks that killed 174 people.

Other stories about India, MUMBAI and Pakistan:

- **Bombay attacks: 'They were very young, like boys really'** (5 articles) [UPDATE]
- **New York's top cop warns of copycat Mumbai attacks** (10 articles) [UPDATE]

US spells out Iraq mission under new pact (World, 26 articles) [UPDATE]

The top U.S. commander in Iraq warned his troops Friday to expect subtle changes in combat operations - including obtaining warrants before searching homes and detaining people - when the newly approved U.S.-Iraq security agreement takes effect on Jan. 1. Iraq's presidency council Thursday approved the U.S.-Iraq security agreement the final step for the agreement to be ratified by the Iraqi government, a council spokesman said. Iraq's Cabinet on Sunday approved a security pact with the United States that will allow American forces to stay in Iraq for three years after their U.N. mandate expires at the end of the year, the government said.

Pentagon plans troop surge in Afghanistan (World, 10 articles)

The military is beginning a big building effort in Afghanistan to house the roughly 20,000 additional troops who are expected to begin pouring in early next year, a top military officer said Friday. Three Canadian soldiers have been killed by a bomb in south Afghanistan, bringing to 100 the number of Canadian troops to die in the conflict. Canada's top commander in Afghanistan, Brig Gen Denis Thompson, defended the country's mission, saying his troops were bringing "peace and stability".

Other stories about Afghanistan, Pakistan and troops:

- **ABC News: Car Bomb Kills 20 in Northwest Pakistan's Peshawar** (8 articles)

Abbildung 10: Newsblaster: Kategorie *world* am 6.12.2008

3.2.3 Nachrichtenüberblicke

3.2.3.1 EMM News Explorer

Das Joint Research Centre der Europäischen Union beobachtet seit 2002 mit dem European Media Monitor (EMM, siehe: <http://press.jrc.it/> und Best u. a. (2006)) nahezu in Echtzeit Nachrichten in insgesamt 11 Sprachen zu beliebigen Themen auf den Websites von Zeitungen und Fernsehsendern. Zu besonders auffälligen Themen werden Nachrichten zusammengefasst und in einer Übersicht auf einer Weltkarte dargestellt (Abbildung 11). Umfangreiche Faktenbasen im Hintergrund ermöglichen die Darstellung von Zusammenhänge zwischen Named Entities (Abbildung 12) auch über Sprachgrenzen hinweg unter Einbindung von thematisch relevanten Nachrichten, Bildern und Zitaten. Weiters bietet die Seite einen Newsbrief genannten Nachrichtenüberblick nach Top-Themenclustern (Abbildung 13), Angaben zur Verteilung von Themen, Artikeln und darin erwähnten Personen (Abbildung 14) sowie Newsalerts beim Vorkommen vorher festgelegter Stichwörter an.

3.2.3.2 Textmap

Textmap heißt eine Entity Suchmaschine, die aus dem Lydia-Projekt (Lloyd u. a., 2005) der New York State University entstanden ist. <http://www.textmap.com/> Die Website beinhaltet zahlreiche Visualisierungen zu Netzwerken zwischen Named Enti-

Clustered news for Friday, December 5, 2008

[Read more...](#)

Abbildung 11: EMM: Clusteransicht

Angela Merkel

Information about this person was last updated on Mittwoch, 3. Dezember 2008.



Abbildung 12: EMM: Schreibungen (inklusive Fehlschreibungen), Titel und Bild zur Named Entity „Angela Merkel“

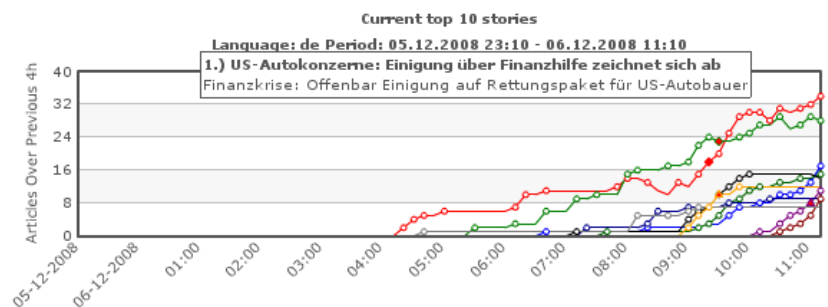


Abbildung 13: JRC Newsbrief: Top-10 Nachrichtenübersicht

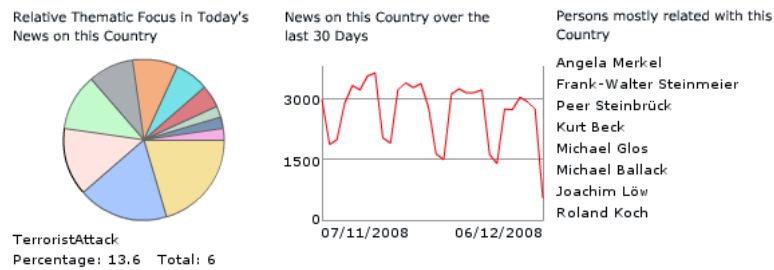


Abbildung 14: JRC Newsbrief: Themen, Artikel und Personen (Deutschland, 6.12.2008)

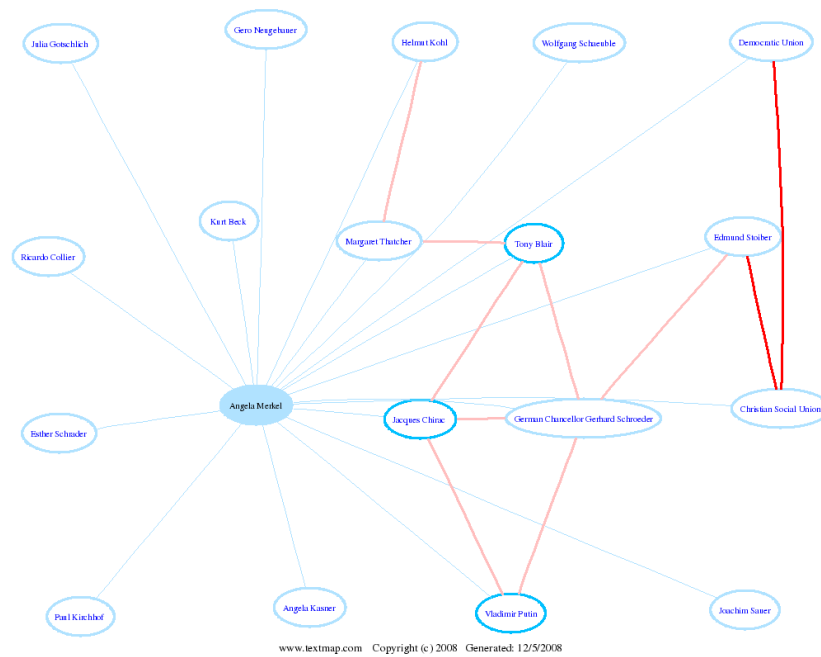


Abbildung 15: Textmap: Netzwerk für „Angela Merkel“

Popularity Time Series (What is this?)(More Time Series.)
Total 9769 references in 2414 articles

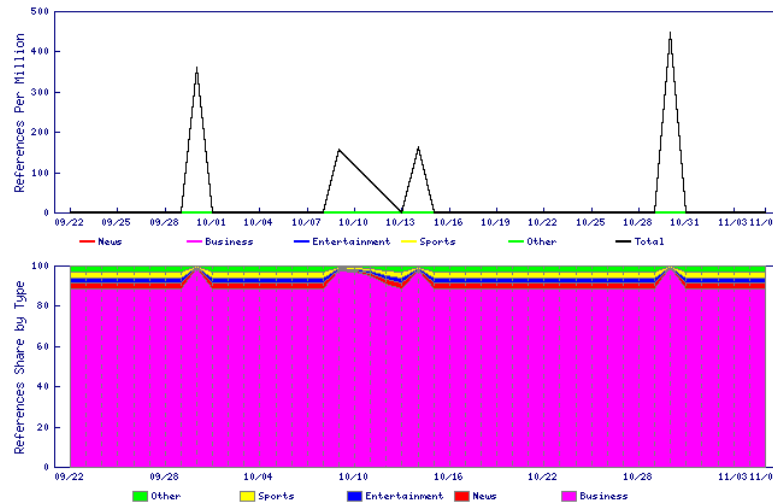


Abbildung 16: Textmap: „Angela Merkel“ – Verlauf der Verteilung der Berichte auf Oberkategorien

Juxtapositions for Angela Merkel: (What is this?)

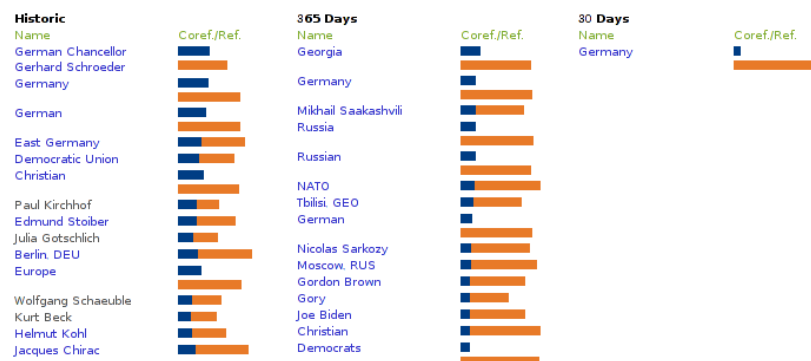
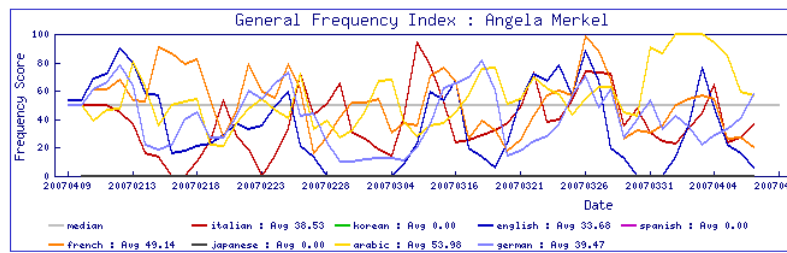


Abbildung 17: Textmap: Juxtapositions für „Angela Merkel“ in unterschiedlichen Zeitintervallen

Time Series for International News(Frequency) (What is this?)



Time Series for International News(Sentiment) (What is this?)

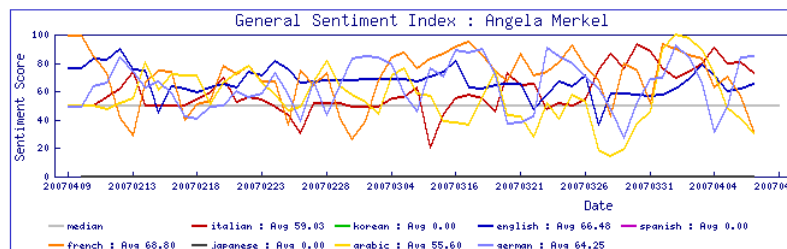


Abbildung 18: Textmap: Gegenüberstellung des zeitlichen Verlaufs von Frequenz und Sentiment

ties (Abbildung 15), Verteilung der Vorkommen von Names Entities über allgemeine Kategorien wie *business* oder *sports* (Abbildung 16) und Suchmöglichkeiten für in erster Linie englischsprachige Nachrichten sowie insbesondere Meinungsanalyse (Abbildung 18) im sehr großen Maßstab (Godbole u. a., 2007). Die so genannten Juxtapositions (Abbildung 17) welche auch in Abbildung 15 mit einfließen stellen statistisch auffällig Assoziationen dar – deren zeitliche und inhaltliche Darstellung wirkt auf den Benutzer allerdings eher verwirrend.

DIE WÖRTER DES TAGES

4.1 EINORDNUNG UND URSPRUNG

Die Wörter des Tages finden im Zusammenhang mit dem Projekt Deutscher Wortschatz in der Abteilung Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig statt. Zur Geschichte des Projekts Deutscher Wortschatz stehen mit Quasthoff (1998) und Quasthoff und Wolff (1999) frühe Quellen zur Verfügung. Seit Ende 2000 kann der Verfasser dieser Arbeit zudem auf eigene Erfahrungen und Erinnerungen aus der Mitarbeit am Projekt zurückgreifen. Zunächst soll über dieses ein kurzer Überblick verschafft werden, bevor die Idee und Entstehung der Wörter des Tages beleuchtet sowie verwandte Projekte und Intentionen beschrieben werden.

4.1.1 *Projekt Deutscher Wortschatz*

Das Projekt Deutscher Wortschatz verbindet die Erstellung umfangreicher Korpora und Methoden zur Be- und Verarbeitung dieser mit der Sammlung und Integration von Wörterbuchdaten.

4.1.1.1 *Motivation*

Anfang der 1990er Jahre waren zum deutschen Wortschatz keine frei verfügbaren größeren maschinenlesbaren Sammlungen vorhanden. In der Abteilung Automatische Sprachverarbeitung (ASV) des Instituts für Informatik der Universität Leipzig wurde daher begonnen, zunächst eine „... Wortliste mit sporadisch vorhandenen Angaben zu Grammatik und Sachgebiet ...“ (Quasthoff, 1998) zusammenzustellen.

Neben der Kompilation vorhandenen Materials zu sprachlichen Belangen wurden zudem Vollformen von Wörtern aus Texten gesammelt, um eine Datengrundlage für das maschinelle Erlernen von Flexionsschemata zu erhalten. Zum standardisierten Zugriff auf und zur Verwaltung der wachsenden Datenmengen wird seit 1994 ein relationales Datenbankmanagementsystem benutzt, zum Zeitpunkt der vorliegenden Arbeit das Open Source Produkt MySQL.

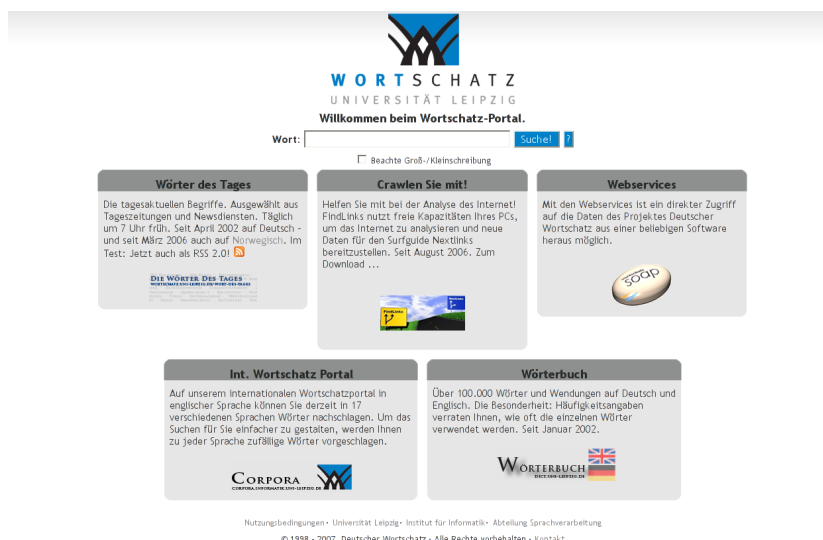


Abbildung 19: Das Wortschatz-Portal (Stand: 13. Dezember 2007)

4.1.1.2 Umsetzung und Entwicklung

Da stets zahlreiche Fragen, etwa zur Orthographie, Herkunft, Bedeutung usw. neu gefundener Formen zu klären sind, wurde ab 1996 automatisiert jeweils ein Beispielsatz mitgesammelt. Die neu vorhandenen Daten ermöglichten es, gemeinsame Vorkommen von Wörtern zu untersuchen, seien es im konkreten Fall idiomatische Fügungen, Kollokationen im Sinne der Linguistik, oder statistisch auffällige Kookkurenzen.

Da die Datengrundlage aufgrund des Auswahlverfahrens extrem unbalanciert war, wurden (und werden immer noch) seit 1998 neue Sätze gesammelt; ab etwa 2000 geschah dies auch tagesaktuell. Zur automatisierten Aufbereitung und Auswertung der Quellen, sowie zum Datenbankaufbau wurde mit dem ConceptComposer eine erste Version einer weitestgehend einzelsprachunabhängig arbeitenden Software entwickelt (Biemann u. a., 2004). Grundvoraussetzung für den Einsatz auf einer Sammlung von Daten ist nur, dass in dieser nach Gruppen (z. B. Wörtern) und Blöcken (z. B. Sätzen) unterscheidbare Zeichen auftreten.

Die Software wurde sowohl zur Analyse von Text in anderen Sprachen, wie Englisch, Französisch, Koreanisch, usw. als auch nicht-natürlichsprachiger Daten, etwa gleichmäßig gruppierten Genomsequenzen oder Links auf Seiten im Internet, erfolgreich angewandt. Inzwischen steht mit Medusa (Büchler, 2006) eine weiterentwickelte Neuimplementierung der Software unter einer Open Source Lizenz zur Verfügung.

In einer damit erzeugten Textdatenbank kann neben dem im

Volltext indexierten und statistisch ausgewerteten Korpus (genannt: *Wortschatz-Korpus*) auch weitere Information aus Quellen zu den Einträgen und Sätzen vermerkt sein (genannt: *Wortschatz-Lexikon*). Statistische Daten sind insbesondere die Frequenz der extrahierten Vollformen bzw. Wortgruppen, die statistische Signifikanz auftretender Paare von Einträgen sowohl in direkter Nachbarschaft als auch im Satz und darauf aufbauend berechnete ähnliche Wörter. An sonstigen Daten sind in erster Linie Angaben zu Quelle und Einordnung der Beispielsätze (z. B. nach Sachgebieten) und sprachliche Angaben zu Einträgen vorgesehen; dies ist jedoch keine grundsätzliche Beschränkung.

4.1.1.3 Präsentation und Nutzung

Neben der zu internen Zwecken möglichen direkten Abfrage der Datenbank wurde der Netzöffentlichkeit im März 1998 unter der Adresse <http://wortschatz.uni-leipzig.de> ein zunächst ca. 100 Millionen laufende Wörter umfassendes Korpus aus annähernd 7,5 Millionen Beispielsätzen verbunden mit einer Vielzahl von Lexikonangaben zugänglich gemacht und seitdem wiederholt aktualisiert. Die aktuelle Fassung zeigt Abbildung 19. Seit Oktober 2004 ist der maschinelle Zugriff auch über Webservices nach dem vom World Wide Web Consortium (W3C) vorgeschlagenen Protokoll SOAP möglich. Im Mai 2006 wurden dann auf <http://corpora.uni-leipzig.de> Normgrößenkorpora in 15 unterschiedlichen Sprachen (Quasthoff u. a., 2006) zusammen mit einer Corpus Browser Software veröffentlicht, welche auch die Offline-Benutzung abdeckt.

Soweit die Zahlen noch zu ermitteln waren, zeigt Tabelle 5 die Entwicklung bei den Seitenabrufen und unterschiedlichen Besuchern der Website pro Monat: Das überproportionale Wachstum der Seitenaufrufe von 2000 auf 2001 rührt von einer Layoutänderung her, infolge der mehr einzelne Dateien übertragen wurden. Eine weitere Layoutänderung Mitte 2006 führt zu einem ähnlichen Effekt in umgekehrter Richtung. Für September 2005 sind wegen eines Hardwareschadens keine Zahlen mehr ermittelbar. Die Unterscheidung der Benutzer findet anhand der IP-Adressen statt. Der starke Abfall im Verlauf des Jahres 2007 ist direkt und indirekt Folge des restriktiven Ausschließens diverser Suchmaschinenspider im Zusammenhang mit den in Abschnitt 4.2.2.3 geschilderten Problemen.

Die Zahl der Anfragen an die Datenbank erhöhte sich von ca. 500 pro Tag Anfang 1999 über ca. 32 500 im Oktober 2004 auf ca. 19 500 im Juni 2007; das prozentuale monatliche Wachstum verlangsamte sich dabei im Lauf der Jahre von ca. 20 % Anfang

Monat	Seitenabrufe	Besucher
1999-10	93 789	n.a.
2000-03	107 099	11 234
2000-09	217 689	22 419
2001-09	1 254 783	63 069
2002-09	2 507 529	117 097
2003-09	4 704 003	171 229
2004-09	7 189 967	236 574
2006-02	8 549 574	566 479
2006-09	9 591 760	637 448
2007-03	12 512 426	860 906
2007-09	2 264 928	376 549

Tabelle 5: Seitenabrufe und Besucherzahlen im Lauf der Zeit

1999 auf durchschnittlich ca. 3 % während des Jahres 2004.

4.1.2 Idee zu „Wörtern des Tages“

Die Idee einer tagesgenauen Datensammlung existierte im Projekt mindestens schon seit 1999. Im Zuge der Ereignisse des 11. September 2001 entstand die Idee einer tagesaktuellen Auswertung der gesammelten Quellen zur Erzeugung einer abstrakten Art von Pressespiegel. Im Frühjahr 2002 wurde mit der Entwicklung eines Vorläufers der heutigen Wörter des Tages begonnen und die Berechnung schließlich rückwirkend bis zum 1. Januar 2001 ausgedehnt. Für die Allgemeinheit unter <http://wortschatz.uni-leipzig.de/wort-des-tages/> online einsehbar sind wegen der in Unter-Unterabschnitt 4.2.2.3 beschriebenen Umstände wiederum nur jeweils die Daten für das aktuelle Jahr.

Die Wörter des Tages wurden zunächst für Deutsch (Quasthoff u. a., 2002b, 2003; Richter, 2005), dann für Norwegisch (Eiken u. a., 2006) sowie Texte aus der Domäne *Sparkasse, Bank und Finanzaufsicht* und schließlich für deutschsprachige Weblogs aufgesetzt. Die technologisch umfangreichste, modernste und am besten dokumentierte Fassung stellen die norwegischen Wörter des Tages dar, weshalb diese auch des öfteren als Referenz dienen.

4.1.3 *Verwandte Ansätze und Arbeiten*

Die Wörter des Tages stehen im Bereich Medienbeobachtung bei weitem nicht allein. Es gibt Vorläufer und konkurrierende Systeme im wissenschaftlichen wie auch im kommerziellen Einsatz.

4.1.3.1 *Vorläufer in der tagesaktuellen Analyse*

Die automatische Beobachtung von Presse und Inhaltsanalyse mit dem Computer hat eine Vorgeschichte seit den 1960ern (siehe auch Unterabschnitt 3.1.4). Die *Pionierleistung* auf dem Gebiet tagesaktueller Auswertung haben DeWeese III und McCombs (1977) vollbracht. Die Autoren untersuchten mit Hilfe des *General Inquirer* und des *Harvard Dictionary* die Tageszeitung *Detroit News* inhaltsanalytisch auf Themen, wobei sie auf den Datenstrom an die Setzmaschinen zugreifen konnten. Schönbach (1984) schreibt dazu: „Gerade die Aufgabe, Themen aufzuspüren, ist aber mit relativ schlichten Kategoriedefinitionen zu lösen: In der Regel genügt eine Liste einfacher Begriffe, die ein Computer im Text suchen muß, um festzustellen, daß ein bestimmtes Thema in ihm behandelt wurde.“ (S. 138) und stellt drei wesentliche Aspekte heraus, bei deren Untersuchung bereits solche einfachen täglichen Analysen helfen können:

- Langfristige Beobachtung von Themen
- Veränderung von Themenprioritäten
- Entstehung neuer Themen

Dies werden später zentrale Elemente der Topic Detection bzw. Topic Tracking Tasks von TREC (siehe Wayne (2000) und Fiscus und Doddington (2002)). Weiterhin haben Kontostathis u. a. (2003) eine umfassende Übersicht zu Geschichte und Stand der Extraktion von Trends mit Hilfe von Text Mining zusammengestellt, die hier nicht wiederholt werden soll.

4.1.3.2 *Der Zugriff auf Nachrichten*

Seitdem Nachrichten im Internet zu finden sind, wurden sie nicht nur über die Homepages der ursprünglichen Anbieter gelesen, sondern auch von Dritten genutzt. Etwas abstrahiert zeigen sich in der Evolution dieser Weiternutzung folgende mit Beispielen¹ angereicherten unterscheidbaren Qualitäten:

¹ Die hier genannten Beispiele dienen explizit nicht einer Marktübersicht.

- **Indexierung:** Suchmaschinen indexierten Nachrichten und es bildeten sich spezielle Nachrichtensuchmaschinen wie *Paperboy* (inzwischen bedeutungslos) heraus. Eine Übersicht zu Nachrichtensuchmaschinen liefert <http://searchenginewatch.com/showPage.html?page=2156261>².
- **Arrangement:** Nachrichteninhalte wurden als Ware begriffen und auf *Portalen* als Komplettpakete oder an kundeninteressen angepasste Auswahlen und Zusammenstellungen syndiziert. Ranking wird dabei neuerdings oftmals durch Benutzerinteraktion mit Web 2.0 Angeboten wie <http://digg.com> oder <http://del.icio.us> implementiert.
- **Stellungnahme und Neuverknüpfung:** Eine Funktion der ersten *Blogger* war es, dass sie Nachrichten aus unterschiedlichen Quellen in einen neuen Zusammenhang stellten und Stellung dazu bezogen.
- **Quantitative Analyse:** Gerade im Hinblick auf Weblogs, erlebt die rein quantitative Analyse eine Renaissance. Angebote wie <http://technorati.com> oder Nielsen Buzzmetrics (<http://www.nielsenbuzzmetrics.com/>) sind hier stark am Markt vertreten.
- **Aggregation und Clustering:** *Nachrichtenübersichtsseiten* wie *Google News* und Medienbeobachtungsdienste wie *Ausschnitt* oder *Landau Media* bilden Cluster bzw. Sammlungen zu Stichwörtern von Nachrichten aus unterschiedlichen Quellen.
- **Summarization, Anreicherung und Analyse:** Durch Einsatz zusätzlicher Text- und Visualisierungstechnologie werden zum Beispiel beim *California Newsblaster* (McKeown u. a., 2002) oder bei NewsInEssence (Radev u. a., 2005) Nachrichten und Cluster textuell zusammengefasst und mit etlichen zusätzlichen Informationen und Analysen versehen weitergereicht, wie dies zum Beispiel beim *European Media Monitor* (Best u. a., 2006) geschieht. Auch die Nachrichtenagentur Reuters (über die aufgekaufte ClearForest) und Dow Jones' Factiva sind in diesem Bereich engagiert.

4.2 ARCHIVIERUNG

Bevor die Umsetzung der Wörter des Tages in Abschnitt 4.3 näher beschrieben wird, sollen hier noch einige grundsätzliche Fra-

² Zugriff am 13. September 2007

gen zur kulturellen Funktion und juristischen Lage von Archiven geklärt werden.

4.2.1 *Zur Funktion von Archiven*

Im Zusammenhang mit Wörterbüchern und Enzyklopädien wurde bereits auf die Funktion von Archivmedien eingegangen. Dort wurde zusammenfassend festgestellt: „Die Grundlage der Fähigkeit des Menschen zu Kultur ist die Fähigkeit nicht nur zu vergessen, sondern sich im Unterschied dazu zu erinnern und diesem Umstand Besonderheit zuzumessen.

[...]

Erinnern geschieht, betrachtet man es in heutiger Denkweise auf Systemebene, biologisch, etwa in Form veränderter Reaktion auf wiederholte Reize, psychisch in Form von Wissen, dem Gedächtnis und als Konstituente von Zeit, ebenso wie sozial als ermöglichende Konstituente von Kommunikation und zugleich auch als deren Inhalt.“ Richter (2004)

Archive sind, wie auch Museen, Bibliotheken, Denkmäler etc., eine Form des organisierten und kollektiven Erinnerns. Sie treffen Auswahlen zu erfassender Medien, erschließen und erhalten diese nicht nur, sondern werten sie auch aus und machen sie ihren Öffentlichkeiten zugänglich. Ihre Funktion als Stütze des gesellschaftlichen Gedächtnisses kann insbesondere in der heutigen Wissens- und Informationsgesellschaft nicht vernachlässigt werden, ebensowenig wie ihre Macht und der Gefahr des Mißbrauchs dieser Macht. Die wesentlichen Pole zwischen denen sich Archive und Gesellschaft hier bewegen sind Archivpflicht – wie z.B. bei der Deutschen Nationalbibliothek – einerseits und Datenschutz andererseits.

4.2.2 *Rechtliche Rahmenbedingungen*

Eine Warnung vorweg: Der Autor dieser Arbeit ist kein Jurist. Die im Folgenden gegebene Darstellung stellt also keine Rechtsberatung dar und kann auch nicht alle Eventualitäten vorwegnehmen. Vielmehr soll laienhaft versucht werden, einen Überblick darüber zu geben, welche rechtlichen Probleme beim Betreiben einer archivartigen Sammlung von digitalen Belegstellen aus öffentlich zugänglichen Quellen bis dato zu bewältigen waren und welche Konsequenzen sich daraus ergaben.

4.2.2.1 *Archivprivileg und elektronische Archive*

Archive haben einerseits Privilegien, etwa in Sachen Urheberrecht, andererseits aber als Institution für Informationsauswahl auch Prüfpflichten, etwa zum Schutz personenbezogener Daten (dazu weiter unten mehr). Die Privilegien elektronischer Archive gehen laut BGH-Urteil vom 10. Dezember 1998 (I ZR 100/96 (OLG Duesseldorf)) nicht so weit, dass (auch nur zum Hausgebrauch bestimmte) digitale Pressearchive ohne explizite Lizenzierung der Inhalte erlaubt wären, da eine Digitalisierung die Möglichkeiten der Weiternutzung nahezu grenzenlos erweitert und somit ein unlauterer Wettbewerbsvorteil dann entsteht, wenn der Rechteinhaber ebenfalls eine digitale Form seines Inhalts vertreibt. In diesem Zusammenhang ist auch das Urteil des Oberlandesgerichts München 1 vom 10. Mai 2007 zum Dokumentenlieferdienst Subito einzuordnen.

Die im Rahmen dieser Arbeit beschriebene und im Projekt Deutscher Wortschatz benutzte Form der Archivierung muss der Rechtslage Rechnung tragen. Dies geschieht mittels der im folgenden Abschnitt beschriebenen Lösung der Anforderungen einer weiteren Norm, dem Urheberrecht.

4.2.2.2 *Belegstellen und Urheberrecht*

Das deutsche Urheberrecht schützt die Rechte der Urheber von Werken der Literatur, Kunst und Wissenschaft. Ein Werk ist dabei eine persönliche geistige Schöpfung, welche eine bestimmte minimale Schöpfungshöhe erreichen muss. Für Texte wird zwischen Sprachwerken der Literatur und Schriftwerken, welche einem praktischen Gebrauchszweck dienen, unterschieden. Zeitungsartikel sind in der Regel letzteres und der Anspruch an die eigene geistige Schöpfung liegt dort in Abgrenzung von der rein schreibhandwerklichen Leistung höher.

Die via Verwertungsgesellschaften vergütungspflichtige Nutzung von Artikeln zu politischen, wirtschaftlichen oder religiösen Tagesfragen aus der Tagespresse ohne explizite Genehmigung wird in §49 Urheberrechtsgesetz (UrhG) geregelt. Digitale Pressespiegel fallen unter diese sehr restriktiv formulierte Regelung. Für die angestrebte Form einer Messung von Wortauftreten über die Zeit hinweg ist diese Form der Nutzung freilich völlig unbrauchbar, da jegliche textuelle Verwertung oder Indexierung explizit verboten ist.

Es wird also ein anderer Modus benötigt: Während vollständige Zeitungsartikel in der Regel die notwendige Schöpfungshöhe aufweisen, muss dies bei Passagen nicht mehr zwingend der Fall

sein. Das Urteil des Landgerichts München I vom 15. November 2006 zur „Übernahme eines Zeitungsartikels im Internet als Urheberrechtsverletzung“ (AZ 21 O 22557/05) illustriert diese Rechtsauffassung exemplarisch. Daran anschließend dürfte eine Verkürzung der Länge der Passage auf einen Satz effektiv sicherstellen, dass man mit an Sicherheit grenzender Wahrscheinlichkeit davon ausgehen kann, es bei der Passage nicht mehr mit einem Werk zu tun zu haben. Die Passage kann demnach ohne urheberrechtliche Probleme verarbeitet, gespeichert und analog zu den Ergebnissen von Suchmaschinen auch angezeigt werden, sofern der ursprüngliche Werkszusammenhang nicht mehr sinnvoll hergestellt werden kann. Eine zufällige Durchmischung sehr vieler solcher Sätze dürfte dieses Kriterium effektiv und effizient erfüllen.

Dennoch kann die Verbindung zum Artikelkontext noch via Hyperlink hergestellt werden, solange die ursprüngliche Quelle weiterhin öffentlich verfügbar ist. Die Entscheidung des Bundesgerichtshofs vom 17. Juli 2003 im Fall Paperboy (AZ I ZR 259/00) zu so genannten Deep Links, also Links direkt auf konkrete Inhalte z.B. auch aus einer Datenbank, statt auf eine Startseite eines Internetangebots bestätigt diese Vorgehensweise mit Hinweis auf §87b UrhG und §1 des Gesetzes gegen den unlauteren Wettbewerbs.

4.2.2.3 *Das allgemeine Persönlichkeitsrecht in Deutschland*

Das allgemeine Persönlichkeitsrecht in Deutschland ist kein eigen in speziellen Gesetzen fixiertes Recht. Vielmehr wurde es durch die Rechtssprechung, vor allem die des Bundesverfassungsgerichts, als Ausdifferenzierung der ersten Sätze der ersten beiden Artikel des Grundgesetzes entwickelt. Es schützt besonders den persönlichen Lebensbereich und verleiht der Selbstbestimmung im Zusammenhang mit Beschaffung, Verwaltung und Veröffentlichung von Daten großes Gewicht.

Für Presse und Rundfunk ist der Umgang mit dem allgemeinen Persönlichkeitsrecht an der Tagesordnung. Insbesondere gibt es in diesem Zusammenhang ein Recht auf zeitnahe und adäquate Gegendarstellung (BVerfGE 63, 131, 142 f.) und eine aus der Berichterstattung über Prominente (BVerfGE 101, Caroline von Monaco) bekannte Praxis per einstweiliger Verfügung bestimmte Äußerungen, ob wahr oder unwahr, zu unterbinden.

In einem Archiv mögen Bericht und Gegendarstellung nacheinander auftauchen, gelöscht wird in der Regel nichts. Solange das Archiv quasi unzugänglich genug auf Papier vorliegt, stört sich daran bislang auch niemand. Gefährlich wird das allgemei-

ne Persönlichkeitsrecht für ein Online-Archiv dann, wenn über das Archiv Sätze zugänglich sind, welche zum Zeitpunkt der Archivierung unproblematisch sind, z.B. weil darin über das rechtskräftige Urteil in einem aktuellen Kriminalfall berichtet wird.

Es gibt in Deutschland für jeden dieser Fälle einen späteren Zeitpunkt, zu dem es effektiv nicht mehr legal ist, die Verknüpfung von Person und Tat öffentlich zu nennen, ohne Persönlichkeitsrechte des Täters zu verletzen und, wie das Abmahn schreiben des Anwalts dann festzustellen hat, *auf unerträgliche Weise* dessen Resozialisierungsaussichten zu gefährden. Eine gesetzmäßige Regelung, ab wann nach der Urteilsverkündung dieser Zeitpunkt genau erreicht ist, existiert bis dato jedoch noch nicht. Durch allgemeine Beobachtung der Tagespresse lässt sich nur einschränken: Für verurteilte Mörder, auch von Prominenten, liegt der Zeitpunkt nach eigenen Messungen offensichtlich irgendwann vor der Haftentlassung. Am Landgericht Hamburg wurde am 7. September 2007 in der Sache 324 O 104/07 einer Klage stattgegeben, nach welcher der fragliche Zeitpunkt spätestens 6 Monate nach dem Urteilsspruch erreicht wäre; in einer Reihe ähnlich gelagerter Fälle (324 O 712/06 und 5 weitere) wurden aber am 18. Dezember 2007 solche Urteile vom zuständigen Oberlandesgericht revidiert. Es ist insgesamt auf dem derzeitigen Stand noch nicht abzusehen, wie dies künftig Bestand haben wird und ob die Schwelle vielleicht noch weiter herabgesetzt wird. Das Bundesverfassungsgericht hat in einem Kammerbeschluss vom 25.11.1999 klargestellt, dass das Lebach-Urteil und andere nicht dahingehend interpretiert werden dürften, dass eine „[...] vollständige Immunisierung vor der ungewollten Darstellung persönlichkeitsrelevanter Geschehnisse [...]“ (NJW 2000, S. 1859ff., Lebach II) beabsichtigt wäre. Zwar sollte dies alles ursprünglich nur für die neuerliche Veröffentlichung in der Presse gelten; die Zugänglichkeit einer Aussage in einem Online verfügbaren Archiv wird aber allzu leicht als neuerliche Veröffentlichung dieser Aussage gewertet.

Zwei Gerichtsentscheidungen zur Nennung der Mörder Walter Sedlmayrs mit vollem bürgerlichen Namen etwa ein Jahr vor ihrer Haftentlassung verdeutlichen die Problematik. Sie gehen letztlich zurück auf das Lebach-Urteil des Bundesverfassungsgerichts (BvR 536/72) vom 05. Juni 1973 (Lilienthal, 2001):

- Das Landgericht Frankfurt am Main hat in einer Urteilsbegründung (Aktenzeichen 2-03 O 358/06) vom 5. Oktober 2006 ausgeführt, dass es keine Pflicht zur Kontrolle und Entfernung bzw. Anonymisierung archivierter Artikel in einem Archiv gäbe – eine solche würde nicht nur

dem Informationsbedürfnis der Öffentlichkeit zuwiderlaufen, sondern auch die öffentliche Aufgabe des Archivs über Gebühr beeinträchtigen.

- Das Hanseatische Oberlandesgericht Hamburg hat sich jedoch in einer Entscheidung vom 21. März 2007 (Geschäftsnummer 324 O 952/06) allein auf die Tatsache der Veröffentlichung berufen und die Art der Darstellung im Rahmen eines Archivs nicht gewürdigt. Statt dessen wird die Nennung der vollen bürgerlichen Namen im Kontext des Verbrechens rundherum verboten.

Da nach dem Prinzip des fliegenden Gerichtstands (§32 Zivilprozessordnung) eine Klage wegen Verletzung allgemeiner Persönlichkeitsrechte durch eine Veröffentlichung derzeit überall dort in Deutschland verhandelt werden kann, wo die fragliche Veröffentlichung verfügbar ist, bei Veröffentlichungen im Internet also überall, kann ein entsprechend gerichtekundiger Anwalt die Klage natürlich bei demjenigen Gericht einreichen, an welchem die Erfolgsaussichten erfahrungsgemäß am größten sind. Ende 2008 denkt das Bundesministerium der Justiz gerade über eine Neuregelung nach, um mißbräuchlichen Auswüchsen des fliegenden Gerichtstands Herr zu werden, wie und mit welchen Folgen dies künftig konkret in Recht umgesetzt werden wird, steht noch zu erwarten.

Als Konsequenz für ein öffentlich zugängliches Online-Archiv, welches sich derzeit vor rechtlichem Ungemach schützen möchte, heißt das: Entweder sind hinreichend alte Belegstellen vorsorglich zu tilgen – wobei hinreichend alt von den Gerichten noch nicht endgültig definiert wurde, oder aber vorsorglich sind alle Eigennamen zu anonymisieren. Andernfalls muss der Zugang zum Archiv mindestens durch technische Maßnahmen auf einen kontrollierten Benutzerkreis eingeschränkt werden.

4.3 IMPLEMENTIERUNG

Die Implementierung kann in vier große Aufgabenbereiche unterteilt werden: Datenbeschaffung (Unterabschnitt 4.3.1), Aufbereiten der Daten mit statistischen (Unterabschnitt 4.3.2) und linguistischen Verfahren (Unterabschnitt 4.3.3), tägliche Analyse (Unterabschnitt 4.3.4) und Präsentation der Ergebnisse (Unterabschnitt 4.3.5). Abbildung 20 zeigt einen groben Überblick der wesentlichen Schritte und redaktionellen Eingriffsmöglichkeiten.

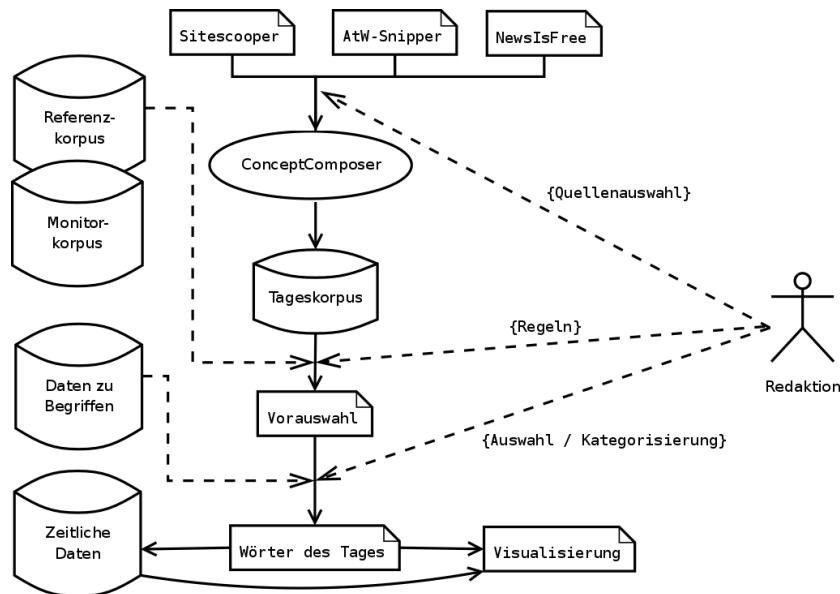


Abbildung 20: Grober Überblick: Ablaufplan Wörter des Tages

4.3.1 Daten und Datenacquire

Für die Implementierung werden grundsätzlich zwei Arten von Daten benötigt einmal ein sehr großes Referenzkorpus, welches den allgemeinen Gebrauch der beobachteten Sprache modelliert, dann für das untersuchte Zeitintervall die Texte der zu analysierenden Medien, zum Beispiel als tagesgenau unterscheidbarer Strom an Artikeln. Bei den unterschiedlich gelagerten Anwendungsfällen gab es viele Gemeinsamkeiten (siehe dazu Unter-Unterabschnitt 4.3.1.1), aber auch einige abweichende Parameter, die in Unter-Unterabschnitt 4.3.1.2 – Unter-Unterabschnitt 4.3.1.5 erläutert werden.

4.3.1.1 Allgemeines Vorgehen

Um an tagesaktuelle Daten zu kommen, gibt es zwei Wege, entweder man bekommt die Daten zur Weiterverarbeitung übergeben wie in Unter-Unterabschnitt 4.3.1.5, oder sie müssen erst einmal beschafft werden. Der naheliegende Weg hierzu ist Spidering und Crawling, also das gezielte automatisierte Herunterladen der neuen Seiten einer vorher ausgewählten Liste von Domains. Zu lösen sind dafür zwei Probleme: Wie erkennt ein Algorithmus neue Seiten und woher kommt die Liste der akzeptablen Domains.

Beim Beobachten von Tageszeitungen ist letzteres für einen

Bereich wie den deutschen Sprachraum eine relativ leichte Aufgabe, gibt es doch Verzeichnisse von Tageszeitungen und Online Medien; besonders das der Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V. (IVW), welches auch gleich noch nützliche Daten zur Verbreitung der einzelnen Medien bereitstellt, ist für Deutschland hier hervorzuheben. Für zahlreiche weitere Sprachen hat sich das Angebot von NewsIsFree (<http://www.newsisfree.com/sources/bylang/>) bewährt.

Nachrichtenmedien haben zudem die Eigenschaft, die aktuellen Nachrichten besonders sichtbar zu platzieren sowie Übersichtsseiten und Newsfeeds anzubieten. Hierdurch lässt sich mit relativ wenig Aufwand eine Liste der zu prüfenden bzw. herunterzuladenden neuen Seiten generieren. Ein gewisser manueller Pflegeaufwand bleibt freilich bestehen, weniger weil so viele Medien neu gegründet werden, sondern vor allem, weil es bei bestehenden Medien gelegentlich Adressänderungen oder komplette Relaunches gibt.

Insofern online frei verfügbare Texte ausgewertet werden, liegen sie meistens meistens als HTML oder in einem anderen weiterverarbeitbaren digitalen Textformat vor. Beschäftigt man sich allerdings mit Nachrichten in Sprachen aus etwas weiter entlegenen Gebieten der Welt wie zum Beispiel Urdu in Pakistan, kann es aber durchaus vorkommen, dass auch im Jahr 2008 Quellen, so sie denn online vorliegen sind, nicht etwa als Text sondern als Bild³ ausgeliefert werden. Ebenfalls zunächst nur Bildinformation liefern Quellen, die nur auf Papier vorliegen. Mit Hilfe einer OCR-Software könnte auch hieraus grundsätzlich digitaler Text ausgelesen werden, bei den Wörtern des Tages ist es aber außerhalb des Interesses auch diese schwieriger erschließbaren Quellen zu nutzen.

4.3.1.2 *Deutsche Presse*

Das Crawling für die deutschen Wörter des Tages hat sich im Lauf der Jahre mehrfach geändert, sowohl von der Quellenlage her als auch was die Umsetzung betrifft. Ziel war es jeweils mit den begrenzten verfügbaren Ressourcen an Arbeitszeit und Material eine vertretbar gute Quellenabdeckung zu erreichen und einen konstanten Textstrom zu gewährleisten. Dem Qualitäts-

³ Geschriebenes Urdu etwa hat eine mit Deutsch in lateinischen Buchstaben verglichen sehr komplexe Typographie mit ca. 20000 Ligaturen (Gulzar und ur Rahman, 2007) – und bis vor einigen Jahren wurden viele Zeitungen noch im Nastaliq-Stil von Hand kalligraphiert. Vergleiche auch: http://www.wired.com/culture/lifestyle/news/2007/07/last_calligraphers (abgerufen am 25. Januar 2008).

und Vollständigkeitsanspruch eines mit kommerziellen und/oder wissenschaftlichen Interessen geführten Archivs sollte dabei keine Konkurrenz gemacht werden, vielmehr stand grundsätzlich die verwendete Technologie im Vordergrund. Gleichwohl aber blieb die Absicht bestehen, dass die aufgebaute Ressource bei der exemplarischen Beantwortung von Fragen des Alltags grundsätzlich nützlich sein möge.

SED-SKRIPTE

Ein erster Schritt zur Automatisierung des Crawlings stellte die Eigenimplementierung eines Downloadskripts für eine spezifische Website dar. Hierzu war es nötig sich einen Überblick über die Hierarchie und den Aufbau jeder einzelnen Website zu machen, um zum Beispiel nach Eingabe eines Datums alle Links mit Artikeln, die an diesem Tag veröffentlicht wurden, zu erhalten und in einem zweiten Schritt herunterzuladen und anschließend den Text zu extrahieren. Skriptsprachen wie bash und sed oder Perl bieten sich für die Implementierung an. Genaue Anpassungen an die individuelle Website sind unvermeidlich und dementsprechend ist der Pflegeaufwand sehr hoch. Solche Skripte waren im Wortschatz-Projekt und anfänglich bei den Wörtern des Tages in den Jahren 1999 – 2003 im Einsatz.

SITESCOOPER

Die Software Sitiescooper (<http://sitescooper.taint.org/>) kapselt das Herunterladen der aktuellen Artikel von typischen Zeitungswebsites im Textformat in einer einfach gehaltenen Syntax. Es sind im wesentlichen die Grenzen zwischen Artikel und sonstigen Elementen sowie wenige perlkompatible reguläre Ausdrücke, zum Beispiel zum Extrahieren der gewünschten Links oder der Überschrift anzugeben. Die Programmausführung geschieht dann zu einem definierten Zeitpunkt per cron-job. Sitiescooper kümmert sich sowohl um die Extraktion des Textes als auch um die Verwaltung bereits heruntergeladener URLs. Problematisch kann dies bei Seiten werden, welche unter immer gleichen URLs zu unterschiedlichen Zeiten unterschiedlichen Content anbieten. Zudem ist die Verwendung von sitescooper wegen der Anpassung an individuelle Website-Strukturen und -Layouts immer noch pflegeaufwändig. Verglichen mit den sed-Skripten stellt sitescooper bereits aber eine deutliche Erleichterung dar. Für die Wörter des Tages und das Wortschatz-Projekt wurde Sitiescooper in den Jahren 2002 – 2006 eingesetzt.

ATW-SNIPPER

Yahoo bietet bei der Suchmaschine *All the web* eine Nachrichtensuche nach Stichwörtern, Sprache und einem Zeitfenster an, die nicht mit Stoppwörtern gefiltert wird. Da quasi jeder Artikel mindestens ein Stoppwort enthält, ergibt sich de facto per Stoppwortsuche eine Liste der Artikel pro Sprache und Tag, so man All the web in hinreichend kurzen Abständen abfragt. Um nicht die Ergebnisseiten wie bei den bisher beschriebenen Ansätzen individuell angepasst parsen zu müssen, wurde ein Text-Extraktor implementiert, welcher lange Textblöcke, welche dicht zusammenstehen in der Seite findet und auswählt. Damit reduziert sich der Pflegeaufwand effektiv auf die mit regulären Ausdrücken und statistischen Mitteln einfache „Säuberung“ des extrahierten Texts (siehe Unterabschnitt 4.3.2). Der ATW-Snipper wird seit 2006 für die Wörter des Tages und das Wortschatz-Projekt eingesetzt.

RSS-CRAWLER

Viele Nachrichtenmedien bieten selbst RSS-Feeds für ihre neuen Artikel an, die leicht automatisiert eingesammelt und weiterverarbeitet werden können. Hierzu wurde ein Crawler implementiert, welcher eine Liste von RSS-Feeds abrufen und aus den darin neuen Artikeln, analog zum ATW-Snipper, Text erzeugt. Falls einzelne Medien keine Feeds anbieten, ließen sich diese zum Beispiel über Google News erzeugen, wie in Listing B.2 skizziert, insofern dies nicht gegen die Nutzungsbedingungen von Google News verstößt. Der RSS-Crawler wird seit 2007 für die Wörter des Tages und das Wortschatz-Projekt eingesetzt.

4.3.1.3 Deutsche Blogosphäre

DEFINITION: CONSUMER-GENERATED MEDIA (CGM)

Consumer-Generated Media können als Ergänzung zum klassischen Journalismus gesehen werden, besonders wenn es darum geht, öffentliche und veröffentlichte Meinung zu beobachten und zu vergleichen. *Weblogs* sind eine Form von CGM, die mittlerweile weit verbreitet ist und deren Meinungsmacht aktuell untersucht wird (Zerfaß und Boelter (2005)). Große Korpora von Weblog-Einträgen sind kostenpflichtig oder unfrei verfügbar, etwa über TRECo6 und Nielsen Buzzmetrics; sie sind gemischt-sprachlich oder haben ein Übergewicht auf Englisch:

- http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

- <http://www.icwsm.org/data.html>

Auch Mishne (2007) beschreibt in seiner Dissertation die Anwendung von automatischen sprachverarbeitenden Verfahren auf englischsprachigen Weblogs. Ein frei verfügbares oder zumindest öffentlich zugängliches Korpus deutschsprachiger Weblogs gibt es bis dato nicht.

DATEN AUS DER DEUTSCHEN BLOGOSPHERE

Anders als bei den etablierten Medien gibt es bei Weblogs nur sehr unvollständige Verzeichnisse und Statistiken; daher wurde von Dirk Olbertz (<http://blogscout.de>) und Jens Schröder (*Deutsche Blogcharts*) das Blogcensus-Projekt gestartet, mit dem Ziel die deutschsprachigen Weblogs möglichst umfassend zu katalogisieren. Seit November 2007 werden auf der Seite <http://blogcensus.de/> monatlich Anzahlen der in den vorangegangenen 2, 4 und 6 Monaten aktiven und dem System bekannten deutschsprachigen Blogs angegeben. Die Zahlen lagen Anfang 2008 bei ca. 125 000, 170 000 bzw. 200 000 mit leicht abnehmenden Tendenz.

Aus Datenschutzgründen stehen die Daten von blogcensus nicht zur Verfügung. Statt dessen galt es für die *Wörter des Tages: Weblogs* eine eigene Liste von deutschsprachigen Weblogs zusammenzustellen. Gefordert war dabei zusätzlich, dass die Blogs auch über Newsfeeds verfügen, um das spätere Einsammeln der Einträge zu vereinheitlichen. Ein Teil der ursprünglichen Liste stammte aus öffentlich verfügbaren Quellen wie der Blogger-Karte von Bloghaus (<http://www.blogworld.de/karte/karte-index.php>), ein weiterer aus musterbasierter Analyse der Daten des Findlinks-Projekts⁴. Beide Quellen konnten aber letztlich nicht in ausreichender Menge und hinreichender Aktualität die nötigen Daten liefern. Als dritte Quelle wurden daher schließlich noch das Blogverzeichnis Technorati und schließlich auch Google Blogs hinzugezogen. Analog zum Vorgehen beim ATW-Snipper und der Datenbereitstellung für den RSS-Crawler werden dort nun aktuelle Blogbeiträge gesucht, welche typische deutsche Stoppwörter enthalten. Das Perl-Modul `XML::Feed` erleichtert dann die Ermittlung von Newsfeeds zu dem Postings. Die Newsfeeds werden dann regelmäßig abgefragt.

Für die *Wörter des Tages: Weblogs* werden so alle 12 Stunden ca. 60 000 Newsfeeds vom BlogCorpusCrawler, einem auf den

⁴ Das Findlinks-Projekt implementiert einen verteilten Crawler, welcher unter <http://wortschatz.uni-leipzig.de/findlinks/> öffentlich verfügbar ist. Ziele des Projekts sind insbesondere die Erkundung der Struktur des Web und das Sammeln von Text in seltenen Sprachen.

Perl-Modulen `LWP::UserAgent` und wiederum `XML::Feed` beruhenden Newsfeed-Crawler besucht und pro Tag ca. 30 000 neue Einträge⁵ heruntergeladen. Diese stammen nicht alle aus Blogs und sind zum Teil Spam, zum Teil auch fremdsprachig; für die Auswertung bleiben aber in der Regel noch weit über 15 000 Dokumente pro Tag übrig.

4.3.1.4 Finanzdomäne

Im Rahmen einer Machbarkeitsstudie wurde ein Ausschnitt der Finanzdomäne anhand von Veröffentlichungen der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), des Deutschen Sparkassen- und Giroverbands (DSGV) und der Deutschen Bundesbank im Zeitraum Anfang 1995 bis April 2006 untersucht. Insgesamt gingen ca. 3000 Dokumente mit 220 000 Sätzen in die Auswertung ein. Eine Besonderheit bei dieser Studie war, dass keine Trennung von Referenzkorpus und den Zeitscheibenkorpora gegeben war, sondern die untersuchten zeitlichen Teilmengen mit der Gesamtheit der Daten als Referenz verglichen wurden.

4.3.1.5 Norwegische Presse

Für die norwegische Fassung der Wörter des Tages wird als Referenzkorpus das *Norsk avis korpus* (<http://www.avis.uib.no>, vgl. (Hofland, 2000)) verwendet. Dieses Zeitungskorpus wird von der *Avdeling for kultur, språk og informasjonsteknologi* (AKSIS) an der Universität Bergen gesammelt und hatte Anfang 2006 einen Umfang von 440 Millionen laufenden Wörtern aus 1,3 Millionen Texten. Jeden Tag wird es automatisch um ca. 200 000 laufende Wörter aus tagesaktuellen Quellen erweitert.

4.3.2 Vorverarbeitung

Um den Ertrag der Webcrawler in den eigentlichen Artikeltext zu transformieren, ist ein gewisser Aufwand nötig. Abbildung 21 zeigt dies anschaulich: Nur der in den gepunkteten Kästchen enthaltene Text soll weiterverarbeitet, die Navigations- und Werbeelemente dagegen sollen ignoriert werden. Im Rahmen der Arbeit an Korpora, vgl. z.B. (Quasthoff u. a., 2006), wurden muster-

⁵ Insgesamt wurden ca. 8 Millionen Einträge von November 2006 – Januar 2008 gecrawlt. Das Crawling von Feeds und Einträgen sowie die Verwaltung der Feeds in einer Datenbank lastet einen PC (Pentium IV, 3 GHz, 512 MB RAM) fast vollständig aus.



Abbildung 21: Beispielartikel von Financial Times Deutschland (22. Januar 2007).

basierte und statistische Verfahren entwickelt, welche dies kapseln. Inhaltlich wird dabei auf vielerlei gezielt: unerwünschte Überreste der HTML-Gestalt, fehlerhaft kodierte Zeichen, fremdsprachiges Material und regelmäßig sich wiederholendes Rauschen, z.B. in Form von Werbung.

4.3.2.1 Von HTML-Files zu Sätzen

Eine Reihe von Verarbeitungsschritten befassen sich damit, wie aus den gecrawlten Webpages Sätze extrahiert werden. Scheinbar ist dies eine einfache Aufgabe. In der Praxis gilt es jedoch einige Probleme zu lösen, die nicht alle unmittelbar klar sind.

Ein erster Schritt ist es, Überreste der HTML-Darstellung zu entfernen. Nicht die HTML-Tags sind hier das Problem, sondern mangelhaft ausgezeichnete Script-Elemente und Stylesheets, eingebettete Meta-Daten, Objekte, Tabellen, Kommentare usw., die ohne Kontrolle irrtümlich als Text behandelt werden könnten.

Der Text wird dann in Sätze zerlegt. Dies ist nicht trivial: Nicht jeder Satz endet mit einem Satzzeichen (z.B. Überschriften). Nicht jeder Punkt ist (eindeutig) ein Satzende (z.B. bei Abkürzungen). Wie genau sind unterschiedliche Formen von Anführungszeichen zu behandeln? Für bekannte Sprachen kann ein Verfahren, welches eine Liste bekannter Abkürzungen mit

	10	100	1 000	10 000
Dänisch	19,63	42,58	62,74	83,10
Deutsch	26,69	48,45	65,97	82,54
Estnisch	11,61	25,92	47,62	73,28
Finnisch	10,98	20,72	37,48	62,39
Französisch	21,38	45,73	66,25	88,65
Isländisch	21,62	40,74	61,22	82,39
Italienisch	17,88	40,59	62,41	85,93
Katalanisch	24,31	45,30	65,20	87,82
Koreanisch	5,68	17,54	37,16	64,33
Niederländisch	22,53	45,23	65,78	85,54
Norwegisch	19,42	41,96	62,05	82,05
Schwedisch	18,76	40,25	60,59	80,93
Serbisch	22,25	41,57	62,20	82,19
Sorbisch	15,95	35,37	58,70	79,99
Türkisch	9,75	19,69	38,80	67,12

Tabelle 6: Prozentuale Textabdeckung der top n types

Kriterien wie „Am Satzanfang werden Wörter groß geschrieben.“ kombiniert, eine sehr gute Erfolgsquote erreichen. Steht ein Part-of-Speech Tagger zur Verfügung, kann zusätzlich der Satzbau als Kriterium einfließen.

Nicht jeder Satz im Text ist in der beobachteten Sprache. Um Artefakte auszuschließen, sollte fremdsprachiges Material entfernt werden. Tabelle 6 aus (Quasthoff u. a., 2006) zeigt, dass schon ein relativ kleines Inventar an 1 000 – 10 000 Wortformen einen sehr großen Anteil des jeweiligen Textes ausmacht. Die satzbasierte Spracherkennung⁶ in der Implementierung von Biehnann und Teresniak (2005) von macht sich dies zunutze.

Die Kookkurrenzstatistik, besonders die für niederfrequente Wörter, kann durch nahezu identische Sätze stark verfälscht werden. Deswegen werden Duplikate und Beinahe-Duplikate erkannt und entfernt. Drei typische Klassen von Beinahe-Duplikaten gibt es dabei: Solche, die sich nur in der Art der verwendeten Anführungszeichen unterscheiden, solche, die sich nur in Zahlen unterscheiden und solche, die sich nur durch einen kleinen und irrelevanten Einschub unterscheiden.

⁶ Vergleiche z.B. Dunning (1994) für eine Übersicht von Verfahren.

4.3.2.2 Von Sätzen zum Korpus

Die vorher extrahierten Sätze werden anschließend noch weiter bereinigt.

Es sollen nur vollständige Sätze erfasst werden. Typischerweise beginnen Sätze mit einem groß geschriebenen Wort und enden mit einem Satzendezeichen (d.i. Punkt, Fragezeichen, Ausrufezeichen). Optional dürfen vor dem ersten Wort im Satz und nach dem Satzendezeichen noch Anführungszeichen stehen. Außer dem Zollzeichen (") können dies auch unterschiedliche Varianten der einfachen und doppelten typographischen Anführungszeichen bzw. Umschreibungen dieser sein⁷: „, „‘, ’, „„, „”, „„“, „„», „«, „‹, ›, „„‘, ’’.

Beim Textsatz wird Text bisweilen gesperrt dargestellt. Dabei kann der Zusammenhang von Wörtern für die Maschine verloren gehen, weil die Sperrung in HTML- und PDF-Dateien häufig durch das Einfügen von Leerzeichen und nicht von Spatia (Sperrungsabständen) vorgenommen wird. Um auszuschließen, dass hierdurch Artefakte entstehen, werden Satzdarstellungen ab sechsmaliger Folge von Buchstabe und Leerzeichen als defekt angesehen.

Listen von Wörtern oder Abkürzungen wie etwa die Aufstellungen von Fussballmannschaften überlagern durch ihr häufiges ähnliches Auftreten die übrigen Kookkurrenzen. Sie enthalten viele Punkte oder Kommas und eine Obergrenze (5 für Punkte, 9 für Kommas) soll dem Abhilfe schaffen.

Bei Tabellen ist nicht ersichtlich, ob die darin enthaltene Information sprachlicher oder numerischer Art ist. Die Sätze, welche aus „unerwünschten“ Tabellen wie zum Beispiel Schachpartien, Sportresultaten oder Aktienkursen entstehen, zeichnen sich durch sehr viele sehr kurze Wörter, also einen hohen Anteil Leerzeichen pro Text aus. Eine manuelle Überprüfung ergab, dass ein Anteil ab etwa 30% Leerzeichen ein gutes Ausschlusskriterium ist.

Typische Überreste des Codes vieler Programmiersprachen enthalten die Zeichen bzw. Zeichenfolgen «, ++, *, ~, |, [, & und / in sehr hoher Zahl und Sätze, welche diese enthalten stammen quasi ausschließlich aus der Programmiersprache und nicht aus der natürlichen Sprache. Daher werden Sätze mit diesen Zeichen ausgeschlossen.

Insofern Standardsprache betrachtet wird, gilt es einige netz-

⁷ Eine Liste der Anführungszeichen inklusive der häufig anstelle dieser verwendeten Zeichen steht im Wikipedia-Eintrag zu Anführungszeichen: <http://de.wikipedia.org/w/index.php?title=Anf%C3%BChrungszeichen&oldid=41429794> (Stand 22. Januar 2007).

typische Sprachvarianten wie so genannte l33t5p34k⁸, Sätze mit vervielfachten Satzzeichen⁹ etc. zu entfernen.

Manuelle Evaluation zeigte, dass Sätze, welche sehr lange Ziffernsequenzen (mehr als 15) oder Folgen von Großbuchstaben (mehr als 20) enthalten, inhaltlich zu einem sehr hohen Anteil nichts sagend waren, somit werden auch diese entfernt.

4.3.2.3 Vom Korpus zur Textdatenbank

Die Referenzdatenbank verwendet das im Wortschatzprojekt übliche Datenbankschema. Dieses besteht aus Listen für Wörter, Sätze und Quellen mit Zuordnungen dieser zu eindeutigen Nummern (IDs), inversen Verknüpfungen der IDs von Wörtern mit Sätzen und von Sätzen mit Quellen, sowie Kookkurrenzen und anderen Verknüpfungen zwischen den IDs von Wörtern untereinander. Es kann zusammen mit einer Software zum Betrachten solcher befüllten Datenbanken von <http://corpora.uni-leipzig.de/download.html> heruntergeladen werden.

Für die wie weiter unten beschrieben täglich anfallenden Daten sind hieran einige zusätzliche Maßnahmen nötig. Zum einen müssen die ID-Listen kontinuierlich gepflegt und erweitert werden. Dann müssen die Tabellen jeweils um eine Datumsspalte ergänzt werden, was die Größe der Tabellen über die Jahre auf ein Vielfaches der Größe der Tabellen des Referenzkorpus wachsen lässt. Da dies früher oder später (abhängig von der Leistungsfähigkeit der verfügbaren Hardware und dem betriebenen MySQL-Tuning) zu Performanceproblemen führt, sollte für sehr große Datenmengen (Daten mehrerer Jahre) eine angepasste Lösung zur Datenspeicherung und -bereitstellung verwendet werden, wie sie mit WCTAnalyze (siehe Unter-Unterabschnitt 4.4.1.2) verfügbar ist.

4.3.3 Linguistische Aufbereitung

Neben der Vorverarbeitung und -filterung, die musterbasiert auf der Ansicht des Textes geschieht, zielt die linguistische Aufbereitung darauf, den Text mit zusätzlicher Information anzureichern und dann, darauf aufbauend, aus dem Text Wissen zu extrahieren.

⁸ In der *Leetspeak* sind Buchstaben oftmals durch Ziffern oder Zeichenfolgen ersetzt. Siehe auch den Eintrag zur Leetspeak in der englischsprachigen Wikipedia <http://en.wikipedia.org/w/index.php?title=Leet&oldid=185550821> (abgerufen am 21. Januar)

⁹ „Multiple exclamation marks,’ he went on, shaking his head, ‘are a sure sign of a diseased mind.’“ (Terry Pratchett – Eric)

4.3.3.1 *Part of Speech Tagging und Grundformenreduktion*

Part of Speech (POS) Tagging, das Zuordnen von Wortklassen zu Wortformen, und Grundformenreduktion, die Rückführung der verschieden geschriebenen Formen eines Wortes auf eine gemeinsame Grundform, sind wesentliche Schritte der linguistischen Aufbereitung.

Da für die Auswahl der Wörter des Tages ein vor allem schnelles Tagging und nur eine grobe Unterscheidung in Nomina, Verben, Adjektive und Sonstiges benötigt wird, ist der Einsatz eines qualitativ hochwertigen, aber vergleichsweise langsamen Taggers nicht sinnvoll. Die Ergebnisse eines im Einsatz schnellen unüberwachten Taggers wie er in (Biemann, 2007) beschrieben wird, genügen völlig, wenn nach der Trainingsphase die wenigen interessanten Wortklassen markiert werden.

GRUNDSÄTZLICHES VORGEHEN

Für die *Ord i Dag* wurde ein verwandtes Verfahren eingesetzt: Das vereinfachte Reengineering eines bekannt guten Taggers anhand von dessen Output. Der *Oslo-Bergen-taggeren* (Johannessen u. a., 2000) ist ein in LISP implementierter, qualitativ hochwertiger Tagger, der regelbasiert auf Grundlage logischer Constraints arbeitet: Neben der Zuordnung von Wortklassen kann dieser auch Grundformen und syntaktische Funktionen angeben; zum einen ist dies mehr Information als für Wörter des Tages benötigt wird, zum anderen ist er vergleichsweise langsam, so dass keine Aktualität mehr gewährleistet wäre, müssten die Ausgaben des Taggers abgewartet werden – Wörter von Vorgestern wären für Benutzende jedoch nicht besonders attraktiv.

Die Reimplementierung des Taggers geschah wie folgt: Zunächst wurde ein Teil des Referenzkorpus getaggt. Mit den erhaltenen Tripeln (Wort, Tag, Grundform) wurden Klassifikatoren trainiert, die mittels der Implementierung von Compact Patricia Tries (CPT) aus (Witschel und Biemann, 2005) zu einer Zeichenkette eine Klasse zurückliefern. Die Anzahl der Klassen ist dabei frei wählbar und die Zuordnung geschieht so, dass die Trainingsdaten exakt reproduziert werden. Da CPTs eine kompakte Repräsentation von Daten darstellen, auf welche sehr effizient zugegriffen werden kann, eignen sie sich hervorragend für die Bereitstellung sehr umfangreicher lexikalischer Komponenten. Eine weitere nützliche Eigenschaft von CPTs ist ihr generalisierendes Verhalten durch welches anhand bekannter Wörter auch vorher unbekannte Zeichenketten klassifiziert werden können. Hat der Klassifikator zum Beispiel gelernt, dass der Begriff *Nasenbären* ein Nomen mit der Grundform *Nasenbär* ist, wird er

den einmal als unbekannt angenommenen *Ameisenbären* anhand der langen gleichen Wortenden *senbären* der Klasse Nomen und der Grundform *Ameisenbär* zuordnen. Sollten mehrere Klassifikationen für ein neu gesehenes Wort möglich sein, so wird eine Verteilung der zutreffenden Klassen zurückgeliefert.

TAGGEN MIT CPTS

Da nur rudimentäre Wortklassen-Angaben für die Filterung benötigt werden, wurde das Tagset des Oslo-Bergen-Taggers auf die Grundlegenden Kategorien Nomen (N), Verb (V), Adjektiv (A), Adverb (AV), Zahl (C), Satzzeichen (IM) und Sonstige (S) reduziert und anhand der 100 000 häufigsten Wortformen wurde dann je ein Wortanfangs- und ein Wortende-CPT trainiert. Dabei wurde nur eine Wortklasse pro zu lernender Zeichenkette erlaubt, falls ein Wort mit mehreren POS Tags vorkam, wurde der häufigste Fall angenommen, was für die zu leistende Aufgabe ausreicht. Die verwendete Wortliste deckt norwegischen Text etwa zu 96,3% ab. Für all diese Wörter gibt der Klassifikator genau ein Tag zurück. Gibt es für unbekannte Wörter mehr als eine passende Klasse, entscheidet der Algorithmus anhand der Verteilung der potentiellen Wortklassen, welche davon am wahrscheinlichsten ist. Der so reimplementierte Unigramm POS Tagger ist verglichen mit dem Original sehr schnell, hat aber eine mindere Qualität, die für andere Aufgaben als die hier erforderlichen möglicherweise nicht ausreicht. Bessere Qualität bei sonst ähnlichem Vorgehen könnte z.B. mit einem Tagger, der auf einem Hidden-Markov-Model und dem Viterbi-Algorithmus (Viterbi, 1967) beruht, erzielt werden, was im Fall einer Neuimplementierung auch zum Zuge käme.

GRUNDFORMENREDUKTION MIT CPTS

Im Gegensatz zur Qualität des Part-of-Speech-Tagging, kann die Grundformenreduktion mittels CPTs mit dem *state of the art* mithalten. Der CPT wird auf einer Liste von Vollform-Grundform-Zuordnungen auf Regeln zur Umformung von Vollformen zu Grundformen trainiert.

Diese Reduktionsregeln bestehen aus zwei Teilen: Einer Zahl, welche angibt, wie viele Zeichen am Wortende der Vollform abzuschneiden sind und einer optionalen Zeichenkette, welche nach dem Abschneiden wieder angehängt wird. Damit läßt sich die Flexion von Wörtern kompakt beschreiben und somit das Verhalten von Wörtern mit gleicher Flexion zusammenfassen. Für die offenen Wortklassen Nomen, Verb und Adjektiv aus dem Tagging weiter oben, wird je ein Wortende-CPT trainiert.

norw. Vollform	stå	stående	står	stås	stått	sto	stod
Reduktionsregel	0	4	1	1	2	1å	2å

Tabelle 7: Reduktionsregeln für die Vollformen des norwegischen Wortes „stå“ (dt. *stehen*).

4.3.3.2 Extraktion von Eigennamen

Die Identifikation und Extraktion von Eigennamen ist ein weiterer wichtiger Aspekt der linguistischen Verarbeitung der Nachrichten.

Für Benutzende sind nämlich Eigennamen in allen Themenbereichen von besonderem Interesse, da sie ihnen zum einen konkrete Handlungen, Ereignisse usw. zurechnen können und sie sich zum anderen auch direkt für sie in als Prominente interessieren. Bei *Ord i dag* wurde die in (Quasthoff u. a., 2002a) beschriebene Methode zur Eigennamenerkennung auf dem AVIS-Korpus angewandt. Diese kommt mit sehr wenig Aufwand an Vorwissen und Überwachung aus: Der Algorithmus nimmt Listen von bekannten Vor- und Nachnamen, Titeln und anderen Namensbestandteilen sowie eine Menge an Klassifikationsregeln und Extraktionsmustern als Eingabe, insgesamt nur wenige hundert Einträge. Dann werden iterativ neue Namensbestandteile auf einem unannotierten Korpus gelernt. Da die Form von Namen einen sehr hohen Grad an Regularität aufweist, funktioniert der Bootstrapping-Prozess sehr gut und liefert letztlich eine sehr umfangreiche und Liste von Namen¹⁰ mit sehr wenigen Fehlern.

Die Klassifikationsregeln spezifizieren Muster nach denen Namen getaggt werden. Zum Beispiel würde eine Regel wie TIT CAP* LN $\rightarrow$ FN bedeuten, dass ein groß geschriebenes Wort, welches zwischen einem Titel und einem Nachnamen steht, ein Vorname ist. Um die *precision* des Klassifikators zu verbessern, wird so eine Entscheidung nicht anhand von nur einem Vorkommen getroffen, sondern muss in anderem Kontexten Bestätigung finden, um akzeptiert zu werden. Per Iteration dieser Regeln, wurde der sehr umfangreiche korpuspezifischer GAZETTER aufgebaut, welcher bei *Ord i Dag* zur Eigennamenerkennung genutzt wird. Durch Anwendung des Verfahrens auf die Eingangsdaten der täglichen Verarbeitung wächst dieser dynamisch weiter; die extrahierten Eigennamen (zum Beispiel *Jens Stoltenberg*) und Verbindungen von Titeln mit Eigennamen (zum Beispiel *statsminister Jens Stoltenberg*) werden für die in nächsten Abschnitt beschriebene tägliche Analyse als Mehrwortbegriffe verwendet.

¹⁰ Im Falle des AVIS-Korpus waren dies ca. 300 000.

4.3.4 *Tägliche Verarbeitung*

Auch bei der täglichen Verarbeitung werden die Daten wie vorher beschrieben vorverarbeitet. Die Tageskorpora werden analog zum Referenzkorpus aufgebaut und aus ihnen dann dann Kandidaten ermittelt (siehe Unter-Unterabschnitt 4.3.4.1), geordnet (siehe Unter-Unterabschnitt 4.3.4.2) und präsentiert (siehe Unterabschnitt 4.3.5).

4.3.4.1 *Auswahl von Kandidaten*

Es ist zentralem Interesse bei den Wörtern des Tages, eine Menge von Wörtern so auszuwählen und anzuordnen, dass ein Mensch daraus die wesentlichen Aspekte der analysierten Nachrichtmenge erkennen kann. Die Menge soll eine überschaubare Zahl hierfür möglichst aussagekräftiger Wörtern enthalten. Zudem soll die Vernetzung dieser Wörter untereinander betrachtet werden:

- Nomina und Phrasen eignen sich in besonderem Maße als Kandidaten, da sie ein hohes Maß Information pro Einheit tragen können. Daher hat auch z.B. (Witschel, 2005) Nomina als Baseline für die Evaluation von Methoden zur Terminologieextraktion benutzt. Als Kandidaten für die Wörter des Tages werden ausschließlich Nomina betrachtet.
- Wörter, welche nicht allgemein bekannt sind, wirken auf den Benutzer nicht erhellend, folglich wird eine Häufigkeitsschwelle im Referenzkorpus, ähnlich wie für die Aufnahme in ein Wörterbuch, für sinnvoll erachtet und es gibt für Wörter, die gar nicht im Referenzkorpus vorkommen, eine zusätzliche Schwelle zu überwinden, um als Kandidat akzeptiert zu werden.
- Allzu häufige Wörter aus dem Referenzkorpus tragen meistens zu wenig spezifische Information, folglich werden sie als Stoppwörter von der Auswahl ausgeschlossen.
- Neu auftretende Wörter, die nicht in mehreren verschiedenen Quellen vorkommen, sind vermutlich Artefakte dieser bestimmten Quelle, z.B. eines einzigen Interviews und sollten folglich ignoriert werden. Wenn aber andererseits die gleichen Wörter in vielen verschiedenen Quellen neu auftreten, tragen diese besonders wichtige Information, vermutlich über ein aktuelles Ereignis.

- Die Kookkurrenzen gleichmäßig verteilten mittelfrequenter Wörter im Referenzkorpus sagt besonders viel über die gewöhnliche Verwendung dieser Wörter aus. Taucht bei solch einem Wort plötzlich ein Peak auf, sind die Kookkurrenzen des Wortes zum Zeitpunkt des Peaks besonders interessante Kandidaten zur Beschreibung des zugehörigen Ereignisses.
- Wortformen, welche die gleiche Entität bezeichnen, werden zusammengefasst betrachtet werden und erhalten die Grundform als angezeigtes Label.

Die eigentliche Auswahl der Kandidaten geschieht wie folgt gestaffelt nach den Unterscheidungen, ob ein Wort im Referenzkorpus enthalten ist und ob es ein Mehrwortbegriff ist. Für Mehrwortbegriffe werden generell niedrigere Schwellwerte bei den absoluten Frequenzen verwendet, da diese durchwegs deutlich weniger frequent als vergleichbare übrige Wörter sind. Im Fall des Vorkommens im Referenzkorpus wird eine Differenzanalyse (vgl. Heyer u. a. (2006), Kap. 4.2) basierend auf einer Kombination des Poissonmaßes¹¹ und den absoluten Frequenzen angewandt. Ein solches Wort wird Kandidat, wenn die Poissonsignifikanz einen Schwellwert übersteigt und bei den absoluten Häufigkeiten jeweils Mindestanzahlen ebenfalls überschritten werden. Der letztere Wert ist von der Größe des Referenzkorpus und der durchschnittlichen Größe der Zeitscheiben abhängig. Mit dem ersten Wert lassen sich die Sicherheit der Auswahl gegen die Größe der ausgewählten Menge abwägen. Falls das Wort nicht im Referenzkorpus vorkommt, kann es zunächst einmal nur Kandidat werden, falls es die doppelte geforderte absolute Mindesthäufigkeit in der Zeitscheibe hat.

In einem zweiten Differenzanalyse-Schritt wird ein Monitor-korpus zum Einsatz gebracht. Damit ist eine Art zweites Referenzkorpus gemeint, welches nicht aus einer historischen sehr großen Menge an Dokumenten besteht, sondern aus den zeitlich nächsten Zeitscheiben gebildet wird, zum Beispiel bei Ord i dag aus den Dokumenten der vorangegangenen fünf Wochen. Sinn des Monitorkorpus ist es den aktuellen Trend in der Verwendung von Wörtern zu berücksichtigen. Speziell soll dadurch vermieden werden, dass Veränderungen mit langfristige Bestand, die nach der Erstellung des Referenzkorpus geschehen sind, aktuelle Ereignisse überschatten.

¹¹ Neben dem Poissonmaß wurden auch andere Maße wie log-likelihood für die Selektion ausprobiert, die Resultate unterschieden sich im Ranking quasi nur bei Wörtern, die so selten sind, dass sie am Häufigkeitskriterium scheitern.

Folgendes Beispiel aus der Politik illustriert die Problematik: Angenommen der Schwerpunkt des Referenzkorpus liegt auf dem Zeitraum, als Gerhard Schröder noch Bundeskanzler war. Nachdem dann Angela Merkel Bundeskanzlerin geworden ist, wird über Gerhard Schröder nach einer Phase weniger Erwähnungen auf einmal im Zusammenhang mit der Ostsee-Pipeline wieder mehr berichtet. Das Referenzkorpus erwartet aber, dass Gerhard Schröder (als Bundeskanzler) noch viel häufiger erwähnt wird, die Erwähnung als Aufsichtsrat an diesem Tag wird also fehlerhafterweise nicht ausgewählt. Angela Merkel andererseits wird nach der Wahr quasi jeden Tag ausgewählt werden, weil alleine durch den (bald nicht mehr neuen) Umstand, dass sie nun Bundeskanzlerin ist, sehr viel über sie geschrieben wird. Das Monitorkorpus korrigiert beide Fehler: Da Vorkommen von Gerhard Schröder in dem vom Monitorkorpus abgedeckten Zeitraum sehr selten geworden sind, wird er aufgrund des lokalen Peaks korrekterweise zusätzlich ausgewählt. Wenn es bei Angela Merkel keinen lokalen Peak gibt, wird sie aus der im ersten Schritt gebildeten Liste wieder entfernt, weil ihre Anzahl der Nennungen sich zwar verglichen mit der historischen Referenz signifikant verändert hat, diese Veränderung nur eben nicht am aktuellen Tag geschehen ist und folglich nicht zu den relevanten Nachrichten dieses Tages gehört.

Prinzipiell ist es auch möglich, die Auswahl ausschließlich mit den Daten des Monitor-Korpus vorzunehmen, da dieses jedoch tendenziell von geringem Umfang ist, sind Abdeckung und Qualität der statistischen Aussagen beim großen Referenzkorpus besser.

4.3.4.2 *Ordnung der Kandidaten*

Die Kandidatenauswahl erzeugt eine unübersichtliche Menge von Begriffen. Je nach Größe des Zeitscheibenkorpus, der darin widergespiegelten Themenvielfalt und den gewählten Parametern kann der Umfang der Menge variieren. Um im Sinne einer Übersicht auf eine Bildschirmseite zu passen, sollten zwischen 100 und 200 Terme ausgewählt werden. Damit der Benutzer sich in der Menge zurechtfindet, sollte diese nach einem Ordnungsschema gruppiert und ausgezeichnet werden. Für zwei Strategien wurden dazu Lösungen implementiert, einmal eine Klassifikation nach typischen Zeitungskategorien, dann ein Clustering nach Themen.

KLASSIFIKATION

Der ad-hoc Ansatz zur Klassifikation der deutschen Wörter des Tages war, dass unklassifizierte Wörter von Hand genau einer Kategorie zugeordnet wurden. War ein Wort einmal klassifiziert, wurde diese Information automatisch wiederverwendet. Neben dem offenkundigen Nachteil, dass dieses Verfahren tägliche manuelle Eingriffe erfordert, ist es auch noch unflexibel und führt bei mehrdeutigen Wörtern und auch bei solchen, die in mehreren Kategorien auftauchen können zu unbefriedigenden Ergebnissen.

Deutlich flexibler ist und bessere Ergebnisse liefert eine Klassifikation, welche den Kontext des jeweiligen Wortes betrachtet und nach Indizien für die jeweiligen Kategorien sucht. Dies wurde für *Ord i Dag* mit Hilfe eines einfachen Sprachmodells implementiert, welches auf der Grundlage einer bekannten Zuordnung von Artikeln zu Themen gelernt wurde.

Das AVIS-Korpus enthält zu jedem Artikel auch die Quell-URL und man konnte es sich zunutze machen, dass in den Pfaden der meisten dieser URLs für die jeweiligen Kategorien sprechende Bestandteile vorkommen. Zu Lösen blieb noch die Aufgabe, die etlichen tausend entstehenden Wörter und Wortfragmente in ein übersichtliches Klassifikationsschema zusammenzuführen. Da die meisten davon eher selten sind, konnte eine Vorfilterung auf nur noch 145 besonders häufige Bezeichner vorgenommen werden. Eliminierung von unsinnigen Kandidaten und Zusammenfassung von unterschiedlichen Schreibweisen für den gleichen Sachverhalt, lieferte hier schließlich 11 Kategorien: Regional (Regionales), Innenriks (Inland), Utenriks (Ausland), Politik (Politik), Økonomi (Wirtschaft), Kultur (Kultur), Sport (Sport), Utdanning (Erziehung und Forschung), Bil (Auto), Forbruker (Verbraucher) und Teknologi (Technologie). Für Deutsch wurden analog auch Klassifikationen nach *DIE ZEIT* und *SPIEGEL ONLINE* erstellt.

Für jedes Paar von Wort und Kategorie wird dann auf der Grundlage des Korpus die statistische Abhängigkeit mit Hilfe der likelihood ratio (Dunning, 1993) getestet. Dies liefert umso größere Werte, wenn ein Wort in einer gegebenen Kategorie signifikant häufiger auftaucht als im Durchschnitt und negative Werte, wenn ein Wort unterrepräsentiert auftritt. Durch den Einsatz eines Schätzers wie dem Good-Turing Estimator (Gale und Sampson, 1995) könnte auch für Wörter, welche in einer Kategorie gar nicht vorkommen ein Wert berechnet werden.

Ein neuer Text, zum Beispiel die Belegstellen für ein Wort und damit quasi dieses über seinen Kontext kann nun anhand dieses Sprachmodells klassifiziert werden. Für die Darstellung der

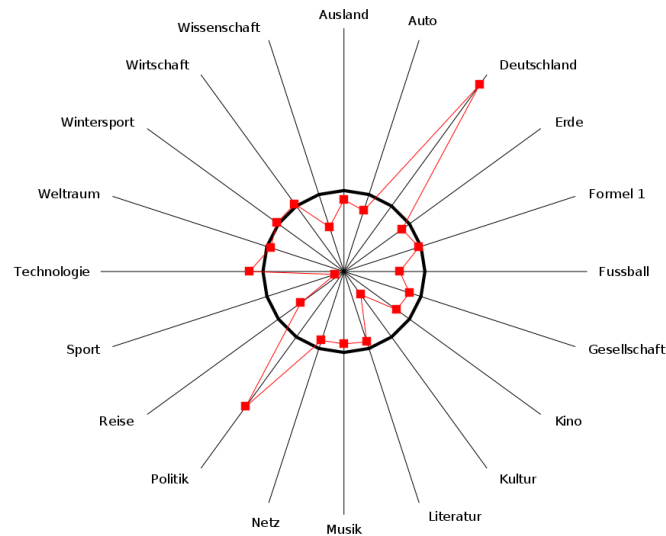


Abbildung 22: Spinnennetzdiagramm für den Begriff „Subventions-Heuschrecke“ am 17. Januar 2008 mit Kategorien von Spiegel Online. Die fette schwarze Linie ist die Nulllinie.

*Ord i Dag*¹² geschieht dies mit Hilfe eines Rocchio-Klassifikators (Cohen und Singer, 1996) und einigen Gewichtungs-Parametern für die Signifikanzwerte im Sprachmodell und die Frequenzen der Wörter im Referenzkorpus, die so gewählt wurden, dass in einer zehnfachen Kreuzvalidierung mit jeweils 90 000 Sätzen Trainingsmenge und 10 000 Sätzen Testmenge auf Basis von Sätzen stabil eine *precision* von 90% und ein *recall* von 50% erreicht wurde. Da es in der Anwendung zu jedem Begriff mindestens drei und meistens mehr als zehn Beispielsätze gibt, über welche summiert wird, kann quasi jedes Wort zugeordnet werden und es sind grundlegende Fehl kategorisierungen weitgehend ausgeschlossen.

Alternativ zur Auswahl der häufigsten Kategorie, kann das rechnerische Ergebnis der Klassifikation auch visualisiert werden, vielleicht am ehesten intuitiv in einem normierten Spinnennetzdiagramm: Für jede Kategorie wird dabei die ermittelte Maßzahl relativ zum Maximum betrachtet und an der jeweiligen Achse abgetragen, wie das in Abbildung 22 exemplarisch für den Begriff „Subventions-Heuschrecke“ am 17. Januar 2008 in der Klassifikation nach Spiegel Online dargestellt ist. Durch den Vergleich solcher Diagramme lässt sich auf einen Blick ablesen, ob in welchem Maß zwei Wörter in unterschiedliche oder gleiche Kategorien und Kontexte fallen.

¹² (vgl. hierzu Listing Listing B.3)

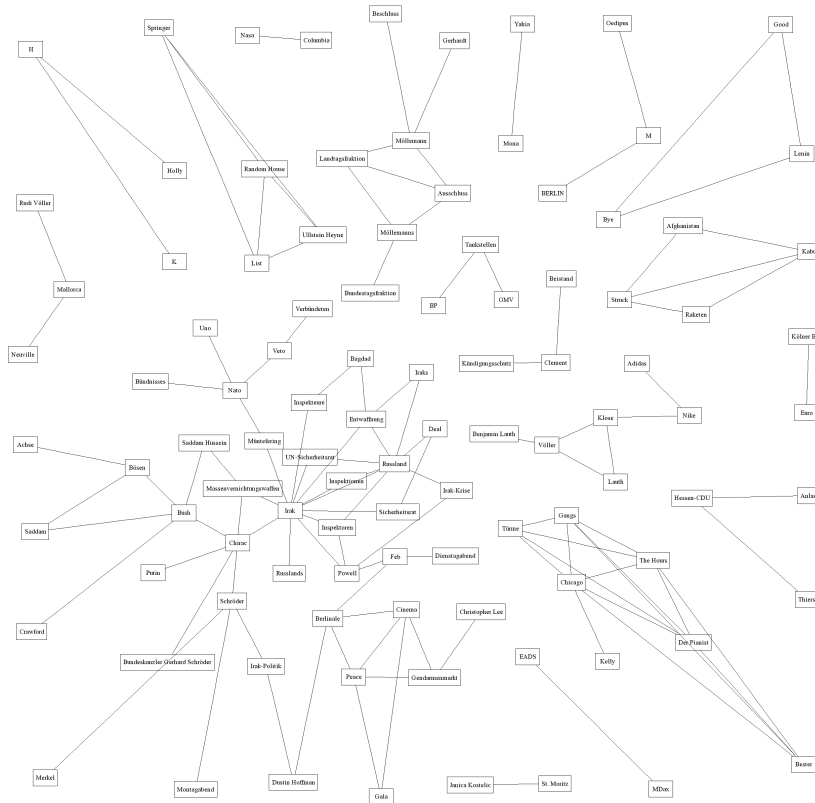


Abbildung 23: Kookkurrenz-Cluster der Wörter des Tages vom 12. Februar 2003.

CLUSTERING

Bei der Ordnung der Kandidaten per Clustering geht der Fokus vom einzelnen Wort über auf Gruppen von Wörtern, die in einem Zusammenhang stehen. Formal ist dieser nur statistisch, tatsächlich steckt aber in der Regel auch etwas thematisches dahinter. Die grundsätzliche Praktikabilität einer Präsentation der Wörter des Tages in Clustern motiviert Abbildung 23: eine direkte visuelle Umsetzung der Kandidatenmenge mit ihren Kookkurrenzen durch das graphviz-Paket¹³. In der Darstellung als Kookkurrenz-Graph sind allerdings noch einige Komponenten verbunden, die hierzu besser aufgebrochen würden, zum Beispiel steht *Dustin Hoffmann* zwischen den Kontexten *Berlinale* und *Irak-Krieg* (weil er sich am Rande der Berlinale 2003 zum Irak-Krieg geäußert hat) ohne, dass diese Kontexte alleine durch ihn zu einem großen Thema zusammengefasst werden sollten. Schon ein einfaches Clusteringverfahren wie k-means (MacQueen, 1967) kann diese Unterscheidung leisten.

¹³ <http://www.graphviz.org/>, Zugriff am 17. Januar 2008.

Die Cluster sind somit normalerweise von feinerer Granularität als die Kategorien der vorher betrachteten Klassifikation und umreißen genau ein Thema. Für eine übersichtliche Darstellung ist es sehr hilfreich, wenn die Cluster möglichst knappe und treffende Beschreibungen bzw. Überschriften haben und wenn diese über die Zeit vererbt werden können, um den allmählichen Wandel von Clustern verfolgen zu können. In Abbildung 24 (b) ist durch die Überschrift *Israel* eine sehr viel genauere Aussage getroffen als in (a), wo *israeler*, *Jeriko* etc. mit anderen, unverbundenen Begriffen zusammen in der Kategorie *Utenriks* stehen.

Das Clustering und das Vergeben von Überschriften geschieht wie folgt:

- **Auswahl von Features:** Die einzelnen Wörter des Tages werden anhand der Beispielsätze geclustert, in welchen sie im Untersuchungszeitraum vorkommen. Als Features werden dabei diejenigen anderen Wörter des Tages gewählt, die ebenfalls in diesen Sätzen vorkommen; die Gewichtung im Feature-Vektor geschieht nach Frequenz. Im Gegensatz zum Graphen oben werden also alle Wörter des Tages benutzt und nicht nur die kookkurrierenden.
- **Anwendung des Clustering-Verfahrens:** Die Wörter werden anhand ihrer Feature-Vektoren mit Hilfe des Cosinus-Maßes auf Ähnlichkeit überprüft und mit dem *k-means-Verfahren* geclustert. Da für k-means die Zahl der Ergebniscluster vorher feststehen muss, wird diese mit einem einfachen Ähnlichkeitsvergleich vorher abgeschätzt.
- **Zuordnung von Überschriften:** Bei den *Ord i Dag* geschah die Zuordnung von initialen Überschriften zu den ermittelten Clustern von Hand. Anhand der beim Clustern zum Tragen gekommenen Beispielsätze ließe sich aber auch ein automatisches Verfahren, z.B. auf der Basis von Terminologieextraktion (Witschel, 2005), deken. Dieses sollte relativ allgemeine und nicht zu viele Begriffe auswählen. Die Kombination aus Zentroid der Clustervektoren, Überschrift und Datum wird jeweils gespeichert.
- **Vererbung von Überschriften:** Nachdem initiale Überschriften zu Zentroiden zugeordnet sind, können bei folgenden Durchläufen auftretende Cluster ihrer Ähnlichkeit zu bestehenden Clustern nach klassifiziert werden. Zum Einsatz kommt hierbei wieder die bereits weiter oben verwendete Rocchio-Methode. Wenn ein neues Cluster zu mindestens 30% zu mindestens einem bestehenden Cluster ähnlich ist,

wird die Überschrift des ähnlichsten bestehenden Clusters an das neue Cluster vererbt.

Erfahrungen mit dem Verfahren zeigten, dass das System relativ schnell lernte und schon nach wenigen Tagen eine größere Zahl Cluster automatisch gelabelt wurden, einige wenige davon allerdings auch falsch. Ein grundsätzliches Problem dieser Cluster-Analyse bleibt aber der manuelle Pflegeaufwand, den sie erfordert. Denn zwar geschieht das Clustering vollautomatisch und die Titel für viele Cluster werden grundsätzlich über die Zeit vererbt, das Vergeben der Überschriften erfordert dabei aber ein Abwägen zwischen Präzision (*5 Tote bei Bombenanschlag in Tel Aviv*) und Verallgemeinerbarkeit (*Gewalt, Israel*), die sich nicht unbedingt (leicht) algorithmisch ausdrücken lässt und bei der menschliche Klassifikatoren ebenfalls Fehler machen. Einmal gewählte Überschriften können sich auch erst im Lauf der Zeit als ungeschickt gewählt herausstellen, ohne zunächst falsch gewesen zu sein; zum Beispiel kann eine Reihe von *Hitzerekord*-Clustern aus dem Sommer im Winter zu einer *Jahrhundertkälte* sehr ähnlich, aber nicht mehr korrekt, sein. Unter der gemeinsamen Überschrift *Wetterextreme* gefasst, bliebe es richtig. Dies ist freilich ein grundsätzliches Problem von Begriffshierarchien in der Beschlagwortung und keine Besonderheit dieser Implementierung von Nachrichtenclustern.

4.3.5 Präsentation

Die Ergebnisse der Auswahl werden sowohl für menschliche Benutzer (siehe Unter-Unterabschnitt 4.3.5.1) als auch für die Maschine (siehe Unter-Unterabschnitt 4.3.5.3) aufbereitet.

4.3.5.1 Benutzerschnittstelle

Die Benutzerschnittstelle hat unterschiedliche Arten von Komponenten: Solche, welche die im vorigen Abschnitt ausgewählten Terme übersichtlich darstellen, solche, die einen Überblick über Dokumente darstellen, solche, die Information über ein konkretes Wort an einem konkreten Tag liefern, und schließlich solche, die komplexere Benutzeranfragen und freies Browsen für den gesamten Datenbestand unterstützen.

ÜBERSICHTSKOMPONENTEN

In Abbildung 24 sind die beiden Übersichtskomponenten dargestellt, welche die Daten der Kategorisierung (a) und des Clusterings (b) aus dem vorigen Abschnitt verwenden. Ein Klick

Politikk	Djupedal · Jens Stoltenberg · kunnskapsminister Øystein Djupedal · Kvinnherad · Linn · nynorsk · Riis-Johansen · Åslaug
Økonomi	Eustace · Fast · Forbrukerråd · Fosen · Helge Lund · HSH · Industri · Joakim Lystad · kjøtt · kredittsyn · Kutaragi · Lystad · Marine · Mattilsynet · Memo · reklame · salta · råd · Sony · UiB · varehandel
Kultur	Alnæs · Billy Bob Thornton · Boys · DumDum · forfatter · forlag · Gyldendal · Irene Falck Jensen · Jonas Field · kunst · Liones · Peer Gynt · Sheridan · Ullmann
Sport	Aalbu · Aamodt · Ajax · Aksel · Aksel Lund Svindal · Andreas Liones · Antoine Deneriaz · Arnold Palmer · Bay · Beckie · Bell · Bjergen · Björn · Bjørnar · Bjørnstad · Bolton · Bård · Bård Bjerkeland · Carola · Dalby · Dan Pettersen · Deneriaz · Ella · Fjeld · Guro · Guro Strøm · Solli · Henrik · Hill · Iditarod · Int · Jeff · Jonas · Keane · Kjetil André Aamodt · Kjus · Leganger · Lillestrøm · Lind · Lund · Mountain · Nome · Ollers · Owen · Paralympics · Peter Fill · Rønning · Rösler · Scott · sjekkpunkt · Steinar Ege · Svindal · Thon · verdenscup · Vidar Ruud · Villegas · Vladimir Voskoboinikov · White · Åre
Innenriks	ANR · Avis · bakterie · barnehageplass · bensinstasjon · Folkehelseinstitutt · Gilde · gjemingsmann · hagle · innsjø · Is · Jenta · Karasjok · Karsten · Karsten Alnæs · Kjeller · kjøttdeig · Lennart · Lidl · Lofoten · mullah · Nyhetsbyrå · nyresvikt · Oppland · psykiater · Ringstad · Rudshøgda · Schjætvat · skudd · skytedrama · smittekilde · Statoil-stasjon · statsminister Jens Stoltenberg · Thorassen · Ullmann · universitetssykehus
Utenriks	Ahmed · begravelse · besogad · Berlusconi · Bjerkeland · fengsel · fugl · fugleinfluensa · islam · israeler · Jeriko · king · Mahmoud Abbas · Mitrovica · observatør · onsdag · politimann · Riksmeklingsmann · Red-Larsen · Saadat · seldan · smitte · Thaksin · Whitney
Regional	dagmamma · Froland · slakteri · Torget · Yahoo
Bil	Alfa Romeo · best · sek
Forbruker	diabetes · medisin · PlayStation
Teknologi	Google
«14.03.2006» Words of the Day [cluster view]	

(a) Kategorien der *Ord i Dag* am 15. März 2006

Clusters of the day: 15-03-2006	
Google	Eustace · Fast · Google · Yahoo ·
Israel	Ahmed · fengsel · israeler · Jeriko · Mahmoud Abbas · observatør · Saadat ·
Kultur	Alnæs · forfatter · forlag · Gyldendal · Karsten Alnæs · Karsten ·
Kultur	Ullmann · Linn · kunst ·
Nyheter · Innenriks	bensinstasjon · Bjerkeland · Bård Bjerkeland · Bård · hagle · Kjeller · Lillestrøm · politimann · Rösler · skytedrama ·
Nyheter · Innenriks	Statoil-stasjon ·
Nyheter · Innenriks	bakterie · Folkehelseinstitutt · Forbrukerråd · Fosen · Froland · Gilde · Joakim Lystad · kjøttdeig · kjøtt · Lidl · Lystad ·
Nyheter · Innenriks	Mattilsynet · råd · Rudshøgda · slakteri · smitte · smittekilde ·
Nyheter · Innenriks	Dalby · Jenta · nyresvikt · Oppland ·
Politikk · Innenriks	barnehageplass · Djupedal · kunnskapsminister Øystein Djupedal ·
Sport	Bjørnar · Iditarod · Jeff · King · Mountain · Nome · sjekkpunkt · White ·
Sport · ski	Beckie · Bjergen · Ella · Guro · Guro Strøm Solli · Scott ·
Sport · ski	Aamodt · Aksel · Aksel Lund Svindal · Antoine Deneriaz · Deneriaz · Kjetil André Aamodt · Kjus · Paralympics · Peter Fill ·
Sport · ski	Åre · Schjætvat · Svindal · verdenscup ·
Teknologi · spill	Björn · Lind · Rønning ·
	Kutaragi · PlayStation · Sony ·
«14.03.2006» Words of the Day [overview] »16.03.2006«	

(b) Cluster der *Ord i Dag* am 15. März 2006

Abbildung 24: Übersichtsseiten

auf ein Wort führt zur Detailansicht (siehe folgender Absatz). Zudem sind Navigationselemente zum Sprung auf den neuesten Tag, zum Blättern vor vorangehenden und folgenden Tag und zum Wechsel zwischen den beiden Darstellungen integriert. Während bei (a) auf die Clusterüberschrift eine einfache alphabetisch sortierte Liste der Clustermitglieder folgt und nur diejenigen Kandidaten angezeigt werden, welche in einem Cluster mit mehr als zwei Mitgliedern sind, ist die Liste bei (b) durch typographische Mittel weiter strukturiert. Die Darstellung ist an Nutzenden von Web 2.0 Seiten wie zum Beispiel flickr oder last.fm vertraute *Tag Clouds* (Smith, 2008) angelehnt.

- Die Größe des Zeichensatzes gibt an, wie übergewichtet im Vergleich mit den Referenzkorpus das Wort ist, ein Maß das Ahmad u. a. (2000) *weirdness-index* nennen; die Darstellung wird wegen der Verteilung der Meßwerte logarithmiert und aus Gründen der Übersichtlichkeit auf den Bereich 50% – 200% der Standardschriftgröße skaliert.
- Die Farbigkeit kodiert, wie sicher die Zuordnung des Klassifikators zu der angezeigten Kategorie ist. Der Vorteil dieser Darstellung ist, dass die besonders auffälligen und eindeutigen Begriffe besonders deutlich sichtbar sind, während weniger auffällige und unklare erst beim zweiten Hinsehen erfasst werden.
- Indem die alphabetische Liste erhalten bleibt, besteht weiterhin die Möglichkeit systematisch nach interessierenden Wörtern zu suchen.

DETAILANSICHT FÜR EIN WORT

Die Detailansicht in zeigt auf einer HTML-Seite alle Information zu einem Wort an einem konkreten Tag. In Abbildung 25 ist die Detailansicht für den norwegischen Begriff *fugleinfluenza* (Vogelgrippe) wiedergeben. Teil (a) zeigt eine Navigationsleiste, über welche Benutzende auf andere Bereiche der Website, auf Erklärungen zu den Wörtern des Tages und den Belegstellen sowie auf die Übersichtsseite zu dem ausgewählten Tag zugreifen können. Darunter folgt eine Liste mit Beispielsätzen aus der tagesaktuellen Presse zu dem Wort und allen seinen Wortformen. Soweit vorhanden ist der fett hervorgehobene Suchbegriff auch Ankertext für einen Link zurück zum Originalartikel bei der jeweiligen Quelle.

Die übrigen beiden Elemente der Seite, ein Assoziationsgraph (siehe Abbildung 25 (b)) und ein kombinierter Verlaufs- und

Wortschatz: Wörter des Tages: Belegstellen für »fugleinfluensa« am 15.03.2006

1. Ungarske forskere har utviklet en vaksine som kan beskytte mennesker mot den aggressive varianten av **fugleinfluensa** , H5N1 .
2. Beredskapssjefen sier at det nå skal jobbes for at **fugleinfluensaen** ikke skal spres , skriver Jyllands-Posten .
3. Personer som skal inn eller reise fra gårder der det er fjørfehold , skal foreta nødvendige beskyttelsestiltak for å forhindre spredning av **fugleinfluensa** .
4. Stockholm (NTB-AFP-TT) : Et EU-laboratorium bekrefter at den farligste formen for **fugleinfluensa** finnes i Sverige , opplyser svenske landbruksmyndigheter .
5. Eller har vi vært for opptatt av **fugleinfluensa** , aids , tuberkulose og malaria i Afrika ?
6. Det er første gang at **fugleinfluensa** blir påvist i Danmark , men viruset H5N1 er tidligere påvist i både Sverige og Tyskland .
7. I alle vestlige land der **fugleinfluensaen** er konstatert , har det vært tilløp til panikk .
8. - Det er en liten teoretisk mulighet for at **fugleinfluensaen** kan bli en humanitær utfordring , men pandemispøkelset , altså risikoen for at det kommer smitte fra menneske til menneske , er på nåværende tidspunkt ren teori .

(a)

Abbildung 25: Detailansicht für *Ord i Dag* „fugleinfluensa“ (norw. *Vogelgrippe*) am 15. März 2006.

Kookkurrenzgraph (siehe Abbildung 25 (c) und Listing B.4) jeweils für das betrachtete Wort, werden im Anschluss erklärt.

4.3.5.2 Grafisches Interface

Der Assoziationsgraph wird von Schmidt (1999) beschrieben. Er zeigt eine Menge von Kookkurrenzen zum jeweiligen Suchbegriff an. Dazu werden alle Kookkurrenzen dem Wert des Kookkurrenzmaßes nach absteigend geordnet und auf eine definierte Anzahl (16, 32, 48, ...) eingegrenzt. Für die so ausgewählten Begriffe als Knoten werden alle Kookkurrenzen als Kanten eingezeichnet, falls der Knoten in mindestens einem Dreieck vorkommt. Die Kantenstärke visualisiert den Wert des Signifikanzmaßes. Das Graphlayout geschieht mit Hilfe einer *simulated annealing*-Implementierung, wobei der Suchbegriff in der Mitte des Graphen fixiert bleibt. Benutzende erhalten somit einen schnellen Überblick der assoziierten Terme.

Der kombinierte Verlaufs- und Kookkurrenzgraph zeigt auf der rechten Seite oben (1) das Startwort. Zu diesem werden bis zu acht Kookkurrenzen aus der durch (2) hervorgehobenen Zeitscheibe ausgewählt, so dass einerseits der Wert des Kookkurrenzmaßes besonders hoch ist und dass sich andererseits aus Gründen der Darstellbarkeit die Frequenz der Begriffe maximal um drei Zweierpotenzen von der des Startworts unterscheidet.

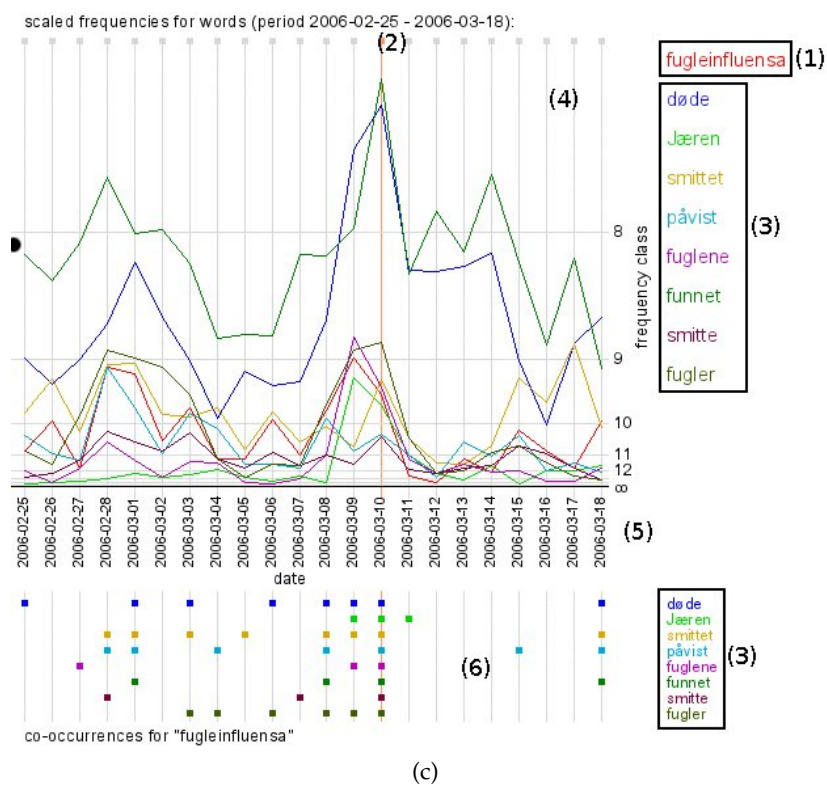
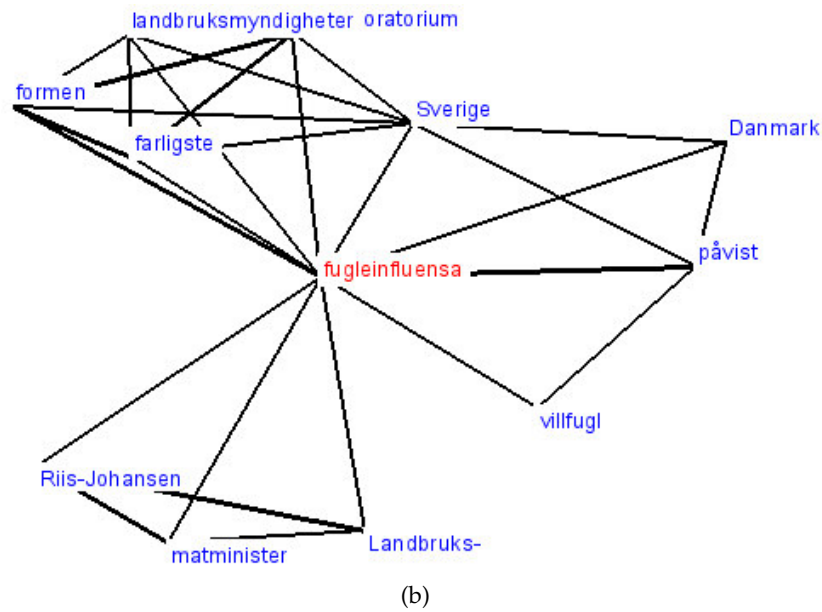


Abbildung 25: Detailansicht für Ord i dag „fugleinfluensa“ (dt. *Vogelgrippe*) am 15. März 2006.

Diese Begriffe werden in (3) angezeigt. Im oberen Bereich (4) wird ein Verlaufsgraph entlang der Zeitleiste (5) eingezeichnet. Dieser ist logarithmisch skaliert und bezüglich der durchschnittlichen Zeitscheibengröße normiert. Im unteren Bereich (6) wird für jede der ausgewählten Kookkurrenzen angezeigt, ob sie auch in dem jeweiligen Zeitscheibenkorpus vorhanden ist. War ein Wort in einer Zeitscheibe als Wort des Tages ausgewählt, so wird dies im Verlaufsgraphen durch ein Rechteck hervorgehoben hinter welchem ein Link auf die jeweilige Detailansicht steckt. Eine interaktive Version des Graphen ermöglicht es zudem auf die bei (3) angezeigten Wörter und die Elemente der Zeitleiste zu klicken, um das fokussierte Wort bzw. das Datum, von welchem die Kookkurrenzen ausgewählt werden, zu ändern.

4.3.5.3 *Maschinenlesbarer Output*

Neben den üblichen HTML-Inhalten, erfreuen sich Newsfeeds in den XML-Formaten RSS bzw. Atom im Netz großer Beliebtheit. Zur Benutzung mit einem Newsreader und zur Content Syndication werden für alle Kategorien RSS-2.0-Feeds angeboten. Diese enthalten für den aktuellen Tag die Liste der Wörter des Tages und Links auf die Seiten in der Webschnittstelle. Die Feeds sind zum einen auf einer Übersichtsseite¹⁴ aktiv verlinkt, zudem wurden sie per <link>-Element auch derart in die Startseite der Wörter des Tages integriert, dass Webbrowser, welche die sogenannte RSS autodiscovery unterstützen, sie dem Benutzer automatisch zum Abonnement vorschlagen können.

Die RSS-Feeds der deutschsprachigen Wörter des Tages¹⁵ werden so von Langenscheidt und dem HanDeDict-Projekt zur automatischen Generierung einer sinnvoll groß bemessenen Liste aktueller Lernwortvorschläge für Sprachlernende des Englischen bzw. des Chinesischen genutzt.

4.3.6 *Evaluation*

Zur Evaluation eines solch komplexes Gebildes wie dem hier vorgestellten zählt zweierlei: eine Evaluation der einzelnen Verfahren und eine Evaluation des Konzepts. Um Verfahren des Information Retrieval und verwandter Disziplinen zu evaluieren gibt es in Lehrbüchern wie Baeza-Yates und Ribeiro-Neto (1999) eine Reihe von Standardmaßen, insbesondere *Precision*, *Recall* so-

¹⁴ So z.B. unter <http://wortschatz.uni-leipzig.de/wdtno/RSS/> für die *Ord i Dag*.

¹⁵ Verfügbar unter: <http://wortschatz.uni-leipzig.de/wort-des-tages/>

wie deren harmonisches Mittel, das *F-Measure* oder *Average Precision*-basierte.

Der Klassifikator in Unter-Unterabschnitt 4.3.4.2 wurde zum Beispiel durch Tests mit zehnfacher Kreuzvalidierung soweit optimiert, dass reproduzierbar eine Precision von 90% bei einem Recall von 50% erreicht wurde. Zur Auswahl der Wörter in Unter-Unterabschnitt 4.3.4.1 wurden die Rankings mehrerer Signifikanzmaße (Poisson, log-likelihood, weirdness-index) verglichen ohne dass ein zu großer Unterschied festgestellt werden konnte.

Ziel dieser Arbeit war es aber explizit nicht, für einzelne Komponenten optimale Lösungen zu finden oder mit Verspätung an einer TREC-Task teilzunehmen und somit unterbleibt eine systematische Evaluation. Ebenso unterbleibt auch eine umfassende Nutzerstudie, die nötig wäre, um das Gesamtkonzept in einer Anwendungsumgebung zu evaluieren, weil es eine solche Anwendungsumgebung (noch) nicht gibt. Vielmehr geht es um eine Präsentation dessen, was möglich wäre, bzw. was im Lauf der vergangenen Zeit auch bereits üblich wurde und was in Zukunft nicht nur für die Medienbranche sondern im Kontext aller eHumanities essentiell sein wird, um Analysierenden besseren Zugriff auf ihre Daten zu gewähren, oder wie es Ayres (2007) formuliert: „The future belongs to the Super Cruncher who can work back and forth and back again between his intuitions and numbers.“

4.4 WEITERENTWICKLUNGEN UND PERSPEKTIVEN

Im folgenden Abschnitt werden Weiterentwicklungen aus den Wörtern des Tages (Unterabschnitt 4.4.1 und Unterabschnitt 4.4.2) sowie Anwendungsfälle (Unterabschnitt 4.4.3) beschrieben.

4.4.1 Anwendungen

Auf der Grundlage der Dokumente, Daten und Bedürfnisse aus den Wörtern des Tages sind eine Reihe von Weiterentwicklungen entstanden, die im Folgenden kurz vorgestellt werden. In Anhang A sind einige illustrierte Anwendungen zu finden.

4.4.1.1 Pressebarometer

Anlässlich der Bundestagswahl 2005 wurden im Vorfeld aus aggregierten Frequenzdaten der Wörter des Tages Charts wie in Abbildung 26 erzeugt und mit der Sonntagsfrage *Wenn am nächsten Sonntag Bundestagswahlen wären ... ?* des Meinungsforschungs-

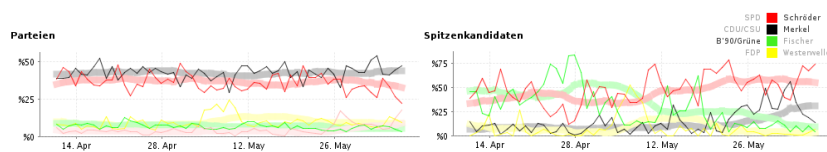


Abbildung 26: Pressebarometer aus Wörtern des Tages (Stand: 6. Juni 2005)

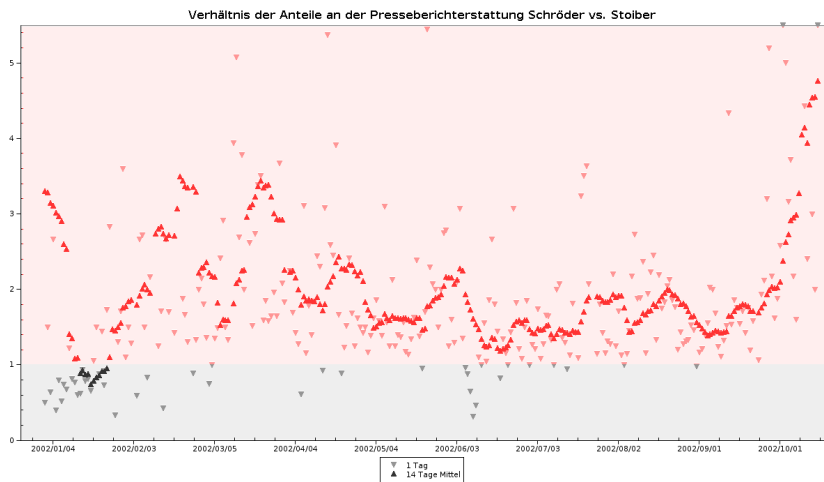


Abbildung 27: Schröder vs. Stoiber (Wahl am: 22. September 2002)

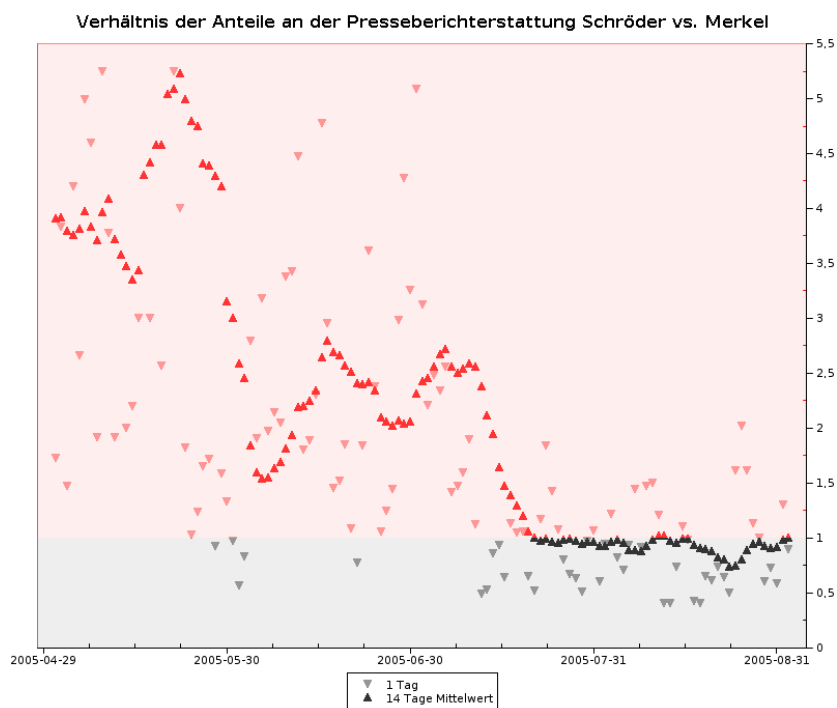


Abbildung 28: Schröder vs. Merkel (Wahl am: 18. September 2005)

instituts Emnid verglichen. Natürlich wäre es hochgradig unseriös einfach aus der puren Zahl der Nennungen auf einen etwaigen Wahlausgang zu schließen. Zu beobachten waren aber dennoch einige interessante Dinge: Die kleineren Parteien wurden im Pressebarometer tendenziell überbewertet. Im Pressebarometer waren grundsätzlich die gleichen Tendenzen wie bei der Sonntagsfrage zu beobachten, wobei die Schwankungen im Pressebarometer auch noch bei Glättung durch ein gleitendes 30-Tage Mittel sehr viel stärker waren. Beim Vergleich der Spitzenkandidaten der Bundestagswahlen 2002 und 2005 im Zeitraum vor der Wahl fiel auf, dass 2002 Gerhard Schröder quasi immer häufiger als Edmund Stoiber genannt wurde. Dank dem so genannten Kanzlerbonus ist dies auch zu erwarten. Angela Merkel konnte aber 2005 bereits einige Wochen vor der Wahl das Verhältnis recht konstant zu ihren Gunsten umkehren – möglicherweise ein Indiz des späteren Wahlausgangs und in Abbildung 27 und Abbildung 28 visualisiert.

4.4.1.2 *WCTAnalyse*

Die Indexierung von und der Zugriff auf Zeitscheibendatenbanken in der beschriebenen Form gestaltet sich im kleineren Rahmen mit aktueller PC-Hardware zufriedenstellend (vgl. Anhang C für das Datenbankschema). Sind sehr umfangreiche Daten, wie etwa die deutschsprachigen Zeitungsdaten, über mehrere Jahre zu bewältigen, zeigt sich aber deutlich die Notwendigkeit über effizientere Datenbereitstellung nachzudenken. (Gottwald u. a., 2007) und ausführlicher (Gottwald, 2007) setzen sich mit diesem Problem auseinander und beschreiben eine auf verketteten Listen basierende erweiterbare Datenstruktur, welche die Probleme der MySQL-Datenbank umgeht und effektiv nur noch durch den verfügbaren Plattenplatz limitiert ist. Zudem implementiert Gottwald (2007) eine Software mit umfangreichen Anzeige- und Analysemethoden für Frequenzen, Kookkurrenzen und Clustern von Wörtern sowie zu so genannten Bedeutungsklassen zusammengefassten Gruppen von Wörtern.

4.4.1.3 *Dokumentenclustering*

Die für die Wörter des Tages verwendeten Texte lassen sich noch mit weiteren Verfahren aufbereiten: Cui (2007) hat die Ausgangsdokumente für die Wörter des Tages mit Hilfe dem Chinese-Whispers-Verfahren (Biemann, 2006) geclustert und erzeugt damit einen Überblick für die Nachrichtencluster des Tages, der

an der Darstellung von Google News orientiert ist. Heine u. a. (2008) wenden die Metapher des Informationsraumes auf in Absätze zerlegte Zeitungsartikel an. Die Absätze und damit die Dokumente werden auch hier mit Chinese Whispers geclustert, mit einem hierarchischen Layout-Algorithmus ausgelegt, mit besonders relevanten Termen beschlagwortet (Witschel, 2005) und dann in einem Mengendiagramm dargestellt.

4.4.1.4 Visualisierung als Treemap

Um die Wörter einer oder aller Kategorien übersichtlich ihrer Gewichtung nach auf begrenztem Raum darzustellen, drängt sich ein Standardverfahren der Visualisierung auf, die Treemap (Shneiderman, 1992). Der zugrunde liegende Algorithmus teilt typischerweise die verfügbare Fläche hierarchisch in Rechtecke mit dem Gewicht der visualisierten Daten entsprechenden Größe ein. Sind diese Größen ähnlich einem *power law* verteilt wie das bei den Wörtern des Tages der Fall ist, ist es sinnvoll, die Werte vor der Übergabe an den Algorithmus logarithmisch zu skalieren, damit nicht zu viele zu kleine und unleserliche Rechtecke entstehen.

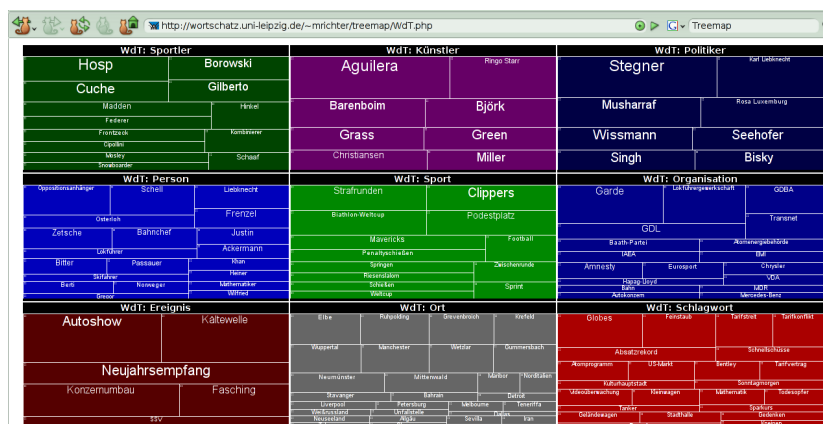
Abbildung 29 (a) zeigt eine exemplarische Implementierung in PHP und HTML der Darstellung der deutschen Wörter des Tages vom 14. Januar 2008 als Treemap. Um allen Kategorien genug Platz einzuräumen wurde auf eine Anordnung der obersten Ebene nach dem Treemap-Algorithmus verzichtet. Eine bekannte Visualisierung mit Hilfe einer Treemap ist die mit Adobe Flash realisierte Google Newsmap¹⁶ in Abbildung 29 (b). Darin sind die Überschriften von Googles News Portal nach den Themen des Portals gruppiert und nach der Anzahl der dazu gefundenen Artikel gewichtet.

Ein Nachteil dieser Darstellung vor der vorher vorgestellten Clustering-Lösung ist, dass die Zusammenhänge zwischen den Wörtern, nicht ersichtlich ist. Dafür ist in jener aber das Gewicht der Absätze und der extrahierten Terme nicht sichtbar, so dass sich eine Kombination der beiden Visualisierungsideen anbietet, analog dazu, wie dies z.B. Fry (2008) beschreibt.

4.4.2 Mashup

Mit dem Web 2.0 sind eine Reihe umfangreicher Datenquellen nicht nur öffentlich verfügbar, sondern als Services konsumier-

¹⁶ <http://marumushi.com/apps/newsmap/newsmap.cfm>, (Zugriff am 14. Januar 2008)



(b) Google Newsmap vom 14. Januar 2008, 17:00 Uhr

Abbildung 29: Treemap-Darstellungen für News.

bar und damit geradezu trivial integrierbar geworden. Insbesondere Geodaten und Landkarten, Fakten, Annotationen, Lexikon-einträge, Buch-, Film- und Musikempfehlungen, aber auch linguistische Daten stehen zur Verfügung. Deren kreative Zusammenführung in einer Benutzerschnittstelle wird auch *Mashup* genannt (Hof, 2005).

4.4.2.1 Grundlagen

Die Daten stammen teils aus zentralen Quellen und teils von Benutzern sogenannter *Social Media*. Im Gegensatz zur formal strengen Lehre des *Semantic Web* (Berners-Lee, 1999) sind letztere weniger verlässlich, sicherlich nicht widerspruchsfrei und weniger komplex strukturiert. Wertlos sind sie aber keineswegs und sie zeichnen sich dadurch aus, dass Nutzer sie gerne und in großem Maßstab generieren. Der einfachen Benutzung zugrunde liegen APIs für Entwickler, welche auf Standards wie RSS, REST, XML-RPC oder SOAP basieren. Durch Tools wie Yahoo Pipes oder Marmite (Wong und Hong, 2007) stellen auf Grundlage diese Schnittstellen auch für Endanwender verständliche und handhabbare Werkzeugkästen bereit.

Während im Geschäftsumfeld die Implementierung komplexer *Service Oriented Architectures* (SOAs) auf Unternehmenslevel gerade erst erfolgt, haben so genannte Mashups bereits die Gunst der Internetnutzer erobert: Mashups werden üblicherweise mit *Asynchronous JavaScript and XML* (AJAX) umgesetzt und laufen im Webbrowser des Benutzers. Sie stellen eine – netztypisch eher untheoretische und nicht-zentralistische – ergebnisorientierte und leichtgewichtige Umsetzung einiger Ideen von SOAs dar:

„Web 2.0 Data mashups are to SOA what NASCAR is to automobile production – a much faster, more free-flowing, results-oriented way of combining complementary components into new applications that have an advantage over traditional months-long application development or production cycles.“

(Trotta, 2006)

„It is widely believed that the key to making the abundant information on the web accessible in a better way than today is to »mash up Web 2.0 and the Semantic Web« (Ankolekar u. a., 2007) – by combining insights from the people-driven, non-authoritarian, bottom-up Web 2.0 with those about knowledge representation, reasoning and interoperability from the Semantic Web initiative.“

(Stefaner, 2007)

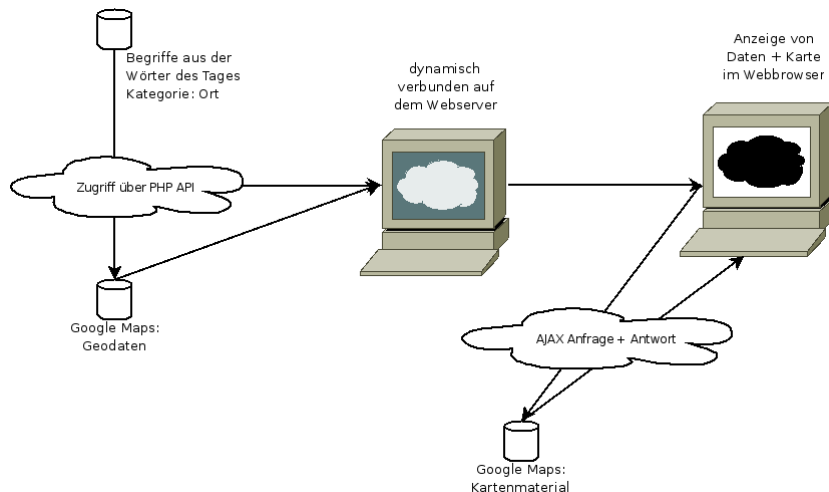


Abbildung 30: Architektur des Mashups von Google Maps und Wörtern des Tages

4.4.2.2 Ein Wörter des Tages-Mashup

In vorigen Abschnitt wurde eine Technologie vorgestellt, welche verspricht auf einfachste Weise die Generierung neuer, datengetriebener Anwendungen zu ermöglichen. Um diesen Anspruch zu testen, wurden die bei den Wörtern des Tages ermittelten Orte auf einer interaktiven Karte angezeigt und dabei an den jeweiligen Orten Popups mit Links auf die Belegstellen auftauchen. Die Architektur der Anwendung skizziert Abbildung 30. Der Beispiel-Code des WdT-Google-Maps-Mashups (siehe Listing B.1, S. 94) ist inklusive Datenbankabfragen und dem umgebenden HTML tatsächlich kaum 30 Zeilen lang und gerade einmal eine handvoll unkomplizierter API-Aufrufe sind für das in Abbildung 31 dargestellte Ergebnis nötig. Die Entwicklungszeit lag bei weniger als einer Stunde inklusive des Einlesens in die knapp gefasste API-Referenz¹⁷. Die These von sehr einfach zu strickenden Anwendungen kann folglich als nicht widerlegt angenommen werden.

Im obigen Beispiel wurden als Wörter des Tages ermittelte Orte angezeigt. Gegeben ein Gazetter (wie er z.B. mit mäßigem Aufwand aus Quellen wie der Wikipedia erstellt werden kann) mit dem die Angabe Ort oder nicht Ort für alle Wörter beantwortet werden kann, sind nahezu ebenso einfach Karten für andere Szenarien denkbar, etwa: „Zeige alle Orte, die mit Wörtern, welche

¹⁷ siehe: <http://www.ajax-info.de/google-maps-api-klasse-in-php> (Zugriff am 14. Juni 2007); die API wurde unter der GNU General Public License veröffentlicht von Philipp Kiszka.

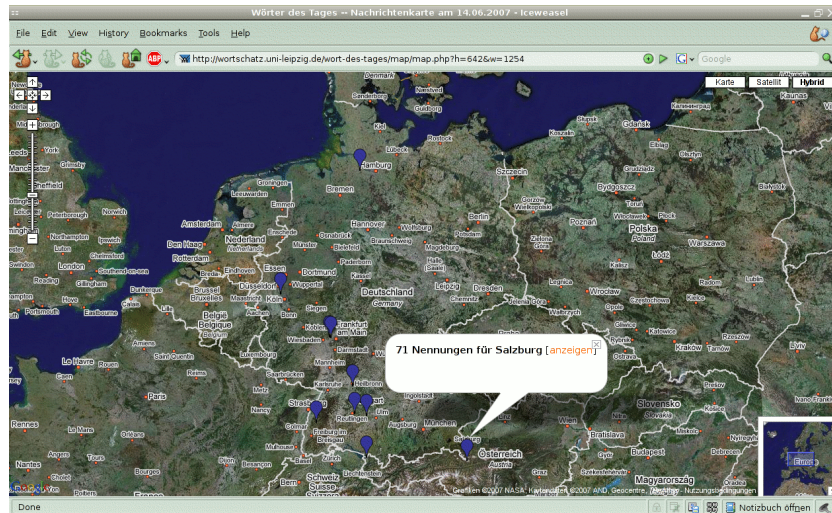


Abbildung 31: Mashup aus Google Maps und Wörtern des Tages: Anzeige im Browser.

mit einem Suchbegriff ko-okkurrieren ebenfalls ko-okkurrieren und zeige die relevanten Belegstellen und zusätzliche Daten an Ort und Stelle an.“ Derartige Darstellungen integrieren dann mehrere Schichten von zusammengehöriger Information z.B. in einem tabMarker. Sie verdichten so das Informationserlebnis auf benutzungs- und zeiteffiziente Weise.

Verglichen mit der statischen Darstellung der Nachrichtensuche von romso.de (Abbildung 32), bei der auf einer Deutschlandkarte Zusammenhänge von Städten und Wörtern markiert werden, ist die Darstellung im „Wörter des Tages“-Mashup flexibler und interaktiv.

4.4.3 Medien- und Trendanalyse

Je nachdem, wo zwischen der eher theoretischen (vgl. z.B. (Merten u. a., 2005)) und der eher praktischen (vgl. z.B. (Besson, 2005)) Sicht auf die Medienresonanzanalyse (siehe Abschnitt 3.1.5) man sich positioniert, und ob man eher an *buzz* oder tieferen Erklärungen und Einsichten interessiert ist, ist das angestrebte Ziel leicht oder nahezu nicht erreichbar. Mit Hilfe der Daten aus den Wörtern des Tages und gegebenenfalls zusätzlichen Verfahren lassen sich aber auf alle Fälle Strategien für Medien- und Resonanzanalysen erproben und insofern die opportunistische Einstellung zur Quellenwahl anstelle einer nach bestimmten objektivierte Kriterien zusammengestellten Stichprobe nicht zu sehr stört, die Ergebnisse auch verwenden.

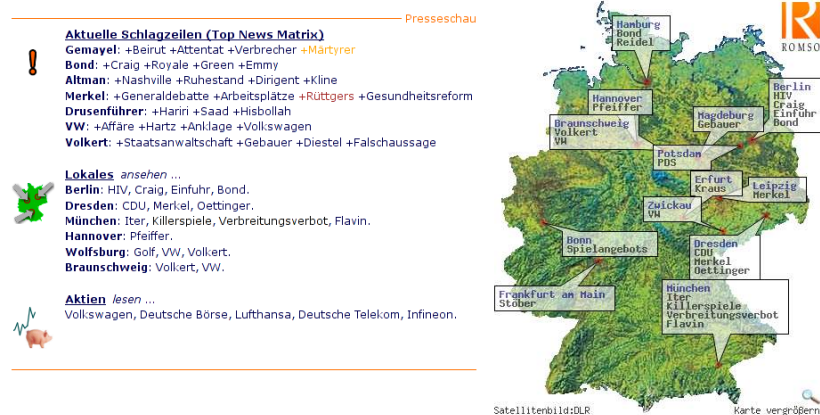


Abbildung 32: Nachrichtenkarte von romso.de für den 22. November 2006, 14:00 Uhr.

4.4.3.1 Zielstellung und Motivation

Die Beobachtung und Vorhersage von Trends hat in den vergangenen Jahren erhebliches kommerzielles Interesse geweckt (vgl. dazu z.B. den Teil III von (Berry, 2003)). Alleine mit dem Zeitscheibenkorpus lassen sich bereits etliche Fragen ohne viel Aufwand beantworten, etwa die folgenden: Wie lange berichten die Medien im Schnitt über ein Ereignis? Wie lange berichten sie über ein spezielles Ereignis? Ist diese Dauer von etwas anhängig und, wenn ja, von was? Gibt es wiederkehrende Themen und, wenn ja, welche? Ist die Wiederkehr periodisch oder lässt sie sich anhand von Indizien vorhersagen?

Die Analysen unterscheiden sich methodisch in einem entscheidenden Aspekt von dem, was man bei typischen Medienbeobachtungsdiensten wie Magenta News in Auftrag geben kann: Anstelle eine vorher definierte Menge von Schlüsselwörtern zu beobachten, kann hier die Sprache der Medien an sich untersucht werden. Die Arbeit eines Medienanalysten wird dadurch in hohem Maße unterstützt, insbesondere die Suche nach möglichen Hypothesen intergiert mit der Möglichkeiten diese live zu überprüfen erinnert sehr an eine These in Ayres (2007): „The future belongs to the Super Cruncher who can work back and forth and back again between his intuitions and numbers.“

4.4.3.2 Veränderung von Begriffen über die Zeit

In Abschnitt 2.3 wurde aufgezeigt, dass über die Gesamtheit der Daten betrachtet, Kookkurrenzen über die Zeit eher nicht bestehen bleiben, dass also Wörter in sich sehr flüssig wandelnden

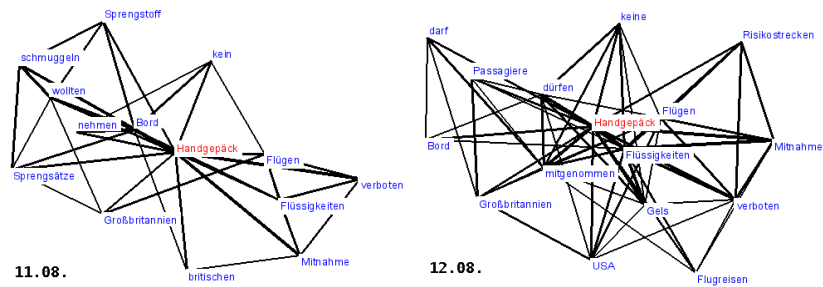


Abbildung 33: Kookkurrenzgraphen zum Begriff »Handgepäck« in der deutschen Presse für den 11. und 12. August 2006.

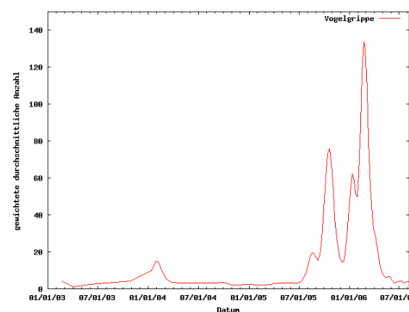
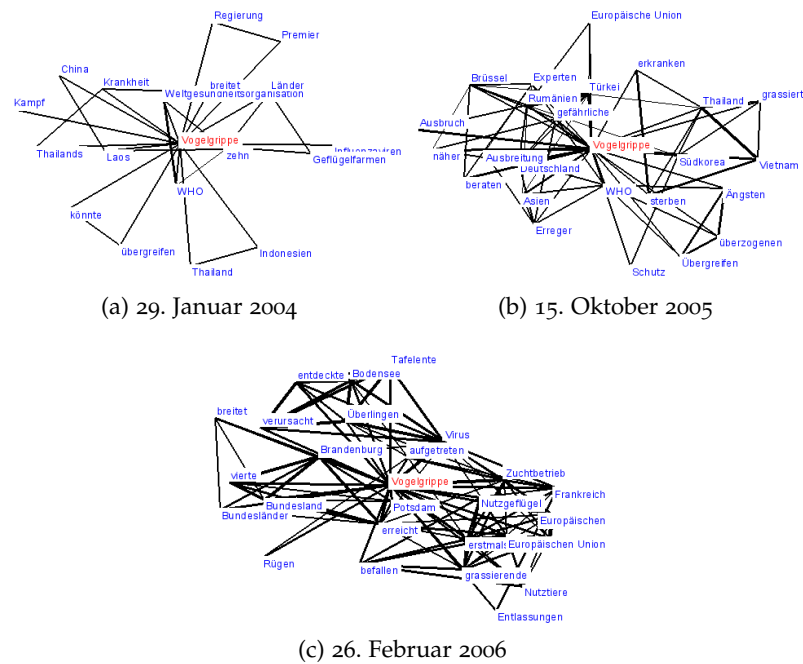


Abbildung 34: Frequenzverlauf des Begriffs »Vogelgrippe« in der deutschen Presse vom 1. Januar 2003 – 1. August 2006.

Kontexten stehen. Dennoch gibt es Kookkurrenzen, die stabil bleiben.

Abbildung 33 illustriert dies in der kurzfristigen Variante am Beispiel *Handgepäck* in der deutschen Presse am 11. und 12. August 2006: Während am 11. noch die Themen versuchter Anschlag und verbotene Mitnahme von Flüssigkeiten im Handgepäck sichtbar sind, hat sich Tags darauf die Berichterstattung auf letzteres eingeschränkt und differenziert es mit *Gels*, *Passagiere*, *Flugreisen*, *Risikostrecken* und *USA* weiter aus.

Über einen längeren Zeitraum betrachtet, lassen sich so durch das Verfolgen veränderter Kookkurrenzen die Wandlungen und Spezifizierungen, welche manche Themen in der Presseberichterstattung erfahren, nachvollziehen. Abbildung 34 zeigt den Frequenzverlauf für den Begriff *Vogelgrippe* im Zeitraum 1. Januar 2003 – 1. August 2006. In den Graphen sind nicht die absoluten Frequenzen für einzelne Tage als Punkte abgetragen, sondern es werden die bezüglich der mittleren Korpusgröße normalisierten Frequenzen mit Hilfe eines Bezierfilters geglättet dargestellt. In Abbildung 35 sind dann die Kookkurrenzgraphen aus den Wörtern des Tages zum Zeitpunkt der drei Peaks am 29. Januar 2004 (a), 15. Oktober 2005 (b) und 26. Februar 2006 (c) aus-

Abbildung 35: Kookkurrenzgraphen für *Vogelgrippe*

gewählt. Während der Verlaufsgraph den Rückschluß auf besonders interessante Tage zulässt, zeigen die Kookkurrenzgraphen in Schnappschüssen, wie in (a) zunächst in Asien eine Krankheit auf Geflügelfarmen auftritt, von der die Weltgesundheitsorganisation glaubt, dass sie auch auf den Menschen übergreifen könnte. Nachdem die Krankheit in (b) schon in Asien grassierte, wird vom Erkranken und Sterben, aber auch von überzogenen Ängsten berichtet. In Europa tagen Experten, um eine Ausbreitung in der Europäischen Union auf dem Weg über die Türkei und Rumänien vielleicht noch zu verhindern. In (c) ist der Erreger dann in vier deutschen Bundesländern angekommen und es werden Details der Ausbreitung in Europa und deren Konsequenzen thematisiert.

4.4.3.3 Vergleich von Kookkurrenzen über Sprachgrenzen

Mit den *Wörtern des Tages* und *Ord i Dag* wird die tagesaktuelle Presse Deutschlands und Norwegens beobachtet. Ein direkter Vergleich der Ergebnisse ist möglich, jedoch nur in den Bereichen sinnvoll, die überregional relevant sind, zum Beispiel also internationale Beziehungen und Großereignisse. Als Beispiel sind in Abbildung 36 die Kookkurrenzen zum Begriff »Handgepäck« für den 11. August 2006 in der norwegischen und deut-

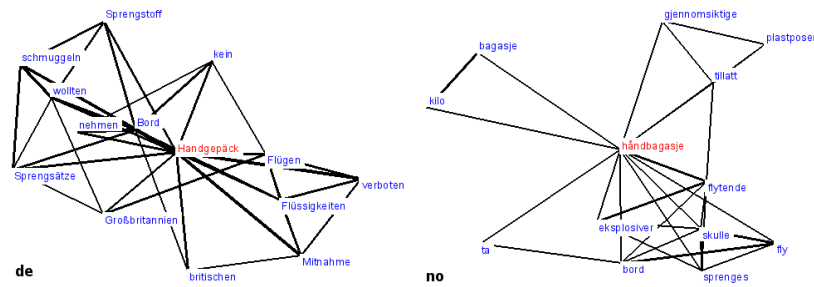


Abbildung 36: Kookkurrenzgraphen zum Begriff »Handgepäck« für den 11. August 2006 in der norwegischen und deutschen Presse.

schen Presse gegenübergestellt. Beide Sichten zeigen grundsätzlich denselben Sachverhalt, nämlich den vereitelten Versuch mit Hilfe flüssiger Sprengstoffe Flugzeuge zum Absturz zu bringen. In der Deutschen Ansicht ist zusätzlich zu sehen, dass der Anschlag in Großbritannien versucht wurde und dass Flüssigkeiten im Handgepäck verboten werden. In der norwegischen Sicht ist davon die Rede, dass durchsichtige (nor. *gjennomsiktige*) Plastiktüten (nor. *plastposer*) gestattet (nor. *tillatt*) seien. Ob dies nun konkret etwas über die Wahrnehmung der Anschläge in den einzelnen Ländern aussagt oder nicht, sei einmal dahingestellt. Im Rahmen einer Studie können solche Fragestellungen aber systematisch operationalisiert und untersucht werden, und die Kookkurrenzen können dabei zu untersuchende Hypothesen nachvollziehbar motivieren sowie modellieren helfen.

4.4.3.4 Ermittlung von Neologismen

Die im Korpus verfügbare Zeitinformation erlaubt es Aufstieg und Fall der Häufigkeiten der Nennungen von Wörtern leicht nachzuvollziehen. Wann ein Wort zuerst benutzt wurde, wie sich die Frequenz und der Gebrauch im Lauf der Zeit gewandelt hat und schließlich, wann ein Wort nicht mehr gebraucht wird, all dies ist unmittelbar ersichtlich. Dies ist nicht nur von allgemeinem Interesse sondern hat auch konkrete Anwendungsfälle, zum Beispiel für diachrone linguistische Untersuchungen: Gegen sei ein Name wie *Google*. Wie lange dauert es, bis die Menschen aus diesem z.B. Verben (*Ich habe mal schnell danach gegoogelt.*) oder Adjektive (*Das sind nur schnell zusammengegoogelte Folien.*) ableiten und gebrauchen? Weiterhin kann es auch eine praktische Anwendung einer zeitgenauen Datenbank sein, die Erstellung oder Überarbeitung von Wörterbüchern zu unterstützen. Sei es, dass es von Interesse ist, ein Inventar an Wör-

tern auf seine Gebräuchlichkeit zu untersuchen und gegebenenfalls anzupassen: Dies war zum Beispiel bei der in Richter (2004) untersuchten Neuauflage des Dornseiff-Wörterbuchs (Dornseiff, 2004) der Fall. Oder sei es, dass ein Neologismenwörterbuch wie (Quasthoff, 2007) entsteht, welches zum Teil auf der Grundlage von im Kontext der Wörter des Tages gesammelten Daten implementiert wurde. Auch die in den vergangenen Unterkapiteln vorgestellten Medienanalysetechniken können für die Wörterbuchproduktion gewinnbringend angewandt werden, zum Beispiel wenn es um den veränderten Gebrauch von Wörtern oder sich wandelnde Wendungen geht. Die chronologische Datenbank ermöglicht in all diesen Fällen Analysierenden einen schnellen Zugriff auf relevante Daten, kann anhand operationalisierter Muster den Blick lenken, Vorschläge für Kandidaten machen und dies mit Belegen unterfüttern. Die Vorarbeit der Operationalisierung und die Aufarbeitung des Materials obliegt freilich der Kompetenz der Analysierenden.

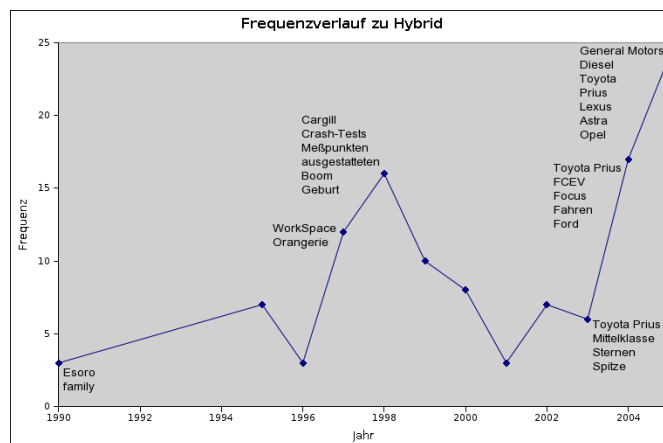
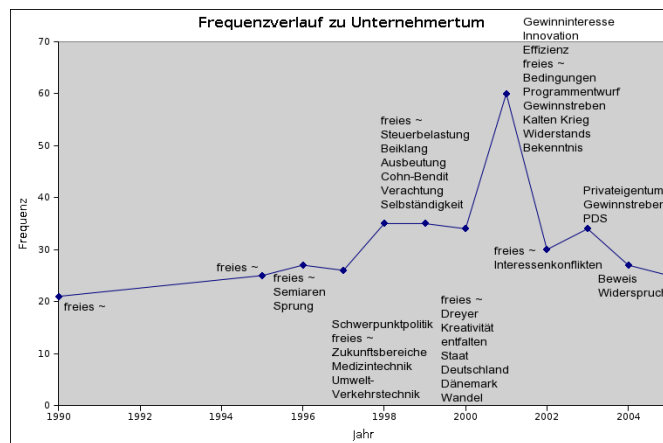
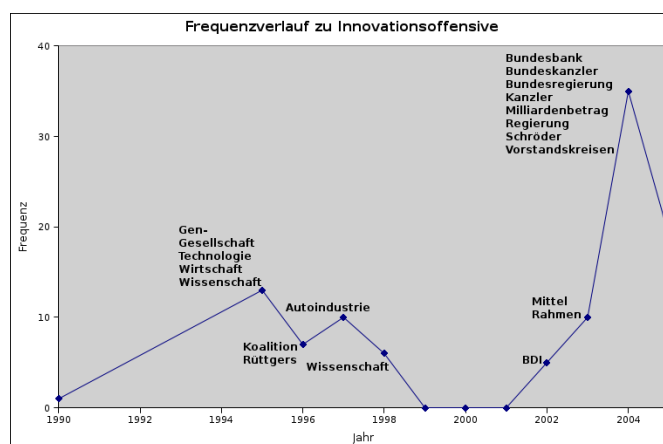
SCHLUSS

Über Inhaltsanalyse, Text Mining und Visualisierung ist ein Apparat nicht nur zum Nacherzählen bebildeter Geschichten sondern auch zur explorativen Datenanalyse entstanden, der mit wenig Aufwand für viele Sprachen einsetzbar ist. Intention war es dabei nie, spezielle Verfahren zu optimieren, sondern vielmehr prototypisch zu beleuchten, wie sich durch die Integration von datengetriebenen Verfahren und fachspezifischen Forschungsinteressen im Bereich einer eHumanity wie der Medienanalyse neue, Gewinn bringende Herangehensweisen an altbekannte Probleme schaffen lassen. Wie die zahlreichen Angebote (wie z.B. die in Kapitel 3 genannten), die zu dem Thema entstanden sind, zeigen, besteht hieran heute massiv Bedarf. Gleichzeitig bedeutet die Angebotsvielfalt aber auch, dass sich der Intellekt schulen muss, um zu einem vorliegenden Problem das adäquate Werkzeug auszuwählen, welches zudem ein Maximum an Eigentransparenz gewährleistet. Denn bei aller Komplexitätsreduktion, die Analysen nun einmal mit sich bringen, hat doch immer ein menschliches Auge wach zu bleiben für die zugrunde liegenden Selektionen und Strukturierungen, die nicht etwa in einer *Black Box* versteckt, sondern im Sinne einer Spezifikation transparent und verfügbar gemacht werden sollen.

WEITERE BEISPIELE AUS DER ANWENDUNG

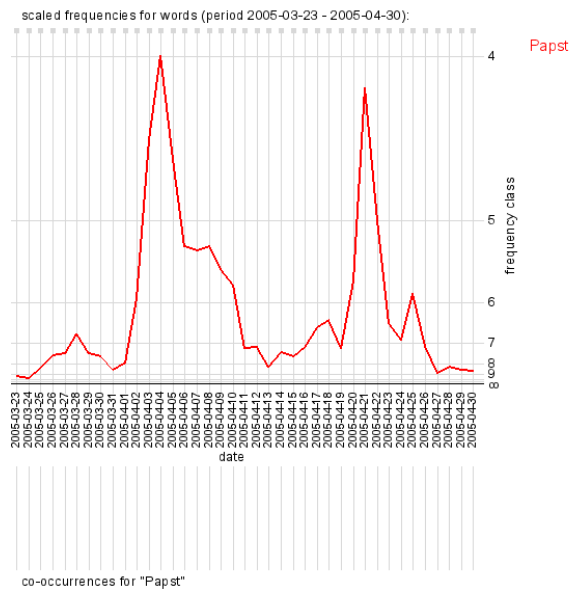
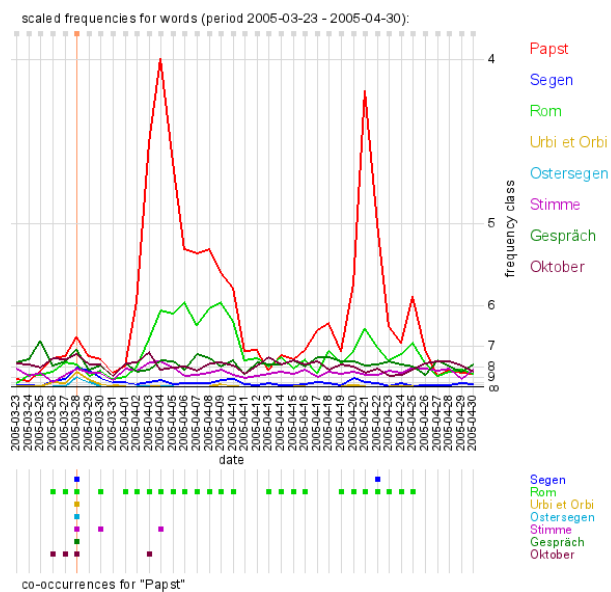
A.1 WIRTSCHAFT

Mit Jahresscheiben statt Tagesdaten wurden für etliche Begriffe Kookkurrenzen auf Frequenzgraphen annotiert, um die Dimension der Vernetzung unmittelbar mit der veränderten Relevanz zusammenzustellen. Beispielhaft seien Abbildung 37, Abbildung 38 und Abbildung 39 zu den Begriffen *hybrid*, *Unternehmertum* und *Innovationsoffensive* herausgegriffen. Die Frequenzangaben sind in diesem Beispiel absolut skaliert, da die Daten auf Korpora normierter Größe (2 Millionen Sätze pro Jahr) erhoben wurden.

Abbildung 37: Annotierter Verlaufsgraph zum Begriff *hybrid*Abbildung 38: Annotierter Verlaufsgraph zum Begriff *Unternehmertum*Abbildung 39: Annotierter Verlaufsgraph zum Begriff *Innovationsoffensive*

A.2 PAPST: TOD UND NEUWAHL

Das Beispiel vom Tode Papst Johannes Paul II. und der Wahl von Joseph Ratzinger zum Papst Benedikt XVI. illustriert das Vorgehen beim Nachvollziehen von Geschehnissen: Die bloße Frequenzübersicht zum Begriff Papst in Abbildung 40 zeigt zwei besonders herausgehobene Tage und einige weitere Peaks, enthält aber noch keine Erklärung. Diese kommt erst durch die Kookkurrenzen zustande. Am 28. März 2005 (Abbildung 41) versagt die Stimme des Papstes beim Ostersegen. Am 4. April 2005 blicken die Kirche, die Welt und insbesondere Polen nach Rom (Abbildung 42), der Johannes Paul II. stirbt und am 5. April beherrscht die Berichterstattung über sein Leben und seinen Tod (Abbildung 43) die Presse. Die Menschen weltweit nehmen Abschied und am 8. April wird das Testament verkündet (Abbildung 44). Zehn Tage später steht das Konklave zur Wahl des Nachfolgers vor der Tür (Abbildung 45). Der Deutsche Kardinal Joseph Ratzinger wird schließlich gewählt (Abbildung 46) und nimmt den Namen Benedikt XVI. an. Am 25. April spendet er den ersten Segen auf dem Petersplatz (Abbildung 47). Ein Umschwenken des Interessenfokus von Papst auf Benedikt XVI. am 21. April zeigt, wer Joseph Ratzinger vorher war (Abbildung 48 und Abbildung 49). Gewählt wird in dieser Zeit freilich nicht nur ein neues Papst, wie Abbildung 50 in zeitlicher Ordnung vor Augen führt.

Abbildung 40: *Papst*: FrequenzverlaufAbbildung 41: *Papst* mit Kookkurrenzen vom 28. März 2005: Die Stimme versagt beim Ostersegen

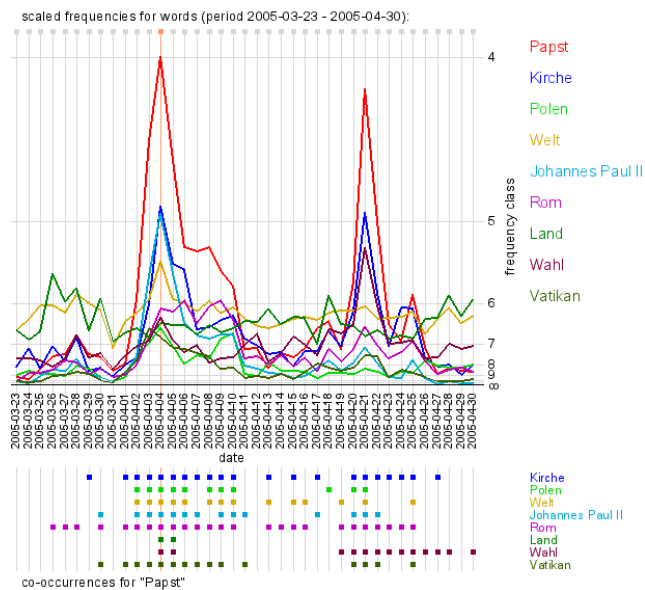


Abbildung 42: *Papst* mit Kookkurrenzen vom 4. April 2005: Die Kirche, Polen, die ganze Welt blickt nach Rom

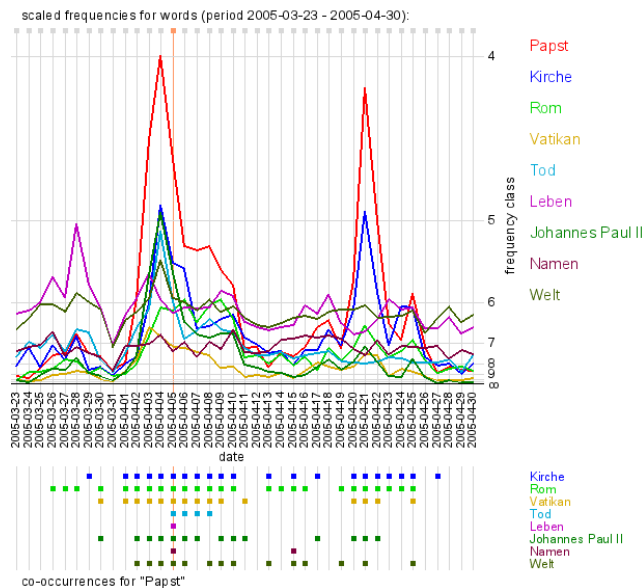


Abbildung 43: *Papst* mit Kookkurrenzen vom 5. April 2005: Papst Johannes Paul II. ist gestorben

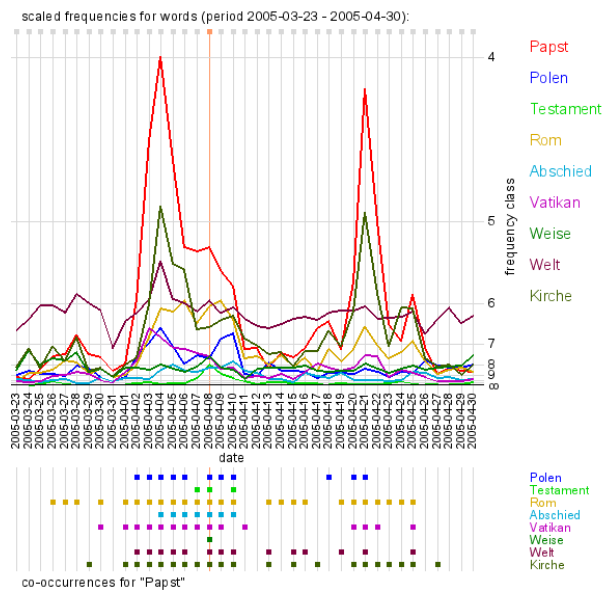


Abbildung 44: *Papst* mit Kookkurrenzen vom 8. April 2005: Testamentsverkündung und Abschied

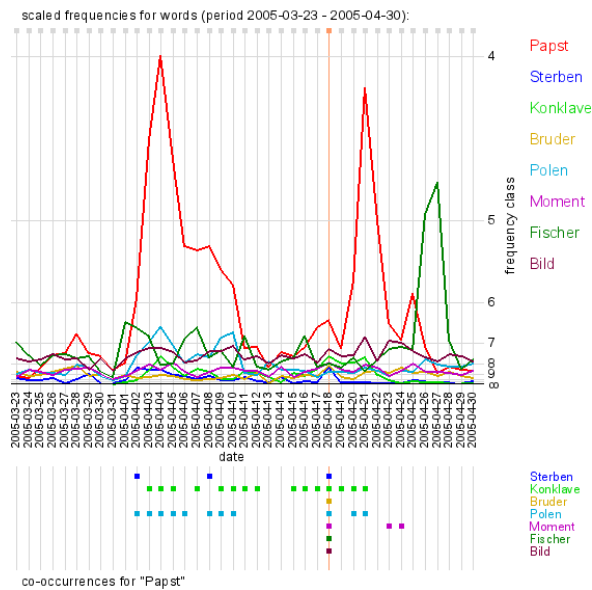


Abbildung 45: *Papst* mit Kookkurrenzen vom 18. April 2005: Das Konklave steht bevor

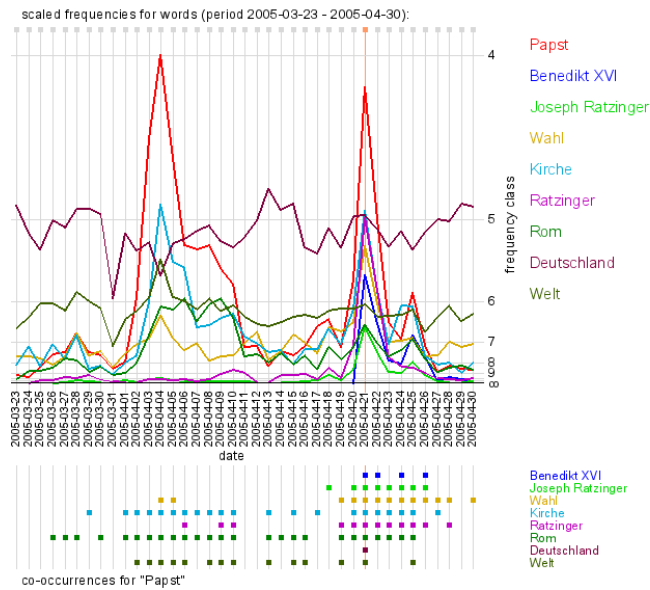


Abbildung 46: *Papst* mit Kookkurrenzen vom 21. April 2005: Der Deutsche Joseph Ratzinger wurde zum Papst gewählt und trägt nun den Namen Benedikt XVI.

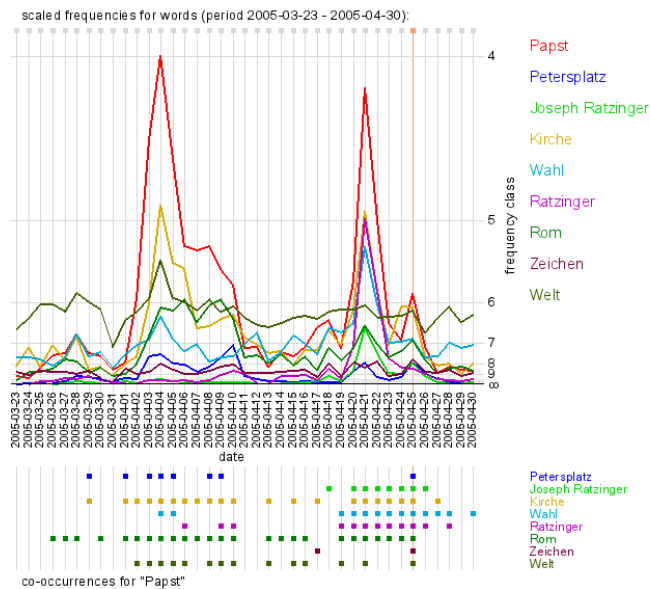


Abbildung 47: *Papst* mit Kookkurrenzen vom 25. April 2005: Der erste Segen auf dem Petersplatz durch Joseph Ratzinger

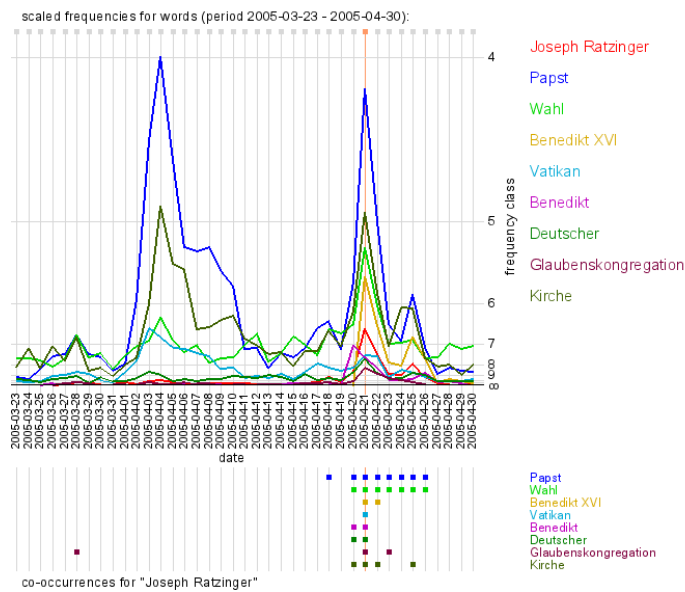


Abbildung 48: *Joseph Ratzinger* mit Kookkurrenzen vom 21. April 2005:
Ratzinger stand vorher an der Spitze der Glaubenskongregation der katholischen Kirche

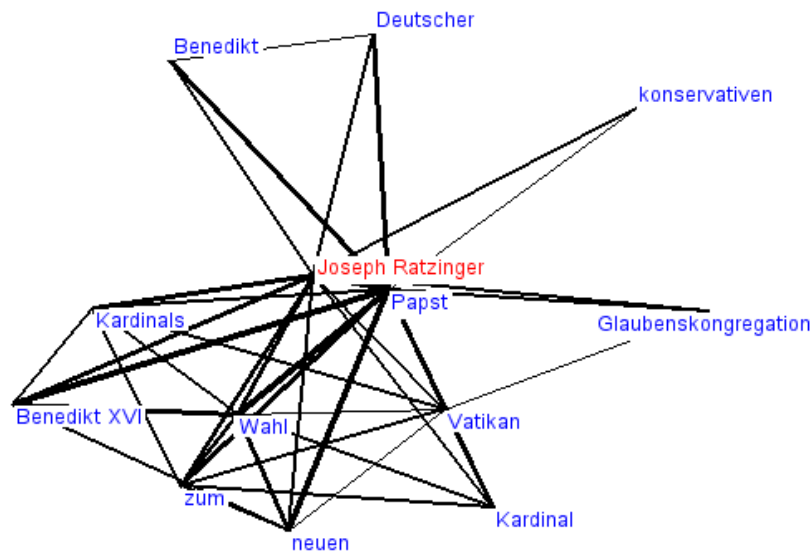


Abbildung 49: Kookkurrenzgraph für *Joseph Ratzinger* am 21. April 2005

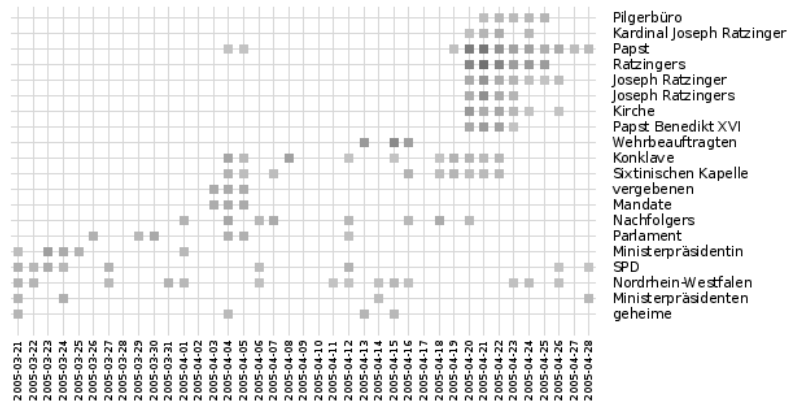


Abbildung 50: Auflösung des Begriffes *Wahl*: Nicht nur ein Papst wurde im März/April 2005 gewählt. Der Grauwert kodiert die Stärke der Kookkurrenz.

A.3 WELTSICHERHEITSRAT

Abbildung 51 ist in Kooperation mit Enrico Rose für die Darstellung am Bildschirm und Scrollen von links nach rechts bei maximal 10000 Pixeln Breite entworfen worden, deswegen ist hier nur ein Überblick, nicht aber Details zu erkennen. Dargestellt sind von oben nach unten die drei Types *Weltsicherheitsrat*, *Iran* und *Irak* sowie von links nach rechts deren gemeinsame Kookkurrenzen über die Zeit. Unmittelbar wird der Schwenk der Aufmerksamkeit der Weltbedrohungslage vom Irak auf den Iran (zwischen Mitte und Ende 2003) deutlich.

Da diese Art der Darstellung im Detail relativ komplex und unverständlich wirkte, wurde sie nicht weiter verfolgt. Im Rahmen des DFG-Forschungsprogramms über Visual Analytics werden in später Fortführung dieser ersten Gehversuche aber künftig weitere, bessere Visualisierungen untersucht.

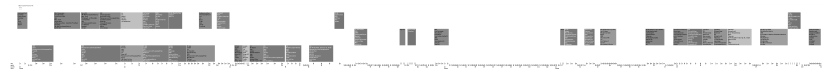


Abbildung 51: Wechsel des Fokus vom Iran zum Irak im Weltsicherheitsrat

LISTINGS

Listing B.1: "News-Mashup (PHP)"

```

1 <html>
  <head>
    <title>Wörter des Tages -- Nachrichtenkarte am <?php echo
      date("d.m.Y"); ?></title>
    <?php
      include("res/class.gmapper.php");
6 $key = "[API KEY]";
      $karte = new gmap($key);
      $karte->headjs();
      ?>
    </head>
11 <body style="bgcolor: #ooo; padding: opx; margin: opx;"
    onunload="GUnload()">
    <?php
      $karte->mapdiv(intval($_GET["h"]), intval($_GET["w"]));
      $karte->bodyjs();
      $karte->map(6, '51.25', '11.5', "hybrid", 1, 17, "large", 1);
16 $db=mysql_connect("woclu4", "[USER]", "[PASS]");
      @mysql_select_db("wdtaktuell", $db);
      $q="select distinct b.wort_bin,b.anzahl from boerse b,
        wort_kategorie k where b.datum > now() - interval 1 day
        and b.wort_bin=k.wort_bin and k.kat_nr=8 and b.wdt='yes'";
      $a=mysql_query($q,$db);
      if($r=mysql_fetch_array($a)){
21 $karte->markstart();
        do{
          list($l,$b)=$karte->getGeoPoint("$r[o]");
          $karte->otherMarker($l,$b,'<strong>' . $r[1] . ' Nennungen
            für ' . $r[0] . ' </strong> [<a href=".." . date("Y/
              m/d/") . $r[0] . '.html">anzeigen</a>] ', "143.png");
          } while($r=mysql_fetch_array($a));
26 $karte->markend();
        }
      ?>
    </body>
  </html>

```

Der PHP-Code in Listing B.1 erzeugt auf einfachste Weise ein Mashup aus Karte und Geodaten von Google Maps und Orten aus den Wörtern des Tages.

Listing B.2: "RSS-Feed Erzeugung (Perl)"

```

#!/usr/bin/perl
# Erzeuge RSS-Feed zu $ARGV[0] aus deutschen Google News
# Benötigte Perl-Module: LWP, Date::Calc, HTML::LinkExtor
# Benutzung: ./createrss.pl $site.de > $site.rss
5 # $site sei ein sinnvoller URL-Bestandteil, z.B. "taz.de"
my $search=$ARGV[0];
use Time::localtime;
use Date::Calc qw(Add_Delta_Days);
use LWP;
10 use HTML::LinkExtor;
# Datum von gestern und heute bestimmen
my $t=localtime;
my ($maxd,$maxm,$y)=($t->mday,1+$t->mon,1900+$t->year);
my ($mind,$minm,$y2)=Add_Delta_Days($y,$maxm,$maxd,-1);
15 # Browser initialisieren
my $browser=LWP::UserAgent->new('agent','Mozilla/5.0 (X11; U;
    Linux i686; en-US; rv:1.8.1.3)');
# Ursprungsurl mit 100 Ergebnissen zu site=$ARGV[0]
my $url="http://news.google.de/news?hl=de&ned=de&q=site:
    $search&ie=UTF-8&as_drrb=q&as_qdr=d&as_mind=$mind&as_minm
    =$minm&as_maxd=$maxd&as_maxm=$maxm&sa=N&num=100";
# Callback-Funktion sammelt Links und zu prüfende Seiten
20 my %links = ();
my %urls = ();
sub cb {
    my($tag,$key,$value) = @_;
    if($tag eq "a" && $key eq "href"){
25         if($value=~/$search\/){
            $links{"$value"}=1;
        }
        if($value=~/http:\/\/news.google.de\/news/ && $value=~/*
            start=\/){
            $urls{"$value"}=1;
30         }
    }
}
# Frage News-Suchseite ab
$p=HTML::LinkExtor->new(\&cb);
35 $res=$browser->request(HTTP::Request->new(GET => $url), sub{
    $p->parse($_[0])});
while(($k,$v)=each(%urls)){
    $p=HTML::LinkExtor->new(\&cb);
    $res=$browser->request(HTTP::Request->new(GET => "$k"),
        sub{$p->parse($_[0])});
}
40 # Schreibe minimales RSS nach STDOUT
print "<?xml version=\"1.0\" encoding=\"UTF-8\"?>
<rss version=\"2.0\">
    <channel>

```

```

45     <title>Google-Newsfeed zu $search</title>
    <link>http://$search</link>\n";
my @out = keys %links;
for($i=0;$i<=$#out;$i++){
    print "<item>
        <title>$out[$i]</title>
        <link>$out[$i]</link>
        <description>$out[$i]</description>
    </item>
    ";
}
55 print "    </channel>
</rss>\n";

```

Der Perl-Code in Listing B.2 erzeugt einen minimalen RSS-Feed von maximal 1 000 in den vergangenen 24 Stunden neuen oder erneuerten Artikel für eine von Google News erfasste Quelle.

Listing B.3: "Klassifikator (PHP)"

```

<?php
$user="wdtno";
$pass="*****";
4 $host="pcaio55";
$dbname="wdtno";
$basedir="/var/www/wdtno";

function date_check($date){
9     if(strlen($date) != 8){
        die("Datum muss YYYYMMDD sein!\n");
    }
    $Y=substr($date,0,4);
    $M=substr($date,4,2);
14    $D=substr($date,6,2);
    return array($Y,$M,$D);
}

class textclassifier {
29     function connect_db($host,$user,$pass){
        $this->db=mysql_connect($host,$user,$pass);
    }

    function getModelString($model){
24        if($model==4){
            return "norwegian newspaper: ";
        }
    }

29    function switchToModelDB($model){
        switch($model){
        case 4:
            mysql_select_db("wdtno",$this->db);
            break;
34        }
    }

    function prepareSentence($sentence){
        $sentence=str_replace("\r","",$sentence);
39        $find = array(".", ":", " ", " ", " ", "!", "?", "\\", '
', '
', '»', '«');
        $replace = array(" ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ");
        $sentence=str_replace($find,$replace,$sentence);
        return "\"" . str_replace(" ", "\",\"",$sentence) . "\""
        ;
    }

44    function storeClassification($model,$tag,$weight){
        $this->classification["k$model"][$tag] += round(

```

```

        $weight,3);
    }

49 function getClassificationSum($model,$num){
    $out=$this->getModelString($model);
    arsort($this->classification["k$model"]);
    $i=0;
    foreach($this->classification["k$model"] as $key =>
        $value){
54         $i++;
        if($i<$num){
            $out .= " $key ($value);\t";
        } elseif($i==$num) {
59             $out .= "$key ($value);\n";
        }
    }
    return $out;
}

64 function getClassificationSumAsArray($model){
    $out=array();
    arsort($this->classification["k$model"]);
    foreach($this->classification["k$model"] as $key =>
        $value){
69         $out[$key]=$value;
    }
    return $out;
}

function getNumberedSubjectAreas($model){
74     $out=array();
    $q="select gr_nr,wort_bin from sachgruppe order by
        gr_nr;";
    $a=mysql_query($q,$this->db);
    while($r=mysql_fetch_array($a)){
79         $out[$r[0]]="$r[1]";
    }
    return $out;
}

function getClassification($sentence,$model,$num){
84     $out="";
    $len=1+substr_count($sentence,",");
    $num=intval($num);
    $sg=$this->getNumberedSubjectAreas($model);
    $q="select s.sa_nr,sum(abs(s.likelihood)/s.likelihood *
89         sqrt(abs(s.likelihood)/w2.anzahl)) as weight
        from wort_sachgebiet s, refwordlist w2
        where s.wort_nr=w2.wort_nr and w2.qual_neg=0
        and w2.wort_bin in ($sentence)
        group by s.sa_nr having weight > 2 order by weight

```

```

        desc limit $num;";
$a=mysql_query($q,$this->db);
94 $c=0;
    if($r=mysql_fetch_array($a)){
        do{
            if(++$c==$num){ $t="\n";} else { $t=";\t";}
            $val=round($r[1],3);
            99 $this->storeClassification($model,$sg[$r[0]],$val*
                log($len));
            $out .= $sg[$r[0]] . " ($val)$t";
            } while ($r=mysql_fetch_array($a));
        }
        return $out;
    }
104 }

function doSilentClassification($sentence,$model){
    $len=1+substr_count($sentence,",");
    $num=intval($num);
    109 $q="select s.sa_nr,sum(abs(s.likelihood)/s.likelihood *
        sqrt(abs(s.likelihood)/w2.anzahl)) as weight from
        wort_sachgebiet s, refwordlist w2 where s.wort_nr=
        w2.wort_nr and w2.qual_neg=0 and w2.wort_bin in (
        $sentence) group by s.sa_nr having weight > 2 order
        by weight desc;";
    $a=mysql_query($q,$this->db);
    while($r=mysql_fetch_array($a)){
        $val=round($r[1],3);
        $this->storeClassification($model,$r[0],$val*log(
            $len));
    }
    114 return $out;
}

function getDetailedClassification($sentence,$model){
    119 $out="";
    $sg=$this->getNumberedSubjectAreas($model);
    foreach(array_keys($sg) as $key){
        $cla[$key]=array();
    }
    124 $q="select w2.wort_bin as wort, s.sa_nr as sachgebiet,
        s.likelihood,
        w2.anzahl,w2.qual_neg
        from wort_sachgebiet s, refwordlist w2 where s.
        wort_nr=w2.wort_nr
        and w2.wort_bin in ($sentence) order by w2.wort_nr,
        s.sa_nr;";
    $a=mysql_query($q,$this->db);
    129 while($r=mysql_fetch_array($a)){
        $cla[$r[1]]["$r[0]"]=array($r[2],$r[3],$r[4]);
    }
    $words=split("\",\"",preg_replace(array('/^'/','/'$/'),

```

```

        array("", ""), $sentence));
$out .= "<table bgcolor=\"grey\" cellspacing=\"5\"><tr
    bgcolor=\"white\"><th></th>";
134 for($i=0;$i<count($words);$i++){
    $out .= "<th>$words[$i]</th>";
}
$out .= "<th>Summe</th>";
$out .= "</tr>\n";
139 foreach(array_keys($sg) as $key){
    $summe=0;
    $out .= "<tr bgcolor=\"white\">";
    $out .= "<td><b> . $sg[$key] . "</b></td>";
    for($i=0;$i<count($words);$i++){
144         if(isset($cla[$key][$words[$i]])){
            $val=$cla[$key][$words[$i]];
            if($val[2]==0){
                $out .= "<td bgcolor=\"green\" align=\"
                    right\"> . round($val[0]/log(1+$val
                        [1]),3) . "</td>";
                $summe += round($val[0]/log(1+$val[1]),3);
149             } else {
                $out .= "<td bgcolor=\"red\" align=\"right
                    \"> . round($val[0]/log(1+$val[1]),3)
                    . "</td>";
            }
        } else {
            $out .= "<td></td>";
154        }
    }
    $out .= "<td>$summe</td>";
    $out .= "</tr>\n";
}
159 $out .= "</table>\n";
return $out;
}

function classify($sentence,$model,$num){
164     if(strlen($sentence)>1){
        $this->switchToModelDB($model);
        echo "$sentence\n";
        $sentence = $this->prepareSentence($sentence);
        echo $this->getDetailedClassification($sentence,4);
169         echo $this->getModelString(4) . $this->
            getClassification($sentence,4,$num);
        echo "\n";
    }
}

174 function classifyWord($wordnr,$model){
    $q="SELECT s.beispiel from sentences s,
        inv_list_full i,word_baseform b where i.bsp_nr

```

```

        =s.bsp_nr and i.wort_nr=b.wort_nr and b.
        base_wort_nr=$wordnr and i.bsp_nr between " .
        $this->Smin . " and " . $this->Smax . ";";
$a=mysql_query($q,$this->db);
if(0==mysql_num_rows($a)){
    $q="SELECT s.beispiel from sentences s,
        inv_list_full i where i.bsp_nr=s.bsp_nr and
        i.wort_nr=$wordnr and i.bsp_nr between " .
        $this->Smin . " and " . $this->Smax . ";";
179    $a=mysql_query($q,$this->db);
    }
    while($r=mysql_fetch_array($a)){
        if(strlen($r[0])>1){
            $sentence= $this->prepareSentence($r[0]);
184            $ignore=$this->doSilentClassification($sentence,
                $model);
        }
    }
    return $this->getClassificationSumAsArray($model);
}
189 }

list($Y,$M,$D)=date_check($_SERVER["argv"][1]);
$classifier = new textclassifier;
$classifier->connect_db($host,$user,$pass);
194 mysql_select_db($dbname,$classifier->db);
$sg=$classifier->getNumberedSubjectAreas($model);

if($_SERVER["argv"][2] != ""){
    $tmp=split(";",$_SERVER["argv"][2],2);
199 $word=array();
    $q="select wort_nr from wortliste where wort_bin=\"${tmp
        [0]}\"";
    $a=mysql_query($q,$classifier->db);
    if($r=mysql_fetch_array($a)){
        $word[$r[0]]=$tmp[0];
204 }
    } else {
        $word=array();
        $q="select w.wort_nr,w.wort_bin from wortliste w,
            wordcounts c where w.wort_nr=c.wort_nr and c.wdt=\"Y\"
            and c.datum=\"${Y}-${M}-${D}\"";
        $a=mysql_query($q,$classifier->db);
209 while($r=mysql_fetch_array($a)){
            $word[$r[0]]=$r[1];
        }
    }
}

214 $a=mysql_query("SELECT min(bsp_nr) from sentences where date
    =\"${Y}-${M}-${D}\";", $classifier->db);
if($r=mysql_fetch_array($a)){

```

```

        $classifier->Smin=$r[0];
    }
    $a=mysql_query("SELECT max(bsp_nr) from sentences where date
        =\"$Y-$M-$D\";", $classifier->db);
219 if($r=mysql_fetch_array($a)){
        $classifier->Smax=$r[0];
    }

    $classifier->store=array();
224 foreach(array_keys($sg) as $key=>$val){
        $classifier->store["$val"]=array();
    }

    foreach($word as $key=>$value){
229 $classifier->classification=array();
        $classifier->classification["k4"]=array();

        $q="SELECT base_wort_nr from word_baseform where wort_nr=
            $key;";
        $a=mysql_query($q, $classifier->db);
234 if($r=mysql_fetch_array($a)) {
            $ret=$classifier->classifyWord($r[0],4);
        } else {
            $ret=$classifier->classifyWord($key,4);
        }
239 list($k,$v)=each($ret);
        $classifier->store[$k][]=array($key,$v);
        unset($classifier->classification);
    }

244 if($_SERVER["argv"][2] != ""){
    print_r($classifier->store);
    } else {
        @mysql_query("DELETE from word_classification where date
            =\"$Y-$M-$D\";", $classifier->db);
        foreach($sg as $key=>$value){
249 if(count($classifier->store[$key])>0){
            echo $sg[$key] . ": ";
            foreach($classifier->store[$key] as $k=>$v){
                $nr=$v[0];
                $str=round($v[1],1);
254 echo "$word[$nr] ($str), ";
                $q="INSERT INTO word_classification VALUES($nr,
                    $key,$str,\"$Y-$M-$D\")";
                @mysql_query($q, $classifier->db);
            }
            echo "\n";
259 }
        }
    }
}
?>

```

Listing B.4: "Frequenz-Kookkurrenz-Diagramm (PHP)"

```

<html>
<head>
3 <meta http-equiv="content-type" content="text/html; charset=
    iso-8859-1">
<base target=_top>
</head>
<body>
<?
8 set_time_limit(60);
  $scriptname=$_SERVER["PHP_SELF"];
  if(isset($_GET["wort"]))
  {
    $url="http://wortschatz.uni-leipzig.de/wdtno";
13 $font="/var/lib/ttf fonts/arialuni.ttf";

    $db=mysql_connect("139.18.2.59","wdtno","*****");
    mysql_select_db("wdtno",$db);

28 $tmp=substr(tempnam("foo","img"),5);
    $im=ImageCreate(770,570);
    $map="<map name=\"$tmp\">\n";
    $white = imagecolorallocate ($im, 255, 255, 255);
    $black = imagecolorallocate ($im, 0, 0, 0);
23 $grey = imagecolorallocate ($im, 216, 216, 216);
    $lgrey = imagecolorallocate ($im, 232, 232, 232);
    $lred= imagecolorallocate ($im, 255, 153, 102);

    $c=array();
28 $c[1]= ImageColorAllocate($im,255,0,0);
    $c[2]= ImageColorAllocate($im,0,0,255);
    $c[3]= ImageColorAllocate($im,0,216,0);

    $c[4]= ImageColorAllocate($im,216,170,0);
33 $c[5]= ImageColorAllocate($im,0,170,216);
    $c[6]= ImageColorAllocate($im,192,0,192);

    $c[7]= ImageColorAllocate($im,0,128,0);
    $c[8]= ImageColorAllocate($im,128,0,62);
38 $c[9]= ImageColorAllocate($im,62,96,0);

    if(isset($_GET["StartOn"])){
        $ST=explode("-",$_GET["StartOn"]);
    } else {
43 $ST=explode("-",date("Y-m-d",mktime(0,0,0,date("m"),date("
        d")-28,date("Y"))));
    }
    $sday="$ST[0]-$ST[1]-$ST[2]";

    if(isset($_GET["EndOn"])){

```

```

48     $ET=explode("-", $_GET["EndOn"]);
    } else {
        $ET=explode("-", date("Y-m-d", mktime(0,0,0,date("m"),date("
            d"),date("Y"))));
    }
    $eday="$ET[0]-$ET[1]-$ET[2]";
53 $seurl="$ET[0]/$ET[1]/$ET[2]";

    if(isset($_GET["FocusOn"])){
        $FT=explode("-", $_GET["FocusOn"]);
        $fday="$FT[0]-$FT[1]-$FT[2]";
58 } else {
        $fday="$ET[0]-$ET[1]-$ET[2]";
        $FT[0]=$ET[0];
        $FT[1]=$ET[1];
        $FT[2]=$ET[2];
63 }

    $stage=round((mktime(0,0,0,$ET[1],$ET[2],$ET[0]) - mktime
        (0,0,0,$ST[1],$ST[2],$ST[0]))/86400);

    if(isset($_GET["woerter"]) && $_GET["woerter"] != ""){
68     $woerter=split(":", $_GET["woerter"]);
    } else if (isset($_GET["wort"]) && $_GET["wort"] != ""){
        $wort=$_GET["wort"];

        $q="CREATE table words$tmp SELECT wort_nr, anzahl from
            wordcounts where datum=\"$fday\";";
73 @mysql_query($q,$db);
        $q="ALTER table words$tmp add index('wort_nr');";
        @mysql_query($q,$db);
        $q="CREATE table coocs$tmp SELECT wort_nr1, wort_nr2,
            signifikanz from kollok_sig_full where date=\"$fday\";";
        @mysql_query($q,$db);
78 $q="ALTER table coocs$tmp add index('wort_nr1');";
        @mysql_query($q,$db);

        $q="SELECT wort_nr from wortliste where wort_bin=\"$wort
            \";";
        $wnr=mysql_result(mysql_query($q,$db),0);
83 $q="SELECT distinct w3.wort_bin from words$tmp w1,
        words$tmp w2, wortliste w3, coocs$tmp c where c.
        wort_nr1=$wnr and c.wort_nr2=w2.wort_nr and c.wort_nr2
        =w3.wort_nr and w1.wort_nr=$wnr and w1.anzahl < 8*w2.
        anzahl and w2.anzahl < 8*w1.anzahl and w3.wort_bin
        not in(\"ta\", \"fâr\", \"inn\") order by c.signifikanz
        desc,w2.anzahl desc limit 8;";

        $a=mysql_query($q,$db);
        if($r=mysql_fetch_array($a)){

```

```

    $wortliste="::$wort::";
88     do {
        $wortliste.="$r[o]::";
    } while ($r=mysql_fetch_array($a));
} else {
    $wortliste="::$wort::";
93 }
$woerter=split(":",$wortliste);
$q="DROP TABLE words$tmp;";
@mysql_query($q,$db);
$q="DROP TABLE coocs$tmp;";
98 @mysql_query($q,$db);
}

// X-Achse (Zeit) geeignet aufteilen
$res=round($tage/28,0);
103 for($i=0; $i<=$tage; $i++)
{
    if($i%$res==0){
        $tag=mktime (0,0,0,$ST[1] , $ST[2] + $i,$ST[0]);
        $durl=date("Y/m/d",$tag);
108        $dstr=date("Y-m-d",$tag);
        $dtag=date("d.m.Y",$tag);
        $x= ($i+0.5)*440/$tage;
        if( $tag == mktime (0,0,0,$FT[1] , $FT[2], $FT[0])) {
            ImageFilledRectangle($im,$x-2,28,$x+2,32,$lred);
            ImageLine($im,$x,30,$x,379,$lred);
113            ImageLine($im,$x,450,$x,549,$lred);
        } else {
            ImageFilledRectangle($im,$x-2,28,$x+2,32,$grey);
            ImageLine($im,$x,30,$x,379,$grey);
118            ImageLine($im,$x,450,$x,549,$grey);
            $map .= "<area shape=\"rect\" title=\"Show graph for
                &quot;$wort&quot; with co-occurences from $dtag
                \" coords=\"\"
                . ($x - 2) . \",28,\" . ($x + 2) . \",32\" href=\"
                    $scriptname?wort=$wort&EndOn=\" . $ET[0] . \"-
                    \"
                . $ET[1] . \"-\" . $ET[2] . \"&StartOn=\" . $ST[0] .
                    \"-\" . $ST[1] . \"-\" . $ST[2]
                . \"&FocusOn=$dstr\"/>\n";
123        }
        ImageTTFText($im,8,90,$x+3,432,$black,$font,date("Y-m-d",
            $tag));
        $x1=round($x-5,0);
        $x2=round($x+5,0);
        $map .= "<area shape=\"rect\" title=\"show words of the
            day from $dtag\" coords=\"$x1,382,$x2,430\" href
            =\"$burl/$durl/\" target=_blank />\n";
128    }
}

```

```

$instring='';
133 $instring.=implode("\","\",$woerter);
$instring='';
$assoc='';
$assoc.=implode("\","\",array_slice($woerter,1));
$assoc='';
138
$q="CREATE table words$tmp SELECT wortliste.wort_bin as
    wort_bin,wordcounts.anzahl as anzahl,wordcounts.datum
    as datum,wordcounts.wdt as wdt from wordcounts,
    wortliste where wordcounts.wort_nr=wortliste.wort_nr
    and wordcounts.datum>=\"$sday\" and wordcounts.datum
    <=\"$eday\" and wortliste.wort_bin IN(\"i\",$instring)
    ";
    @mysql_query($q,$db);

    $q="ALTER table words$tmp add index('wort_bin');";
143 @mysql_query($q,$db);

    $q="CREATE table coo$tmp SELECT w2.wort_bin as wort_bin,c.
    date as datum,c.signifikanz as signifikanz from
    wortliste w1, wortliste w2, kollok_sig_full c where w1
    .wort_bin=\"$wort\" and w2.wort_bin IN($assoc) and c.
    date >= \"$sday\" and c.date <= \"$eday\" and c.
    wort_nr1=w1.wort_nr and c.wort_nr2=w2.wort_nr;";
    @mysql_query($q,$db);
    $q="ALTER table coo$tmp add index('wort_bin');";
148 @mysql_query($q,$db);

    // Faktor bestimmen
    $query="SELECT max(anzahl) from words$tmp where wort_bin IN(
        $instring) and datum <= \"$eday\" and datum >= \"$sday\"";
    $maxzahl=mysql_result(mysql_query($query,$db),0);
153 $query="SELECT max(anzahl) from words$tmp where wort_bin=\"i
        \" and datum <= \"$eday\" and datum >= \"$sday\"";
    $maxder=mysql_result(mysql_query($query,$db),0);
    $factor=abs(round(($maxder/($maxzahl-10)),2));

    // = HKL
158 $outfactor=round(log($factor)/log(2));

    ImageTTFText($im,10, 0, 10, 20,$black,$font,"scaled
        frequencies for words (period $ST[0]–$ST[1]–$ST[2] – $ET
        [0]–$ET[1]–$ET[2]):");
    ImageTTFText($im,10, 0, 10, 565,$black,$font,"co-occurrences
        for &quot;\" . $woerter[1] .&quot;");
    $today=date("d.m.Y");
163
    // Achsenbeschriftungen

```

```

ImageTTFText($im,10,90,485,265,$black,$font,"frequency class"
);
ImageTTFText($im,10,90,470,375,$black,$font,"8");
168 if($outfactor>25) {$outfactor=25;}
    if($outfactor<0) {$outfactor=0;}
    for($i=$outfactor-4;$i<=$outfactor+3;$i++){
        $y=371-(8*$factor)/pow(2,$i-1);
        ImageLine($im,0,$y,455,$y,$grey);
173     if($y<362){
        ImageTTFText($im,10,0,460,$y+4,$black,$font,$outfactor+
            $i);
        }
    }
}
ImageTTFText($im,10,0,200,445,$black,$font,"date");
178
for($i=1;$i<10;$i++){
    $xlast=0; $ylast=0; $started=0;

    if(isset($woerter[$i]) && $woerter[$i]!=""){
183        $wi=$woerter[$i];

        ImageTTFText($im,11,0,500,20+30*$i,$c[$i],$font,$wi);
        $y1=5+30*$i;
        $y2=20+30*$i;
188        $map .= "<area shape=\"rect\" title=\"Show graph
            for &quot;" . urlencode($wi) . "&quot; on $ET
            [2].$ET[1].$ET[0] \" coords=\"500,$y1,750,$y2
            \" href=\"\$scriptname?wort=$wi&EndOn=" . $ET
            [0] . "-" . $ET[1] . "-" . $ET[2] . "&StartOn="
            " . $ST[0] . "-" . $ST[1] . "-" . $ST[2] . "&
            FocusOn=" . $FT[0] . "-" . $FT[1] . "-" . $FT
            [2] . "\" />\n";
        for($j=0;$j<=$tage;$j++) {
            $x=($j+0.5)*440/$tage;
            $thedata=date("Y-m-d",mktime(0,0,0,$ST[1] , $ST
            [2]+($j), $ST[0]));
            $query="SELECT 8*$factor*b1.anzahl/b2.anzahl,
                b1.wdt from words$tmp b1,words$tmp b2 where
                b1.wort_bin=\"$woerter[$i]\" and b2.
                wort_bin=\"$i\" and b1.datum=\"$thedata\"
                and b2.datum=\"$thedata\" and b1.anzahl >
                0;";
193            $answer=mysql_query($query,$db);
            if($result=@mysql_fetch_array($answer))
            {
                $started=1;
                if($j>=1){
198                    $y=34*$result[0];
                    ImageLine($im,$xlast,371-$ylast,$x,371-$y
                        , $c[$i]);

```

```

203         if($result[1]=="yes"){
        if($_GET["wdt"]=="yes"){
            ImageFilledRectangle($im,$x-2,371-$y
                -2,$x+2,371-$y+2,$c[$i]);
        $nam=urlencode($woerter[$i]);
        $map .= "<area shape=\"rect\" title
            =\"Belegseite zu »$woerter[$i]«
            vom $dtag anzeigen\" coords=\"$x1
            , $y1, $x2, $y2\" href=\"$burl/$durl
            /$nam.html\" target=_blank />\n";
        }
        $tag=mktime (0,0,0,$ST[1] , $ST[2] + $j, $ST
            [0]);
        $durl= date("Y/m/d", $tag);
208        //$dtag= date("d.m.Y", $tag);
        $dtag= date("Y-m-d", $tag);
        $x1=$x-2;
        $x2=$x+2;
        $y1=371-$y-2;
213        $y2=371-$y+2;
        }
        $ylast=$y;
        } else {
            $ylast=34*$result[0];
218        }
        } else {
            if($started==1) {
                $y=$ylast/2;
                ImageLine($im,$xlast,371-$ylast,$x,371-$y,$c[
                    $i]);
223                $ylast=$y;
            }
        }
        $xlast=$x;

228        $q="SELECT signifikanz from coo$tmp where
            wort_bin=\"$wi\" and datum=\"$thedata\"";
        $a=mysql_query($q,$db);
        if($r=mysql_fetch_array($a)){
            ImageFilledRectangle($im,$x-2,433+12*$i,$x
                +2,437+12*$i,$c[$i]);
        }
233    }
    if($i>1){
        ImageTTFText($im,9,0,500,440+12*$i,$c[$i],$font,$wi)
        ;
    }
    }
238 }

```

```

ImageLine($im,0,370,455,370,$black);
243 ImagePNG($im,"/var/www/wdtno/images/$tmp.png");
    if(isSet($_GET["copy"])){
        copy("/var/www/wdtno/images/$tmp.png","/var/www/wdtno
        /$url/graph/$wort.png");
        unlink("/var/www/wdtno/images/$tmp.png");
    }
248 $map .= "</map>";
    echo "<img border=\"0\" src=\"/wdtno/images/$tmp.png\">";
    $q="DROP table words$tmp;";
    @mysql_query($q,$db);
    $q="DROP table coo$tmp;";
253 @mysql_query($q,$db);
    } # $_GET["wort"] is set
    ?>

<h2>New Query</h2>
258 <form target=_top action="<?php echo $scriptname; ?>" method=
    "GET"><pre>
        search term: <input type="text" size=40 name="wort"
            value="<? echo $_GET["wort"]?>" />
        start date*: <input type="text" size=10 name="StartOn"
            value="<? echo $_GET["StartOn"]?>" />
    <span style="color: #f93;">focus on** date*:</span> <input
        type="text" size=10 name="FocusOn" value="<? echo $_GET["
        FocusOn"]?>" ?>
263         end date*: <input type="text" size=10 name="EndOn"
            value="<? echo $_GET["EndOn"]?>" />

            <input type="submit">

    </pre></form>

268 * Please enter dates in the format YYYY-MM-DD! (Feb 01 2006
    to Feb 23 2006)<br />
    ** max. the 7 most significant co-occurrences from this day
        are displayed in addition to the search term
    </body>
    </html>

```

DATENBANKSCHEMA

Für die Speicherung der Daten wurde das im Zusammenhang mit Quasthoff u. a. (2006) entwickelte neue MySQL-Schema der Wortschatz-Datenbank um einen zusätzlichen Schlüssel erweitert, um eine ID für die jeweilige Zeitscheibe mit speichern zu können. Leider wuchsen dadurch die zu bewältigenden Datenmengen erheblich an, so sehr, dass bisweilen ein Standard-PC nicht mehr zur flüssigen Bewältigung ausreichte; aus diesem Grunde wurde von Gottwald (2007) das Programm WCTAnalyze (siehe Unter-Unterabschnitt 4.4.1.2) unabhängig von einer Datenbank entwickelt.

In der Wortschatz-Datenbank werden grundsätzlich Quelldokumente (mit URI und Datum in der Tabelle *sources*), Datensätze (in der Tabelle *sentences*), Types (mit Anzahl in der Tabelle *words*) und Kookkurrenzen als Beziehungen zwischen Types (in den Tabellen *co_s* und *co_n*) gespeichert. Die Zuordnung von Sätzen zu Quellen und von Types zu Sätzen findet jeweils als n:m-Relation in Form einer inversen Liste (optional mit Positionsangabe) statt (in den Tabellen *inv_so* und *inv_w*). Da MySQL Foreign Keys ignoriert, sind diese nicht explizit angegeben, ebenso sind die Indizes zur Optimierung spezieller Abfragen nicht angegeben. Die Modifikation zum Hinzufügen der Zeitscheibendaten beschränkt sich auf eine weitere Tabelle *slices* deren ID den Tabellen *words*, *co_s* und *co_n* als Bestandteil des Primärschlüssels hinzugefügt wird.

Listing C.1: "modifizierte Tabellenstruktur der Wortschatz-Datenbank"

```

CREATE TABLE 'slices' (
  'sl_id' int(10) unsigned NOT NULL auto_increment,
  'description' varchar(255) default NULL,
4  PRIMARY KEY ('sl_id')
);

CREATE TABLE 'sources' (
  'so_id' int(10) unsigned NOT NULL auto_increment,
9  'source' varchar(255) default NULL,
  'date' date default NULL,
  PRIMARY KEY ('so_id')
);

14 CREATE TABLE 'sentences' (
  's_id' int(10) unsigned NOT NULL auto_increment,

```

```

'sentence' text,
PRIMARY KEY ('s_id')
);
19
CREATE TABLE 'words' (
  'w_id' int(10) unsigned NOT NULL auto_increment,
  'sl_id' int(10) unsigned NOT NULL default '0',
  'word' varchar(255) default NULL,
24  'freq' int(8) unsigned default NULL,
PRIMARY KEY ('w_id','sl_id')
);

CREATE TABLE 'co_n' (
29  'w1_id' int(10) unsigned NOT NULL default '0',
  'w2_id' int(10) unsigned NOT NULL default '0',
  'sl_id' int(10) unsigned NOT NULL default '0',
  'sig' int(8) unsigned default NULL,
  'freq' int(8) unsigned default NULL,
34  PRIMARY KEY ('w1_id','w2_id','sl_id')
);

CREATE TABLE 'co_s' (
  'w1_id' int(10) unsigned NOT NULL default '0',
39  'w2_id' int(10) unsigned NOT NULL default '0',
  'sl_id' int(10) unsigned NOT NULL default '0',
  'sig' int(8) unsigned default NULL,
  'freq' int(8) unsigned default NULL,
44  PRIMARY KEY ('w1_id','w2_id','sl_id')
);

CREATE TABLE 'inv_so' (
  'so_id' int(10) unsigned NOT NULL default '0',
  's_id' int(10) unsigned NOT NULL default '0',
49  PRIMARY KEY ('so_id','s_id')
);

CREATE TABLE 'inv_w' (
  'w_id' int(10) unsigned NOT NULL default '0',
54  's_id' int(10) unsigned NOT NULL default '0',
PRIMARY KEY ('w_id','s_id')
);

```

WISSENSCHAFTLICHER WERDEGANG

Dezember 2007–Dezember 2008 Wissenschaftlicher Mitarbeiter an der Universität Leipzig im Rahmen des AiF geförderten Projekts NIMS – Neue Interaktive Multimediasuche.

April 2005–November 2007 Stipendiat der Medienstiftung der Sparkasse Leipzig. Promotionsstudium und Bearbeitung der vorliegenden Dissertation bei Prof. Gerhard Heyer in der Abteilung Automatische Sprachverarbeitung am Institut für Informatik an der Universität Leipzig.

Mai 2004–November 2004 Magisterarbeit über „Korpusbasierte Lemmaselektion“ unter Betreuung von Prof. Dr. Dr. Dietrich Kerlen (†), Dr. Thomas Keiderling und Prof. Dr. Gerhard Heyer im Bereich Buchwissenschaft am Institut für Kommunikations- und Medienwissenschaft der Universität Leipzig.

Oktober 1997–März 2005 Magisterstudium: Kommunikations- und Medienwissenschaft (1. Hauptfach), Physik (2. Hauptfach, bis September 1998) und Informatik (2. Hauptfach, ab Oktober 1998) an der Universität Leipzig.

PUBLIKATIONEN

Einige Ideen und Abbildungen wurden bereits in folgenden Publikationen veröffentlicht:

- Richter (2005): Analysis and Visualization for Daily Newspaper Corpora. In: *Proceedings of the RANLP*, Seite 424–428.
- Richter (2006): Künstlicher Sachverstand. In: *Message: Fachzeitschrift für Journalismus*, 1/2006.
- Quasthoff, Richter und Biemann (2006): Corpus Portal for Search in Monolingual Corpora. In: *Proceedings of the LREC 2006*.
- Eiken, Liseth, Witschel, Richter und Biemann (2006): Ord i dag: Mining Norwegian Daily Newswire. In: *Proceedings of the FinTAL 2006*.
- Richter, Quasthoff, Hallsteinsdóttir und Biemann (2006): Exploiting the Leipzig Corpora Collection. In: *Proceedings of IS-LTC 2006*.
- Gottwald, Heyer, Richter und Walde (2007): WCTAnalyze - Collecting, Indexing, Accessing and Visualizing Temporally Indexed Textual Resources. In: *Proceedings of the TIME 2007*.
- Hallsteinsdóttir, Eckart, Biemann, Quasthoff und Richter (2007): Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In: *Proceedings of the NODALIDA 2007*.
- Biemann, Heyer, Quasthoff und Richter (2007): The Leipzig Corpora Collection. Monolingual corpora of standard size. In: *Proceedings of Corpus Linguistics 2007*.
- Gottwald, Richter, Heyer und Scheuermann (2008): Tapping Huge Temporally Indexed Textual Resources with WCTAnalyze. In: *Proceedings of the LREC 2008*.

LITERATURVERZEICHNIS

- [Ahmad u. a. 2000] AHMAD, K. ; GILLAM, L. ; TOSTEVIN, L.: Weirddness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In: *Proceedings of TREC-8*, 2000, S. 717–724 (Zitiert auf Seite 67.)
- [Ankolekar u. a. 2007] ANKOLEKAR, A. ; KRÖTZSCH, M. ; TRAN, T. ; VRANDECIC, D.: The two cultures: mashing up web 2.0 and the semantic web. In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA : ACM Press, 2007, S. 825–834. – ISBN 978-1-59593-654-7 (Zitiert auf Seite 76.)
- [Ayres 2007] AYRES, I.: *Super Crunchers*. New York, NY: Bantam Dell, 2007 (Zitiert auf Seiten 71 und 79.)
- [Baeza-Yates und Ribeiro-Neto 1999] BAEZA-YATES, R. ; RIBEIRO-NETO, B.: *Modern Information Retrieval*. New York: Addison-Wesley, 1999 (Zitiert auf Seite 70.)
- [Berelson 1952] BERELSON, B. (Hrsg.): *Content analysis in communication research*. New York: Hafner, 1952 (Zitiert auf Seite 21.)
- [Berners-Lee 1999] BERNERS-LEE, T.: *Weaving the Web*. Orion Business Books, 1999 (Zitiert auf Seite 76.)
- [Berry 2003] BERRY, M.W.: *Survey of Text Mining: Clustering, Classification and Retrieval*. Springer, 2003 (Zitiert auf Seite 79.)
- [Besson 2005] BESSON, N. A.: *Medienresonanzanalyse DO-IT-YOURSELF*. Edingen-Neckarhausen: Dr. Besson Fachverlag, 2005 (Zitiert auf Seiten 24 und 78.)
- [Best u. a. 2006] BEST, C. ; POULIQUEN, B. ; STEINBERGER, R. ; GOOT, E. Van der ; BLACKLER, K. ; FUART, F. ; OELLINGER, T. ; IGNAT, C.: Towards Automatic Event Tracking. In: *Proc. of ISI 2006*, 2006 (Zitiert auf Seiten 29 und 39.)
- [Biemann 2005] BIEMANN, C.: Ontology Learning from Text - a Survey of Methods. In: *LDV-Forum* 20 (2005), Nr. 2, S. 75–93 (Zitiert auf Seite 27.)
- [Biemann 2006] BIEMANN, C.: Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to

- Natural Language Processing Problems. In: *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, URL <http://wortschatz.uni-leipzig.de/~cbiemann/pub/2006/BiemannTextGraph06.pdf>, 2006 (Zitiert auf Seite 73.)
- [Biemann 2007] BIEMANN, C.: *Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm*, Universität Leipzig, Dissertation, 2007 (Zitiert auf Seite 55.)
- [Biemann u. a. 2004] BIEMANN, C. ; BORDAG, S. ; HEYER, G. ; QUASTHOFF, U. ; WOLFF, C.: Language-independent Methods for Compiling Monolingual Lexical Data. In: *Proceedings of CicLING 2004* Bd. 2945. Seoul, South Korea : Springer, 2004, S. 215–228 (Zitiert auf Seite 35.)
- [Biemann u. a. 2007] BIEMANN, C. ; HEYER, G. ; QUASTHOFF, U. ; RICHTER, M.: The Leipzig Corpora Collection. Monolingual corpora of standard size. In: *Proceedings of Corpus Linguistics 2007*, 2007 (Zitiert auf Seite 113.)
- [Biemann und Teresniak 2005] BIEMANN, C. ; TERESNIAK, S.: Disentangling from Babylonian Confusion - Unsupervised Language Identification. In: *Proceedings of CICLing - Computational Linguistics and Intelligent Text Processing*, 2005 (Lecture Notes in Computer Science 3406), S. 762–773 (Zitiert auf Seite 52.)
- [Bordag 2007] BORDAG, S.: *Elements of Knowledge-free and Unsupervised Lexical Acquisition*, Universität Leipzig, Dissertation, 2007 (Zitiert auf Seite 12.)
- [Büchler 2006] BÜCHLER, M.: *Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten*, Universität Leipzig, Diplomarbeit, 2006 (Zitiert auf Seite 35.)
- [Butscher 2006] BUTSCHER, R.: *Text Mining in der Konsumentenforschung unter besonderer Berücksichtigung von Produktontologien*, Universität Erlangen, Dissertation, 3 April 2006. – URL <http://www.opus.ub.uni-erlangen.de/opus/volltexte/2006/346/> (Zitiert auf Seite 23.)
- [Cipollone 2006] CIPOLLONE, P.: *Boiled Frogs and Your Organization's Reputation*. Factiva from DowJones Whitepaper. 15 November 2006. – URL http://www.factiva.com/collateral/files/whitepaper_reputation_0605.pdf. – Zugriff am 5.7.2007 (Zitiert auf Seite 3.)

- [Cohen und Singer 1996] COHEN, W. W. ; SINGER, Y.: Context-sensitive methods for text categorization. In: *Proc. of SIGIR '96* (1996) (Zitiert auf Seite 62.)
- [Cui 2007] CUI, C.: *Dokument-Clustering von tagesaktuellen Zeitungstexten auf Grundlage der Wörter des Tages*, Universität Leipzig, Diplomarbeit, Februar 2007 (Zitiert auf Seite 73.)
- [De Roeck u. a. 2005] DE ROECK, A. ; SARKAR, A. ; GARTHWAITE, P.H.: Even very frequent function words do not distribute homogeneously. In: *Proceedings of the 5th RANLP*, 2005 (Zitiert auf Seite 11.)
- [DeWeese III und McCombs 1977] DEWEESE III, C. L. ; MCCOMBS, M. E.: Computer Content Analysis of "Day-Old" Newspapers: A Feasibility Study. In: *Public Opinion Quarterly* (1977), Nr. 41, S. 91–94 (Zitiert auf Seite 38.)
- [Diefenbach 2001] DIEFENBACH, D. L.: *Theory, Method, and Practice in Computer Content Analysis*. Kap. Historical Foundations of Computer-Assisted Content Analysis, S. 13–41, Westport, CT: Ablex, 2001 (Zitiert auf Seite 22.)
- [Dornseiff 2004] DORNSEIFF, F. ; QUASTHOFF, U. (Hrsg.): *Der deutsche Wortschatz nach Sachgruppen*. Walter de Gruyter: Berlin, 2004 (Zitiert auf Seite 83.)
- [Dunning 1994] DUNNING, T.: Statistical identification of language / New Mexico State University. Computing Research Lab. 1994. – CRL MCCS-94-273 (Zitiert auf Seite 52.)
- [Dunning 1993] DUNNING, T. E.: Accurate methods for the statistics of surprise and coincidence. In: *Computational Linguistics* 19 (1993), Nr. 1, S. 61–74 (Zitiert auf Seiten 12 und 61.)
- [Eiken u. a. 2006] EIKEN, U. C. ; LISETH, A. T. ; WITSCHER, H. F. ; RICHTER, M. ; BIEMANN, C.: Ord i dag: Mining Norwegian Daily Newswire. In: *Proceedings of the FinTAL*, 2006 (LNAI 4139), S. 512–523 (Zitiert auf Seiten 37 und 113.)
- [Esuli 2008] ESULI, A.: *Automatic generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*, Università di Pisa, Dissertation, 2008 (Zitiert auf Seite 26.)
- [Esuli und Sebastiani 2007] ESULI, A. ; SEBASTIANI, F.: Page Ranking WordNet Synsets: An Application to Opinion Mining. In: *Proceeding of the ACL07*, 2007, S. 424–431 (Zitiert auf Seite 26.)

- [Femers und Klewes 1997] FEMERS, S. ; KLEWES, J.: *PR-Erfolgskontrolle. Messen und Bewerten in der Öffentlichkeitsarbeit. Verfahren, Strategien, Beispiele*. Kap. Medienresonanzanalysen als Evaluationsinstrument der Öffentlichkeitsarbeit, S. 115–134, Frankfurt: IMK, 1997 (Zitiert auf Seite 24.)
- [Fiscus und Doddington 2002] FISCUS, J. G. ; DODDINGTON, G. R.: *Topic detection and tracking evaluation overview*. S. 17–31. Norwell, MA, USA : Kluwer Academic Publishers, 2002. – ISBN 0-7923-7664-1 (Zitiert auf Seite 38.)
- [Früh 1998] FRÜH, W.: *Inhaltsanalyse. Theorie und Praxis*. UVK Medien, Konstanz, Germany, 1998 (Zitiert auf Seiten 21 und 22.)
- [Fry 2008] FRY, B.: *Visualizing Data*. O'Reilly: Sebastopol, CA, 2008 (Zitiert auf Seite 74.)
- [Gale und Sampson 1995] GALE, W. ; SAMPSON, G.: Good-Turing frequency estimation without tears. In: *Journal of Quantitative Linguistics* (1995), Nr. 2, S. 217–237 (Zitiert auf Seite 61.)
- [Godbole u. a. 2007] GODBOLE, N. ; SRINIVASAIAH, M. ; SKIENA, S.: Large-Scale Sentiment Analysis for News and Blogs. In: *Proceedings of ICWSM 2007*, 2007 (Zitiert auf Seite 33.)
- [Gottwald 2007] GOTTWALD, S.: *Strategien, Konzepte und prototypische Entwicklung einer Software für die semiautomatische Analyse chronologischer Textmuster in Zeitschriftenkorpora*, Universität Leipzig, Diplomarbeit, 2007 (Zitiert auf Seiten 73 und 110.)
- [Gottwald u. a. 2007] GOTTWALD, S. ; HEYER, G. ; RICHTER, M. ; WALDE, P.: WCTAnalyze - Collecting, Indexing, Accessing and Visualizing Temporally Indexed Textual Resources. In: *Proceedings of the TIME 2007*, 2007 (Zitiert auf Seiten 73 und 113.)
- [Gottwald u. a. 2008] GOTTWALD, S. ; RICHTER, M. ; HEYER, G. ; SCHEUERMANN, G.: Tapping Huge Temporally Indexed Textual Resources with WCTAnalyze. In: *Proceedings of the LREC'o8 (to appear)*, 2008 (Zitiert auf Seite 113.)
- [Gulzar und ur Rahman 2007] GULZAR, A. ; RAHMAN, S. ur: Nastaleeq: A challenge accepted by Omega. In: *Proceedings of the 17th European TeX Conference, Bachotek, Poland*, 2007 (Zitiert auf Seite 46.)
- [Haller 1993] HALLER, M.: *Theorien öffentlicher Kommunikation. Problemfelder, Positionen, Perspektiven*. Kap. Journalistisches

- Handeln: Vermittlung oder Konstruktion von Wirklichkeit?, S. 137–151, 1993 (Zitiert auf Seite 19.)
- [Hallsteinsdóttir u. a. 2007] HALLSTEINSDÓTTIR, E. ; ECKART, T. ; BIEMANN, C. ; QUASTHOFF, U. ; RICHTER, M.: Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In: *Proceedings of the NODALIDA 2007* (2007) (Zitiert auf Seite 113.)
- [Hamilton 1994] HAMILTON, J.: *Time Series Analysis*. Princeton University Press, 1994 (Zitiert auf Seite 18.)
- [Heine u. a. 2008] HEINE, C. ; BORDAG, S. ; SCHEUERMANN, G. ; HEYER, G.: Euler Diagram Rendering on the GPU for Document Cluster Exploration. In: *Submitted at Eurographics 2008* Bd. 27, 2008 (Zitiert auf Seite 74.)
- [Heyer u. a. 2006] HEYER, G. ; QUASTHOFF, U. ; WITTIG, T.: *Wissenrohstoff Text*. w3l, 2006 (Zitiert auf Seiten 2, 9, 12, 27 und 59.)
- [Hof 2005] HOF, R.: Mix, Match, and Mutate. In: *Business Week* (2005), 25 Juli. – URL http://www.businessweek.com/@76IH*ocQ34AvyQMA/magazine/content/05_30/b3944108_mz063.htm (Zitiert auf Seite 76.)
- [Hofland 2000] HOFLAND, K.: A Self-Expanding Corpus Based on Newspapers on the Web. In: *Proceedings of the LRECo, 2000* (Zitiert auf Seite 50.)
- [Johannessen u. a. 2000] JOHANNESSEN, J.-B. ; HAGEN, K. ; NØKLESTAD, A.: A Constraint-based tagger for Norwegian. In: *17th Scandinavian Conference of Linguistics* Bd. 1 University of Southern Denmark, Odense (Veranst.), 2000, S. 31–47 (Zitiert auf Seite 55.)
- [Kepplinger 1989] KEPPLINGER, H. M.: *Fortschritte der Publizistikwissenschaft*. Kap. Realität, Realitätsdarstellung und Medienwirkung, S. 39–55, Freiburg i. Br: Verlag Karl Alber, 1989 (Zitiert auf Seite 19.)
- [Kleinberg 2002] KLEINBERG, J.: Bursty and hierarchical structure in streams. In: *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA : ACM Press, 2002, S. 91–101. – ISBN 1-58113-567-X (Zitiert auf Seite 11.)
- [Klingemann 1984] KLINGEMANN, H.-D.: *Computergestützte Inhaltsanalyse in der empirischen Sozialforschung*. Frankfurt a.M., 1984 (Zitiert auf Seite 23.)

- [Kontostathis u.a. 2003] KONTOSTATHIS, A. ; GALITSKY, L. ; POTTENGER, W. M. ; ROY, S. ; PHELPS, D. J.: *A comprehensive survey of text mining*. Kap. A survey of emerging trend detection in textual data mining, Springer, 2003 (Zitiert auf Seite 38.)
- [Krippendorff 2004] KRIPPENDORFF, K.: *Content Analysis: An Introduction to Its Methodology*. Sage, 2004 (Zitiert auf Seiten 21 und 22.)
- [Kuttner 1981] KUTTNER, H.-G.: *Zur Relevanz text- und inhaltsanalytischer Verfahrensweisen für die empirische Forschung*. Frankfurt am Main, Bern: Verlag Peter D. Lang, 1981 (Zitiert auf Seite 22.)
- [Landmann und Züll 2004] LANDMANN, J. ; ZÜLL, C.: Computerunterstützte Inhaltsanalyse ohne Diktionär? In: *ZUMA-Nachrichten* 54 (2004), Nr. 05, S. 117–140 (Zitiert auf Seite 27.)
- [Lenz 2005] LENZ, H.: *Universalgeschichte der Zeit*. Marixverlag, 2005 (Zitiert auf Seite 18.)
- [Lilienthal 2001] LILIENTHAL, V.: *Senderfertig abgesetzt. ZDF, SAT.1 und der Soldatenmord von Lebach*. Vistas Verlag, Berlin, 2001 (Zitiert auf Seite 43.)
- [Lissmann 2001] LISSMANN, U.: *Inhaltsanalyse von Texten - Ein Lehrbuch zur computerunterstützten und konventionellen Inhaltsanalyse*. 2. Landau: Verlag Empirische Pädagogik, 2001 (Zitiert auf Seite 22.)
- [Lloyd u. a. 2005] LLOYD, L. ; KECHAGIAS, D. ; SKIENA, S.: Lydia: A System for Large-Scale News Analysis. In: *Proceedings of SPIRE 2005*, 2005 (Zitiert auf Seite 29.)
- [Luhmann 2004] LUHMANN, N.: *Die Realität der Massenmedien*. 3. Aufl. VS Verlag für Sozialwissenschaften, 2004 (Zitiert auf Seiten 19 und 20.)
- [MacQueen 1967] MACQUEEN, J. B.: Some Methods for classification and Analysis of Multivariate Observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967 (Zitiert auf Seite 63.)
- [Maurer und Reinemann 2006] MAURER, M. ; REINEMANN, C.: *Medieninhalte: Eine Einführung*. Wiesbaden, VS Verlag für Sozialwissenschaften, 2006 (Zitiert auf Seite 24.)
- [Mayring 2000] MAYRING, P.: *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. 7. Weinheim: Deutscher Studien Verlag, 2000 (Zitiert auf Seite 22.)

- [McKeown u. a. 2002] MCKEOWN, K. R. ; BARZILAY, R. ; EVANS, D. ; HATZIVASSILOGLU, V. ; KLAUVANS, J. L. ; SABLE, C. ; SCHIFFMAN, B. ; SIGELMAN, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In: *Proceedings of the Human Language Technology Conference, 2002* (Zitiert auf Seiten 28 und 39.)
- [Merten 2004] MERTEN, K.: *Inhaltsanalyse: Einführung in Theorie, Methode und Praxis*. Westdeutscher Verlag, 2004 (Zitiert auf Seiten 2 und 21.)
- [Merten u. a. 2005] MERTEN, K. ; DAHM, C. ; SPRIESTERSBACH, T. ; TOP, J. ; WINTERBERG, P.: *Medien, Dachse & Tenöre*. Münster: Lit Verlag, 2005 (Zitiert auf Seiten 24, 25 und 78.)
- [Merten und Wienand 2004] MERTEN, K. ; WIENAND, E.: *Medienresonanzanalyse*. Manuskript eines Vortrags. 2004. – URL <http://comdat.de/downloads/Medienresonanzanalyse%20Vortrag%202004.07.05%20Berlin.pdf>. – Heruntergeladen am 4. September 2006 (Zitiert auf Seite 24.)
- [Mishne 2007] MISHNE, G. A.: *Applied Text Analytics for Blogs*, Universität van Amsterdam, Dissertation, 27 April 2007 (Zitiert auf Seite 49.)
- [Popper 1995] POPPER, K. R.: *Objektive Erkenntnis*. Hoffmann und Campe, 1995 (Zitiert auf Seite 19.)
- [Pustejovsky u. a. 2003] PUSTEJOVSKY, J. ; CASTAÑO, J. ; INGRÍA, R. ; GAIZAUSKAS, R. ; SETZER, A. ; KATZ, G.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: *Proc. of IWCS-5 (Fifth International Workshop on Computational Semantics)*, 2003 (Zitiert auf Seite 18.)
- [Quasthoff 1998] QUASTHOFF, U.: *Linguistik und neue Medien*. Kap. Projekt Deutscher Wortschatz, DUV, 1998 (Zitiert auf Seite 34.)
- [Quasthoff 2007] QUASTHOFF, U.: *Deutsches Neologismenwörterbuch*. Walter de Gruyter: Berlin, New York, 2007 (Zitiert auf Seite 83.)
- [Quasthoff u. a. 2002a] QUASTHOFF, U. ; BIEMANN, C. ; WOLFF, C.: Named Entity Learning and Verification: Expectation Maximisation in Large Corpora. In: *Proceedings of CoNLL-2002*, 2002, S. 8–14 (Zitiert auf Seite 57.)
- [Quasthoff u. a. 2006] QUASTHOFF, U. ; RICHTER, M. ; BIEMANN, C.: Corpus Portal for Search in Monolingual Corpora. In:

Proceedings of the LREC, 2006 (Zitiert auf Seiten 11, 36, 50, 52, 110 und 113.)

[Quasthoff u. a. 2002b] QUASTHOFF, U. ; RICHTER, M. ; WOLFF, C.: Wörter des Tages - tagesaktuelle wissensbasierte Analyse und Visualisierung von Zeitungen und Newsdiensten. In: HAMMWÖHNER, R. (Hrsg.) ; WOLFF, C. (Hrsg.) ; WOMSER-HACKER, C. (Hrsg.): *Information und Mobilität. Proc. 8. Internationales Symposium für Informationswissenschaft*, UVK, 2002 (Schriften zur Informationswissenschaft 40), S. 369–372 (Zitiert auf Seite 37.)

[Quasthoff u. a. 2003] QUASTHOFF, U. ; RICHTER, M. ; WOLFF, C.: Medienanalyse und Visualisierung: Auswertung von Online-Presstexten durch Text Mining. In: SEEWALD-HEEG, Uta (Hrsg.): *Sprachtechnologie für die multilinguale Kommunikation - Textproduktion, Recherche, Übersetzung, Lokalisierung - Beiträge der GLDV-Frühjahrstagung 2003* Bd. 5, 2003, S. 442–459 (Zitiert auf Seite 37.)

[Quasthoff und Wolff 1999] QUASTHOFF, U. ; WOLFF, C.: Korpuslinguistik und große einsprachige Wörterbücher. In: *Linguistik Online* (1999). – URL http://viadrina.euv-frankfurt-o.de/~wjournal/2_99/quasthoff.html (Zitiert auf Seite 34.)

[Quasthoff und Wolff 2002] QUASTHOFF, U. ; WOLFF, C.: The Poisson collocations measure and its application. In: *Workshop on Computational Approaches to Collocations, Vienna, Austria*, 2002 (Zitiert auf Seite 12.)

[Radev u. a. 2005] RADEV, D. ; OTTERBACHER, J. ; WINKEL, A. ; BLAIR-GOLDENSON, A.: NewsInEssence: Summarizing Online News Topics. In: *Communications of the ACM* 48 (2005), Nr. 10, S. 95–98 (Zitiert auf Seiten 28 und 39.)

[Richter 2004] RICHTER, M.: *Korpusbasierte Lemmaselektion*, Leipzig University, Diplomarbeit, 2004 (Zitiert auf Seiten 8, 40 und 83.)

[Richter 2005] RICHTER, M.: Analysis and Visualization for Daily Newspaper Corpora. In: *Proceedings of the RANLP*, 2005, S. 424–428 (Zitiert auf Seiten 37 und 113.)

[Richter u. a. 2006] RICHTER, M. ; QUASTHOFF, U. ; HALLSTEINSDÓTTIR, E. ; BIEMANN, C.: Exploiting the Leipzig Corpora Collection. In: *Proceedings of IS-LTCo6*, 2006 (Zitiert auf Seite 113.)

- [Rössler 2005] RÖSSLER, P.: *Inhaltsanalyse*. UVK, 2005 (Zitiert auf Seite 22.)
- [Schmidt 1999] SCHMIDT, F.: *Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung*. <http://lips.informatik.uni-leipzig.de/pub/1999-18>, Leipzig University, Diplomarbeit, 1999 (Zitiert auf Seite 68.)
- [Schönbach 1984] SCHÖNBACH, K.: *Computergestützte Inhaltsanalyse in der empirischen Sozialforschung*. Kap. "The Issues of the Seventies": Computerunterstützte Inhaltsanalyse und die langfristige Beobachtung von Agenda-Setting-Wirkungen der Massenmedien, S. 131–154, Frankfurt am Main / New York: Campus Verlag, 1984 (Zitiert auf Seite 38.)
- [Shneiderman 1992] SHNEIDERMAN, B.: Tree visualization with tree-maps: 2-d space-filling approach. In: *ACM Transactions on Graphics* 11 (1992), Januar, Nr. 1, S. 92–99 (Zitiert auf Seite 74.)
- [Smith 2008] SMITH, G.: *Tagging - People-Powered Metadata for the Social Web*. New Riders: Berkeley, CA, 2008 (Zitiert auf Seite 67.)
- [Stefaner 2007] STEFANER, M.: *Visual Tools for the Socio-Semantic Web*, University of Applied Sciences Potsdam, Diplomarbeit, Juni 2007 (Zitiert auf Seite 76.)
- [Teichert u. a. 2008] TEICHERT, Th. ; HEYER, G. ; SCHÖNTAG, K. ; MAIRIE, P.: Co-Word Analysis for Assessing Consumer Associations: A Case Study in Market Research. In: *Proceedings of „Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology (EMOT-08)“*, 2008 (Zitiert auf Seite 26.)
- [Titscher u. a. 1998] TITSCHER, S. ; WODAK, R. ; MEYER, M. ; VETTER, E.: *Methoden der Textanalyse. Leitfaden und Überblick*. Opladen: Westdeutscher Verlag, 1998 (Zitiert auf Seiten 2 und 21.)
- [Trotta 2006] TROTTA, G.: *ActiveGrid: How Mashups Complement SOAs*. 21. Dezember 2006. – URL <http://www.ebizq.net/blogs/firstlook/2006/12/mashups.php>. – Abruf: 7. Juni 2007 (Zitiert auf Seite 76.)
- [Türcke 2005] TÜRCKE, C.: *Vom Kainszeichen zum genetischen Code. Kritische Theorie der Schrift*. C. H. Beck Verlag, 2005 (Zitiert auf Seite 2.)
- [Viterbi 1967] VITERBI, A. J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In:

- IEEE Transactions on Information Theory* 13 (1967), Nr. 2, S. 260–269 (Zitiert auf Seite 56.)
- [Wayne 2000] WAYNE, C.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In: *Proceedings of the LREC*, 2000, S. 1487–1494 (Zitiert auf Seite 38.)
- [Witschel 2005] WITSCHSEL, H. F.: Terminology Extraction and Automatic Indexing. Comparison and Qualitative Evaluation of Methods. In: *Proceedings of Terminology in Knowledge Engineering 2005*, 2005 (Zitiert auf Seiten 58, 64 und 74.)
- [Witschel und Biemann 2005] WITSCHSEL, H.F. ; BIEMANN, C.: Rigorous dimensionality reduction through linguistically motivated feature selection for text categorisation. In: *Proceedings of NODALIDA*, 2005 (Zitiert auf Seite 55.)
- [Wong und Hong 2007] WONG, J. ; HONG, J. I.: Making mashups with marmite: towards end-user programming for the web. In: *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA : ACM Press, 2007, S. 1435–1444. – ISBN 978-1-59593-593-9 (Zitiert auf Seite 76.)
- [Zerfaß und Boelter 2005] ZERFASS, A. ; BOELTER, D.: *Die neuen Meinungsmacher : Weblogs als Herausforderung für Kampagnen, Marketing, PR und Medien*. Nausner & Nausner, 2005 (Zitiert auf Seite 48.)
- [Zipf 1935] ZIPF, G. K.: *The Psycho-Biology of Language*. Houghton-Mifflin, 1935 (Zitiert auf Seite 8.)
- [Züll und Alexa 2001] ZÜLL, C. ; ALEXA, M.: *Inhaltsanalyse: Perspektiven, Probleme, Potentiale*. Kap. Automatisches Codieren von Textdaten. Ein Überblick über neue Entwicklungen, S. 303–317, Köln: Herbert von Halem-Verlag, 2001 (Zitiert auf Seiten 22 und 23.)
- [Züll und Mohler 1992] ZÜLL, C. (Hrsg.) ; MOHLER, P. (Hrsg.): *Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse*. Westdeutscher Verlag, 1992 (Zitiert auf Seite 22.)

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, die vorliegende Dissertation selbstständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 18. Dezember 2008

Matthias Richter, M.A.