Christian Wolff

Zeitbezogene Korpusauswertung – Medienanalyse oder Sprachwandelforschung?

1 Einleitung

In den vergangenen Jahren sind im Zuge der Renaissance der Korpuslinguistik zahlreiche Korpora des Deutschen verfügbar geworden.[1] Zwar sind aufgrund der Vielzahl von Unterscheidungskriterien für elektronische Sprachkorpora (u.a. Sprache, Modalität, Textsorte, Annotation, Quellen) noch längst nicht alle wünschenswerten Korpora vorhanden, die bestehenden Daten eröffnen aber bereits jetzt neue Möglichkeiten für die Sprach- und Medienanalyse. Die wissenschaftshistorisch bedingte partielle Zurückdrängung empirischer Ansätze in der Sprachwissenschaft durch den vor allem in der theoretischen Linguistik vorherrschenden Rationalismus kann mittlerweile als überwunden gelten (vgl. dazu Pereira 2000[2]).

Der nachfolgende Aufsatz widmet sich der Frage, inwieweit mit Hilfe der mittlerweile umfangreichen Korpora des Deutschen Aussagen zu kurzfristigen Änderungen im Sprachgebrauch möglich sind und inwiefern sich solche korpuslinguistischen Auswertungsverfahren auch für die Medienanalyse fruchtbar machen lassen. Da sich die heute vorliegenden – großen – Textkorpora des Deutschen vornehmlich aus Pressetexten zusammensetzen, ist der Zusammenhang von Sprach- und Medienanalyse offensichtlich.

Aufbauend auf einem kurzen Überblick zu den derzeit (online) verfügbaren Korpora des Deutschen werden anhand von Beispielen Möglichkeiten der zeitorientierten Korpusanalyse gegeben (Kap. 3). Ein kurzer Ausblick erörtert abschließend das Verhältnis von Sprach- und Medienanalyse.

2 Elektronische Ressourcen und Textdatenbanken des Deutschen

Für die hier betrachteten Methoden der zeitbezogenen Korpusanalyse kommen vor allem größere Korpora des Deutschen in Betracht, da bei Auf

---

[1] Vgl. die Übersicht bei Lemnitzer/Zinsmeister 2006, 113ff.

[2] Pereira 2000, 1239 drückt es so aus: "In the last forty years, research on models of spoken and written language has been split between two seemingly irreconcilable traditions: formal linguistics in the Chomsky tradition, and information theory in the Shannon tradition", vgl. Lemnitzer/Zinsmeister 2006, 15ff.

des Materialbestands auf die einzelnen „Zeitscheiben" in jedem Teilkorpus noch hinreichend sprachliches Material verfügbar sein sollte, um es für korpuslinguistische Analysen nutzen zu können. Beim aktuellen Stand der Korpusentwicklung umfassen solche Korpora typischerweise wenigstens einige hundert Millionen laufende Wortformen (*token*) und dabei mehrere Millionen verschiedene Wörter (*types*, in der Regel Vollformen). Zu den Korpora, die auch den nachfolgenden Beispielen für zeitbezogene Analysen zugrunde liegen, gehören unter anderem:

- Die Korpora des Instituts für Deutsche Sprache in Mannheim (IDS), mit insgesamt etwa zwei Milliarden laufenden Wortformen.
- Das deutsche Korpus des Leipziger Projekts *Deutscher Wortschatz* mit aktuell ca. 500 Millionen laufenden Wortformen (vgl. Quasthoff/Wolff 2000).
- Das Kernkorpus des Projektes *Digitales Wörterbuch der Deutschen Sprache* (DWDS) an der Berlin-Brandenburgischen Akademie der Wissenschaften (vgl. Klein 2004, Geyken 2005).

In den letzten Jahren ist dabei als Trend der Korpuslinguistik die Analyse von Korpora, die teilweise oder überwiegend aus *online publizierten* Texten bestehen, zu beobachten. Insbesondere die umfangreichen Daten des World Wide Web ermöglichen es mittlerweile, auch sog. *gigaword corpora* und demnächst auch *teraword corpora* mit jeweils Milliarden oder gar Billionen laufender Wortformen zu nutzen.[3] Von den oben genannten Korpora ist vor allem das Leipziger Korpus aus solchen ausgewählten Webquellen zusammengesetzt. Dagegen sind *crossmediale* Korpora, die Quellen nicht nur aus unterschiedlichen Printmedien zusammenstellen, sondern verschiedene Kommunikationsmodalitäten integrieren (*gesprochene Sprache* in Film, Funk und Fernsehen, *Text* aus unterschiedlichen Quellen), bisher aufgrund des hohen Aufbereitungsaufwands noch sehr selten. Sie stellen aber gerade für die Medienanalyse im Sinne einer ganzheitlichen Erfassung und Untersuchung der von einem Mitglied einer Sprachgemeinschaft wahrgenommenen medial vermittelten sprachlichen Daten ein wichtiges Desiderat dar.

3 Zeitbezogene Analysen von (Text-)Korpora

Will man Korpora linguistisch oder medienanalytisch mit Bezug zum Publikationszeitraum von Texten auswerten, kommt zu den bereits einleitend genannten Kriterien des Korpusaufbaus die Frage nach Zeitgranularität, d.h. nach der kleinsten im Korpus unterschiedenen Zeiteinheit hinzu. Derzeit findet sich vor allem der Tag als kleinste Zeiteinheit, was

[3] Einführend zur Nutzung des World Wide Web als linguistisches Korpus vgl. Kilgarriff/Grefenstette 2003.

durch den typischen Publikationsrhythmus Online-Medien als Quellengrundlage bedingt ist. Angesichts der Vielzahl von Online-Medien, die eine kontinuierliche Publikationsmöglichkeit bieten (z.B. e-Mail, Online-Foren, Wikis und Weblogs (*blogs*), vgl. Schlobinski 2006), wäre grundsätzlich auch eine Analyse mit feinerer zeitlicher Auflösung möglich und wünschenswert.

Am anderen Ende des Spektrums zeitlicher Auflösung stehen diachrone Korpora, deren Zeiteinteilung sich über Jahrhunderte erstrecken kann, die aber in der Regel keinen Bezug zur Medienanalyse aufweisen, sondern rein linguistisch motiviert sind. Das Kernkorpus des *Digitalen Wörterbuchs der Deutschen Sprache* weist eine Einteilung nach Dekaden von 1900 bis 2000 auf und sichert vor allem auch eine repräsentative Zusammenstellung nach Textsorten über die verschiedenen Dekaden hinweg (s.u. Abb. 1-4).

Die nachfolgenden Beispiele für zeitbezogene Analysen von Medientexten sollen den aktuellen Stand zeitbezogener Korpusanalyse aufzeigen und gleichzeitig die enge Verschränkung von Medienanalyse und sprachwissenschaftlicher Betrachtung illustrieren.

3.1 Tübinger Wortwarte

Vor allem der Entdeckung von Neologismen widmet sich die Tübinger *Wortwarte*, die als Online-Plattform seit 2000 betrieben wird und tagesgenau neue (deutsche) Wörter aus online publizierten Texten dokumentiert. Zu den wesentlichen Zielen der Wortwarte rechnen Lemnitzer/Uhle 2006 vor allem die Beschreibung und Kommentierung von „Tendenzen der Entwicklung des Deutschen".

Die Auswahl geeigneter Wortkandidaten für Neologismen erfolgt auf der Basis einer in der Regel täglichen Auswertung der Online-Presse (Spiegel online, Süddeutsche online etc.) und basiert auf dem Vergleich mit dem gemeinsam vom Institut für Deutsche Sprache, Mannheim, dem Seminar für Sprachwissenschaft der Universität Tübingen und dem Institut für maschinelle Sprachverarbeitung der Universität Stuttgart erarbeiteten und betriebenen *Deutschen Referenzkorpus* (DEREKO) mit etwa 120 Millionen laufenden Wortformen (*tokens*) und ca. 2,3 Millionen verschiedenen Wörtern (*types,* vgl. Dipper et al. 2002). Um „zufällige" Fehler (z.B. Tippfehler) oder Falschschreibungen auszufiltern, erfolgt die Auswahl von Neologismen aus der nach dem Abgleich mit DEREKO verbleibenden Kandidatenmenge von Hand. Als sprachliche Norm wird hier die neue deutsche Rechtschreibung herangezogen. Die folgenden zwei Beispiellisten zeigen die ausgewählten Wörter für den 11. und 12. Dezember 2006:

Neologismen der Tübinger Wortwarte vom 11. Dezember 2006	***Neologismen der Tübinger Wortwarte vom 12. Dezember 2006***
Aktivgurtsystem, das; Blitzdating, das; Chefkickstarter, der;	Altersteilzeitler, der; Amokankündigung, die; Aquagrafie, die; Atomblues, der; Bayernbox, die;

Deontologe, der; Freegan, der / die; Gewinnabschöpfungsverfahren, das; Icescating, das; Klimakteriumsberater, der; Kompaktkonsole, die; Remixarbeit, die; Snowtube, der / das?; Studivzgruppe, die; Trivialzwerg, der; Vorkammerdiesel, der	Bayernkonsole, die; Biosentimentalität, die; Biotechgenerikum, das; Comicbattle, die; Dispokinetiker, der; Doppelsteckung, die; Fertigglühwein, der; Flachdrahtspulentechnik, die; Fondsifikat, das; Futterpate, der; Gourmetvideo, das; Greenbag, die; Hesylierung, die; Klingeltonranking, das; Meinungsarchiv, das; pegyliert, Adjektiv; Pegylierung, die; submongoloid, Adjektiv; Subraumsatellitenfrequenz, die; telerealistisch, Adjektiv; Vielschwafeltest, der; Zwangsmonopol, das

Tabelle 1: Neologismen der Tübinger Wortwarte vom 11. und 12. 12. 2006[4]

Es ist offensichtlich, dass mit dem Auswahlkriterium „erstmaliges Erscheinen eines Wortes in einer Online-Pressequelle" der Zeitpunkt der tatsächlichen Begriffsschöpfung nur näherungsweise bestimmt werden kann: Die Wortwarte bietet für jeden ihrer Neologismen auch eine Vergleichsrecherche mit der Suchmaschine *Google* an, über die weitere Quellen erschlossen werden können. Während etwa für das Wort *Vielschwafeltest* auch bei Google ausschließlich der Berliner Tagesspiegel als Quelle ersichtlich ist, sind für *Zwangsmonopol* bei Google immerhin 587 ältere Quellen, u.a. in Parlamentsprotokollen ersichtlich und man wird hier eher eine lexikalische Lücke im Referenzkorpus DEREKO entdecken als einen tatsächlichen Neologismus des Dezember 2006 (wohl ähnlich bei *Meinungsarchiv, Aktivgurtsystem* oder *Gewinnabschöpfungsverfahren*). Unterschiedliche Muster des erstmaligen Auftretens in online publizierten Medien sind einer näheren Betrachtung würdig: Während bei einer Vielzahl von Begriffen deutlich ist, dass es sich um möglicherweise bereits länger verwendete Fachterminologie handelt, die eben nur erstmalig im Beobachtungszeitraum 2000-2006 in einem Publikumsmedium verwendet wird (*Zwangsmonopol*, *Gewinnabschöpfungsverfahren*), lassen sich andere Begriffe als Bezeichnungen technologischer Innovationen begreifen (*Aktivgurtsystem*, *Biotechgenerikum*). Journalistische geprägte Sprachkreativität dürfte sich dagegen hinter Wörtern wie *Vielschwafeltest* oder *Trivialzwerg* verbergen, wobei auch Mischformen denkbar erscheinen (*Blitzdating*).

3.2 Zeitverläufe im Korpus des digitalen Wörterbuchs der Deutschen Sprache (DWDS)

Neben einer mit umfassender Funktionalität ausgestatteten Benutzerschnittstelle („Wortinformationssystem des DWDS"), die unter anderem eine präzise zeitliche Einschränkung der Recherche und die Darstellung von Kollokationsmengen zu den Suchbegriffen auch nach unterschiedlichen

[4] Quelle: Universität Tübingen, Seminar für Sprachwissenschaft (Hrsg.) (2006). Die Wortwarte. http://www.sfs.uni-tuebingen.de/~lothar/nw/Archiv/Datum/d061212.html (12. 12. 2006) und http://www.sfs.uni-tuebingen.de/~lothar/nw/Archiv/Datum/d061211.html (11. 12. 2006), Zugriff Dezember 2006.

Kollokationsmaßen erlaubt, bietet die DWDS-Schnittstelle eine Visualisierung des Gebrauchsverlaufs geordnet nach Dekaden und Textsorten. Es lässt sich damit nicht nur die absolute Auftretenshäufigkeit eines Wortes in der jeweiligen Dekade, sondern auch die Verteilung der Treffer auf verschiedene Textsorten im Korpus darstellen. Damit kann man – gewissermaßen „auf einen Blick" – für beliebige Wörter einen schnellen Überblick über ihre Gebrauchspraxis gewinnen. Einige Beispielbilder aus der DWDS-Schnittstelle sollen dies verdeutlichen. Die folgenden drei Abbildungen zeigen die Verlaufscharakteristika für die drei Wörter *Automobil*, *Auto* und *PKW* von 1900 bis 2000. Alle drei Verläufe weisen einen Maximalwert in der Dekade der 1950er Jahre auf, und dabei vor allem in der Textsorte *Gebrauchsliteratur*. Dies dürfte sich durch die Bedeutung des Automobils als Symbol des (west-)deutschen Wirtschaftswunders in den fünfziger Jahren erklären lassen, während *PKW* als *terminus technicus* offensichtlich erst in den Dreißiger Jahren eingeführt wurde. Bemerkenswert ist auch, dass die Kurzform *Auto* gegenüber *Automobil* bereits ab der Dekade 1920-1930 deutlich häufiger im Korpus nachgewiesen wird.

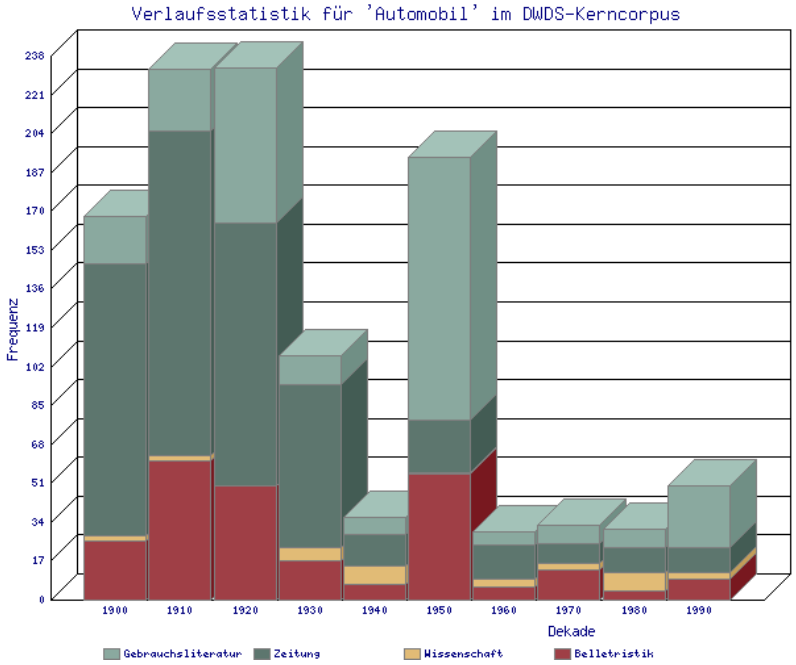

Abbildung 1: Zeitverlauf für Automobil *im DWDS-Kernkorpus*[5]

[5] Quelle für die Abbildungen 1-4: Berlin-Brandenburgische Akademie der Wissenschaften (Hrsg.) (2006). Projekt "Digitales Wörterbuch": online-Ressourcen. Zeitverläufe. Online-Abfragen, http://www.dwds.de/, Zugriff Dezember 2006.

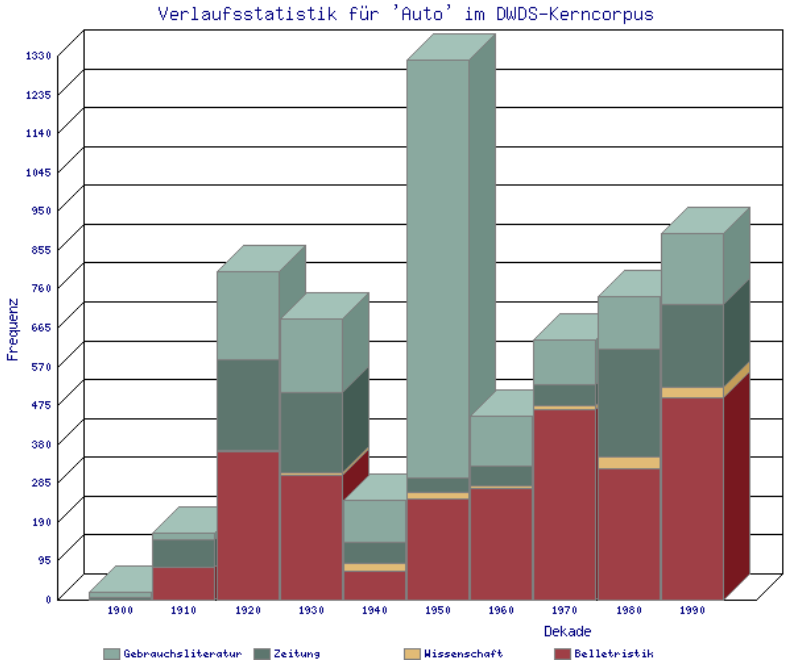

Abbildung 2: Zeitverlauf für Auto *im DWDS-Kernkorpus*

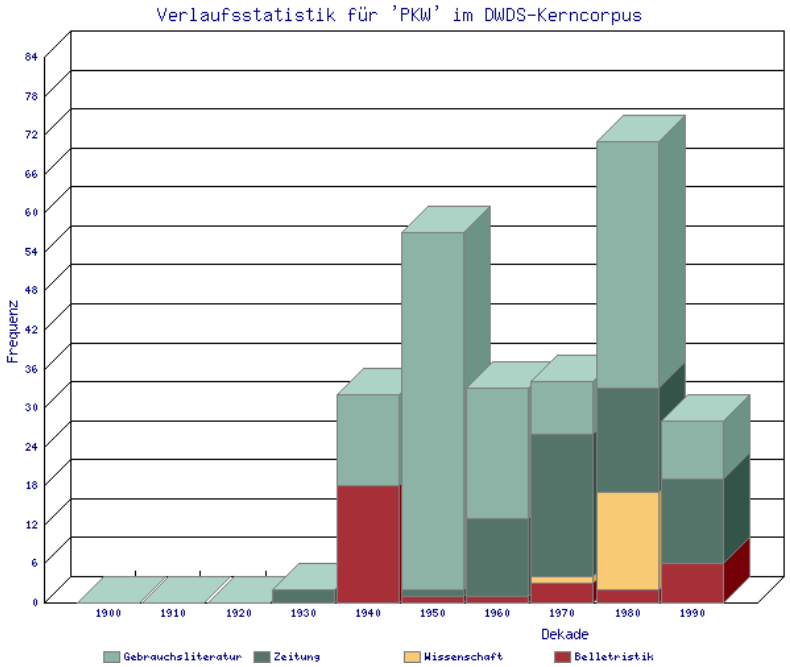

Abbildung 3: Zeitverlauf für PKW *im DWDS-Kernkorpus*

Ein zweites interessantes Beispiel liefert das Wort „Gau“. Als wichtiger Begriff der Sprache des Nationalsozialismus ist das ursprünglich neutral konnotierte Wort[6] seit dem Ende des dritten Reiches kaum mehr im Gebrauch, mit der signifikanten Ausnahme der 80er Jahre, als es als Akronym für *größter anzunehmender Unfall* im Rahmen der Kernkraftdebatte zu neuer Prominenz – nach der Einteilung des DWDS vor allem in wissenschaftlichen Texten – gelangte.

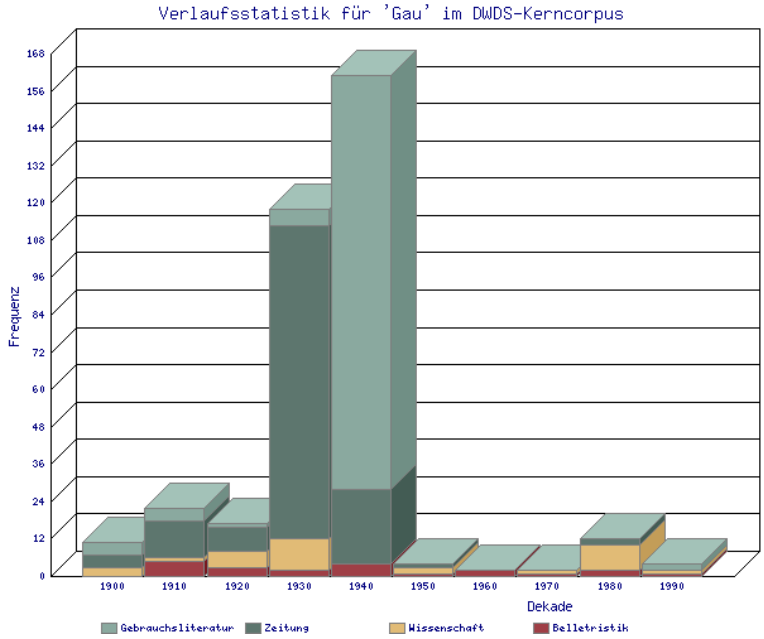

Abbildung 4: Zeitverlauf für Gau *im DWDS-Kernkorpus*

An diesem Beispiel werden zwei methodische Probleme einer zunächst rein statistischen Auswertung von Textkorpora deutlich: Einerseits erfordert die Interpretation der Rohdaten der Korpusanalyse ein hohes Maß an linguistischem Wissen (und an Weltwissen), andererseits liefern die Daten in vielen Fällen lediglich eine Bestätigung dessen, was lexikographisch längst festgeschrieben ist. Im Falle Gau kann immerhin festgehalten werden, dass das Referenzlexikon des DWDS, das Wörterbuch der Deutschen Gegen-

[6] „Gau, der; -(e)s, -e hist. in sich geschlossene Landschaft: Bei der Dürftigkeit des Ackerbaues wurde der Platz der Ansiedlung innerhalb des Gaues ... öfter gewechselt Mehring Dt. Geschichte 5; ich sei nur von einem Gau des alten Alemanniens in den andern hinüber ... gegangen G. Keller 4,480 (Gr. Heinrich)“. Quelle: Berlin-Brandenbrugische Akademie der Wissenschaften [Hrsg.] (2006). Projekt "Digitales Wörterbuch": online-Ressourcen. Wörterbuch der deutschen Gegenwartssprache. Online-Abfrage zu „Gau“, http://www.dwds.de/?woerterbuch=1&corpus=1&kompakt=1&qu=Gau&last_corpus=DWDS&old_corpus=DWDS&kw=on&sort=1&res=-1&cp=1, Zugriff Dezember 2006.

wartssprache (WDG) die Bedeutung des Akronyms nicht enthält bzw. aus zeitlichen Gründen gar nicht enthalten kann, wie das auf der Plattform des DWDS verfügbare Lemma zu Gau im WDG zeigt.

3.3 Zeitanalysen mit den Daten des Projektes Deutscher Wortschatz

Im Rahmen eines weiteren großen Korpusanalyseprojektes – des Leipziger „Projekt Deutscher Wortschatz" – sind in den vergangenen Jahren ebenfalls eine Reihe von Anwendungen entstanden, die deutsche Textkorpora nach temporalen Kriterien auswerten.

3.3.1 Wörter des Tages

Unter dem Motto „Wörter des Tages" werden seit 2002 mit ähnlichen Methoden wie bei der Tübinger Wortwarte täglich Online-Pressetexte gesammelt und durch den Vergleich mit einem großen Referenzkorpus ausgewertet. Grundlage ist hier das deutsche Korpus des Projektes *Deutscher Wortschatz*, das im Vergleich mit DEREKO oder DWDS zwar einen deutlich größeren Umfang (etwa 500 Millionen laufende Wortformen) aufweist, aber – wie Lemnitzer & Zinsmeister 2006:122f nicht ganz zu unrecht konstatieren – durch die weniger systematische Quellenauswahl eine schwerer nachvollziehbare Textgrundlage aufweist.

Anders als bei der Wortwarte ist hier nicht die Entdeckung von Neologismen das Ziel, sondern die Herausfilterung der zu einem Zeitpunkt jeweils besonders häufig gebrauchten Wörter und deren aktueller Kontext (Kollokationsmengen zu ausgewählten Begriffen). Es geht also weniger um die Änderung von Sprache durch Einführung neuer Begriffe, sondern um die Erfassung und Untersuchung des aktuellen Sprachgebrauchs in den Medien. Damit ist der Ansatz – soweit er sich „nur" auf die tagesgenaue Beobachtung beschränkt, wesentlich näher an der Methodik einer quantitativen Medienanalyse, da sich aktuelle Ereignisse unmittelbar in den entsprechenden Wortlisten niederschlagen, wie das nachfolgende (gekürzte) Beispiel für die Wörter des 12. September 2006 belegt:

Kategorie	***Wörter des Tages***
Sportler, Trainer, Funktionäre	Andy Roddick • Asamoah • Borowski • Briatore • Federer • Frings • Gerald Asamoah • Kimi Räikkönen • Magath • Michael Schumacher • Nationaltrainer • Roger Federer • Schumi
Sport	DFB-Pokal • FIFA • Formel 1 • Läufer • Marathon • US Open
Politiker	Biedenkopf • Bundesinnenminister Wolfgang Schäuble • Bundeskanzlerin Angela Merkel • Bundeswirtschaftsminister Michael Glos • CSU-Chef Edmund Stoiber • Castro • Djukanovic • Ismail Hanija […]
Organisation	Aldi • Atombehörde • BaFin • CNN • Chiphersteller • Dell • El Kaida • FIA • Fatah • Financial Times Deutschland • Frankfurter Allgemeinen Zeitung • Greenpeace • Hamas • Hisbollah • KfW […]
Ereignis	Einsturz • Gottesdienst • Hurrikan • Jahrestag • Lohnausgleich • Regensburg • Schweigeminute • Schweigeminuten • Terroranschlag •

	Terroranschläge • Trauerfeiern
Schlagwort	Anti-Terror-Kampf • Archivbild • Atomprogramm • Atomstreit • Auslastung • Dossier • Eckpunkte • Gammelfleisch • Gebet • Geburtshaus • Gedenken • Gesundheitsfonds • Gesundheitsreform […]
Ort	Altötting • Atlantis • Basilika • Beirut • Brunsbüttel • Cape Canaveral • Chelsea • Deggendorf • Golfstaaten • Ground Zero • Helsinki • Inn • Iran • Jemen • Karibik • Katar • Kühlhaus • Libanon […]
Personen aus Kunst, Kultur und Wissenschaft	Anna Nicole Smith • Madonna • Tom Tykwer • Tykwer
sonstige Personen	Attentäter • Bin Laden • Bloomberg • Florence • Gläubige • Heilige • Heilige Vater • Hofer • Jesus • Joseph Ratzinger • Karasek • Kirchenoberhaupt • Mittelständler • Muslime • Natascha • Oberhaupt • Osama bin Laden • Papst Benedikt XVI • Pilger • Pontifex • Regensburger • Ricke • Schlichter • Smith

Tabelle 2: Wörter des Tages (12.9.2006)[7]

Ähnlich wie bei der Tübinger Wortwarte greifen auch hier automatische und intellektuelle Auswahlprozesse ineinander: Ausgehend von einer automatischen Vorauswahl von im Vergleich mit dem Referenzkorpus auffallend häufig an einem Tag gebrauchter Wörter, findet eine intellektuelle Klassifikation statt, die diese Begriffsliste zu einem einfachen Raster zuordnet, das der Ressortaufteilung einer Zeitung nachempfunden ist. Dabei treten natürlich Inkonsistenzen auf (etwa die Zuordnung des geographischen Eigennamens *Regensburg* in die Kategorie *Ereignis* oder die Einordnung des Wortes *Chiphersteller* in die Kategorie *Organisation* statt in die Auffangkategorie *Schlagwort*), insgesamt ist aber bei fortlaufender Aufnahme immer weiterer Begriffe in die verschiedenen Klassen eine zunehmend bessere automatische Klassifikation gewährleistet. Daneben erfolgt auch eine semiautomatische Erkennung von Mehrwortgruppen im Bereich der Eigennamen: Auf der Basis von automatisch erzeugten – auch mehrstelligen – Kollokationsmengen zu allen Wörtern des Referenzkorpus werden Listen potentieller Mehrwortgruppen generiert, die nach intellektueller Durchsicht auch als Wörter in die Wortliste des Korpus aufgenommen werden (z.B. im obigen Beispiel *Kimi Räikkönen*, *Cape Canaveral* oder *Papst Benedikt XVI*).

Neben der Kategorisierung erfolgt auch eine Visualisierung der jeweiligen Wörter des Tages im Zeitverlauf der vergangenen vier Wochen. Dabei wird jeweils in einem Diagramm die relative Häufigkeit eines Wortes des Tages zusammen mit seinen am Stichtag stärksten Kollokationen angezeigt (Abb. 5).

---

[7] Quelle: Universität Leipzig, Institut für Informatik, Abt. Automatische Sprachverarbeitung (Hrsg.) (2006). Projekt Deutscher Wortschatz. Online-Abfrage Wörter des Tages 12. 9. 2006, http://wortschatz.uni-leipzig.de/wort-des-tages/2006/09/12/, Zugriff Dezember 2006.

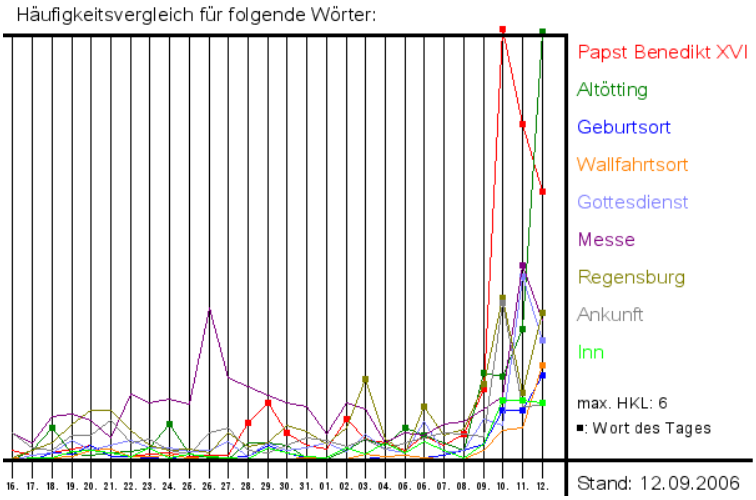

Abbildung 5: Zeitverlauf für Papst Benedikt XVI *(12. 9. 2006)*[8]

Diese Visualisierungsform macht unmittelbar deutlich, wie sich Ereignisse in den Medien spiegeln. Beispielsweise erzeugen in regelmäßigen Zeitabständen wiederkehrende Ereignisse in der Analyse vergleichbare Aktivitätsmuster (z.B. regelmäßig stattfindende Sportereignisse wie Formel 1-Rennen). Auch „mediale Erregungskurven" der politischen und gesellschaftlichen Berichterstattung lassen sich so nachzeichnen

3.3.2 Zeitbezogene Änderung von Gebrauchskontexten

Sind tagesgenaue Frequenzdaten zu Wörtern aus Online-Pressetexten verfügbar, so lassen sich auch über die reine Gebrauchsfrequenz hinausgehende Fragestellungen untersuchen. Geht man mit Firth 1957 davon aus, dass ein Wort durch seinen Kontext charakterisiert wird („a word is charaterised by the company it keeps", Firth 1957), liegt es nahe, über die Betrachtung von Auftretenshäufigkeiten hinauszugehen und zu untersuchen, inwieweit sich die Gebrauchs*kontexte* von Wörtern im Zeitverlauf ändern.

Auf der Basis der Daten des Projektes Deutscher Wortschatzinsbesondere der mittlerweile seit mehreren Jahren verfügbaren Tageskorpora[9] – wurde kürzlich in einer Regensburger Abschlussarbeit in

---

[8] Quelle: Universität Leipzig, Institut für Informatik, Abt. Automatische Sprachverarbeitung (Hrsg.) (2006). Projekt Deutscher Wortschatz. Online-Abfrage Verlaufsvisualisierung zu *Papst Benedikt XVI am 12.9.2006,* http://wortschatz.uni-leipzig.de/wort-des-tages/2006/09/12/Papst+ Benedikt+XVI.html, Zugriff Dezember 2006.

[9] Im Rahmen des Teilprojektes „Wörter des Tages" liegen dabei für jeden Tag der letzten Jahre strukturidentische Datenbanken mit Informationen über die Häufigkeit einzelner

Informationswissenschaft (Mendel 2006) untersucht, mit welchen Metriken sich Änderungen im Gebrauchskontext darstellen lassen. Neben der schon angesprochenen Frage nach der geeigneten Zeitgranularität stellt sich vor allem die Bestimmung eines geeigneten Vergleichsverfahrens für Kollokationsmengen als methodisches Problem heraus. Die Auswahl der Zeitgranularität wird dabei einerseits theoretisch durch die untersuchte Fragestellung (langfristiger Sprachwandel oder kurzfristige Phänomene der Darstellung von Konzepten in den Medien?) als auch von der praktischen Verfügbarkeit hinreichend vieler Daten für kleine Zeiteinheiten beeinflusst. Nach Analyse der vorliegenden Tages- und Jahreskorpora werden Tageskorpora zu Monatskorpora aggregiert, da der Monat als kleinste sinnvoll nutzbare Zeiteinheit erscheint.

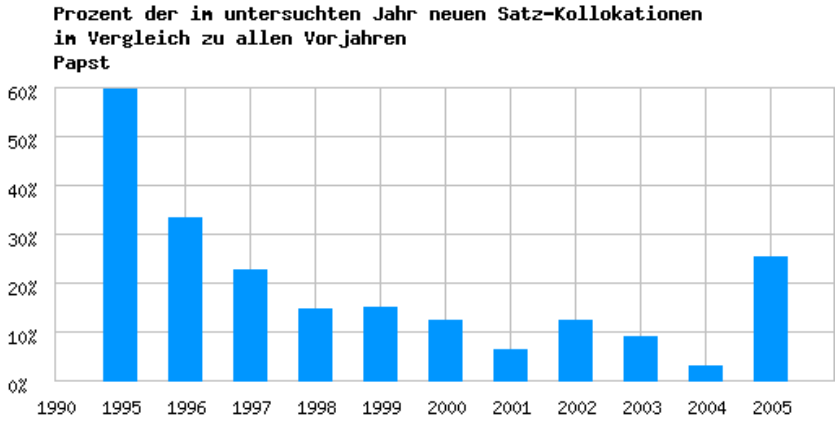

Abbildung 6: Unterschiede der Kollokationen eines Jahres zu Papst *im Vergleich mit allen Vorjahren*[10]

Für den Vergleich von Kollokationsmengen eines Begriffes aus unterschiedlichen Zeitscheiben werden sowohl Daten verwendet, die unmittelbare Nachbarschaft zweier Wörter im Text erfordern, als auch Kollokationen, die sich innerhalb eines Textfensters (Satzebene) ermitteln lassen. Eine Kollokationsmenge eines Wortes zum Zeitpunkt X kann sowohl mit der unmittelbar vorangegangenen Zeitperiode als auch mit der Aggregation *aller* vorangegangener Zeitperioden verglichen werden, wobei mengentheoretische Operationen zur Vergleichsberechnung herangezogen werden (Mendel 2006, 50ff). Mit Blick auf die Medienanalyse können damit unter-

Wörter in den täglich gesammelten Corpora sowie die automatisch ausgewerteten Kollokationen zu diesen Wörtern vor. Zur Berechnung der Wörter des Tages vgl. Quasthoff/Richter/Wolff 2002 und 2003.

[10] Quelle: Mendel, Ulrike (2006). Vergleich der Kollokationsmengen von Wörtern über mehrere Zeitabschnitte. Online-Abscfrageschnittstelle, Universität Regensburg, Institut für Medien-, Informations- und Kulturwissenschaft, http://www-alab.uni-r.de/~mendel/zeit-vergleiche.php, Zugriff Dezember 2006.

schiedliche Fragestellungen beantwortet werden: Bei lediglich lokalem Vergleich zweier Zeitperioden ergibt sich ein Indikator für das Ausmaß kurzfristiger Gebrauchsänderung in der medialen Verwendung eines Wortes; bei Vergleich mit aggregierten Vorperioden sieht man in der Regel eine fallende Kurve, die zeigt, dass mit der Zeit immer weniger neue Kollokate im Kontext eines Wortes zu beobachten sind. Um diese Daten besser interpretieren zu können, steht eine webbasierte Datenvisualisierung bereit, die die Änderungsraten zwischen Kollokationsmengen als Balkendiagramm darstellt (Abb. 6.).

Das Bild zeigt – erwartungsgemäß – für die Jahre 1996 – 2004 fallende Änderungsraten in den Kollokationsmengen und einen anschließenden deutlichen Ausschlag für das Jahr 2005, als sich durch die Wahl eines neuen Papstes auch der Kontext des Wortes *Papst* in den (online-)Medien deutlich änderte.

4 Fazit: Medienanalyse und Sprachwandel

Den voranstehenden Beispielen zeitbezogener Korpusvergleiche sind einige Merkmale gemeinsam:

- Sie operieren über vergleichsweise großen Textbeständen,
- sind online als Datenmaterial für die Forschung verfügbar und
- bieten nur vergleichsweise einfache Phänomene als zeitbezogenes Vergleichskriterium – insbesondere die Häufigkeit von Wörtern und Wortkontexten.

Sowohl für eine weitergehende sprachwissenschaftliche Betrachtung (Morphologie, Syntax, Textgrammatik etc.) als auch für die medienanalytische Textanalyse z.B. des typischen Aufbaus oder Argumentationsgangs in Online-Texten wäre eine weitergehende Aufbreitung und Annotation der Korpora erforderlich, die aber mit den heutigen Mitteln der Korpuslinguistik noch nicht oder nicht für sehr große Textkorpora zu leisten ist. Für die kommenden Jahre ist hier wenigstens für die linguistische Analyse Besserung in Sicht: Nicht nur werden automatische Verfahren der Textanalyse (*part of speech tagging*, morphologische Zerlegung, Syntaxanalyse) immer leistungsfähiger, auch die Standardisierung linguistischer Annotation beginnt sich mittlerweile abzuzeichnen (vgl. Declerck 2006).

Gleichzeitig ist es angesichts der vorhandenen *Massen*daten sinnvoll, die Betrachtung ausgewählter Einzelphänomene – wie auch in diesem Beitrag geschehen – durch die (automatische) Analyse des Gesamtdatenbestands zu ergänzen. Mit Hilfe geeigneter mathematischer Verfahren aus dem Bereich der Zeitreihenanalyse lassen sich typische Verlaufsmuster identifizieren und Wortgruppen bilden, die diesen Mustern entsprechen. Beispielsweise könnten damit Worte ermittelt werden, deren Verwendung besonders stark

zu- oder abnimmt oder deren Gebrauchskontext gleichen Veränderungen unterworfen ist.

Im Bereich der Medienwissenschaft steht man, was die Anwendbarkeit korpuslinguistischer Analysen angeht, einem doppelten Problem gegenüber: Zum einen ist die Medienwissenschaft eine junge Querschnittdisziplin (in diesem Sinne allenfalls vergleichbar der traditionellen Vorstellung einer methodenintegrierenden Philologie, vgl. Rusch 2002, 7), deren Methodenrepertoire noch in Entwicklung begriffen ist. Zum anderen finden sich quantitative Ansätze der Medienwissenschaft bisher vor allem im Bereich der Mediennutzungsforschung (vgl. etwa van Eimeren/Frees 2005) und des Medienmarketing bzw. der Werbeforschung (vgl. Unger 2005), weniger aber in dem hier angesprochenen Bereich der Medienanalyse.[11] Letztlich ist aber Medienanalyse – soweit es sich um die Betrachtung sprachlichen Materials handelt – von linguistischer Analyse nicht zu trennen. Man kann festhalten, dass im Spannungsfeld von Wirklichkeit, sprachlicher Repräsentation und medialer Vermittlung in der zeitbezogenen Auswertung von Sprachkorpora noch ein großes Potential liegt.

5 Literatur

Declerck, Thierry (2006): SynAF: Towards a Standard for Syntactic Annotation. In: Proc. LREC 2006, Fifth International Conference on Language Resources and Evaluation, Genua, Mai 2006.

Dipper, Stefanie et al. (2002): DEREKO (DEutsches REferenzKOrpus). German Reference Corpus. Final Report (Part I). Stuttgart: Institut für Maschinelle Sprachverarbeitung / Tübingen: Seminar für Sprachwissenschaft, Universität Tübingen, online verfügbar unter: http://www.sfs.nphil.uni-tuebingen.de/dereko/DEREKOReport.pdf, Zugriff Dezember 2006.

Eimeren, Birgit van, & Frees, Beate (2005): ARD/ZDF-Online-Studie 2005. Nach dem Boom: Größter Zuwachs in internetfernen Gruppen. In: Media Perspektiven Heft 8 (2005), 362-379.

Firth, John R. (1957): A synopsis of linguistic theory 1930-1955. In: Studies in Linguistic Analysis. Oxford, 1-32. Wiederabdruck in: Palmer, F.R. (ed.) (1968): Selected Papers of J.R. Firth 1952-1959. London.

Geyken, Alexander (2005): Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS). In: BBAW Circular 2005, Heft 32, 40.

Kilgarriff, Adam/Grefenstette, Gregory (2003): Introduction to the Special Issue on the Web as Corpus. In: Computational Linguistics 29(3) (2003), 333-347.

Klein, Wolfgang (2004): Vom Wörterbuch zum Digitalen Lexikalischen System. In: Zeitschrift für Literaturwissenschaft und Linguistik. 136 (2004), 10-55.

Lemnitzer, Lothar/Uhle, Tylmann (2006): Die Wortwarte - auf der Suche nach den Neuwörtern von morgen. Universität Tübingen, Seminar für Sprachwissenschaft,

[11] Dies wird auch in der einführenden Literatur in die Medienanalyse deutlich, vgl. Marcinkowski/Marr 2005, Mikos/Wegener 2005.

2006, http://www.sfs.uni-tuebingen.de/~lothar/nw/Projekt/index.html, Zugriff Dezember 2006.

Lemnitzer, Lothar; Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen.

Marcinkowski, Frank/Marr, Mirko (2005): Medieninhalte und Medieninhaltsforschung. In: Bonfadelli, Heinz/Jarren, Otfried/Siegert, Gabriele [Hrsg.] (2005²): Einführung in die Publizistikwissenschaft. Bern/Stuttgart/Wien (UTB Bd. 2170), 425-467.

Mendel, Ulrike (2006): Untersuchung von Bedeutungswandel mit Hilfe corpuslinguistischer Verfahren. Magisterarbeit, Informationswissenschaft, Universität Regensburg, Juni 2006 (für 2007 zur Veröffentlichung auf dem Dokumentenserver der Universität Regensburg vorgesehen, http://www.opus-bayern.de/uni-regensburg).

Mikos, Lothar/Wegner, Claudia [Hrsg.] Qualitative Medienforschung – ein Handbuch. Konstanz (UTB Bd. 8314).

Pereira, Fernando (2000): Formal grammar and information theory: Together again? In: Philosophical Transactions of the Royal Society, Vol. 358 (Nr. 1769), 1239-1253, April 2000 (Online verfügbar unter: http://citeseer.ist.psu.edu/pereira00formal.html, Zugriff Dezember 2006).

Quasthoff, Uwe/Richter, Matthias/Wolff, Christian. (2002): Wörter des Tages – tagesaktuelle wissensbasierte Analyse und Visualisierung von Zeitungen und Newsdiensten. In: Hammwöhner, Rainer/Wolff, Christian/Womser-Hacker, Christa [Hrsg.] (2002). Information und Mobilität. Proc. 8. Internationales Symposium für Informationswissenschaft, Universität Regensburg, Oktober 2002. Konstanz: UVK (Schriften zur Informationswissenschaft, Bd. 40), 369-372.

Quasthoff, Uwe/Richter, Matthias/Wolff, Christian. (2003): Medienanalyse und Visualisierung: Auswertung von Online-Pressetexten durch Text Mining. In: Seewald-Heeg, Uta [Hrsg.] (2003): Proc. 13. GLDV-Jahrestagung 2003, März 2003, HS Anhalt, Köthen. Sankt Augustin, 424-459 [seitengleich auch erschienen als: LDV-Forum 18(1,2) (2003), 452-459].

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

Rusch, Gebhard [Hrsg.] (2002): Einführung in die Medienwissenschaft. Wiesbaden.

Schlobinski, Peter [Hrsg.] (2006): Von *hdl* bis *cul8r*. Sprache und Kommunikation in den Neuen Medien. Mannheim/Leipzig/Wien/Zürich (Thema Deutsch, Bd.).

Unger, Fritz (2005): Mediaplanung – Voraussetzungen, Auswahlkriterien und Entscheidungslogik. In: Scholz, Christian [Hrsg.] (2005): Handbuch Medienmanagement. Berlin/Heidelberg, 735-760.