MICHAEL SCHARKOW

Lesen und lesen lassen – Zum State of the Art automatischer Textanalyse

Inhaltsanalytisch forschende Kommunikationswissenschaftler sehen sich in zunehmendem Maße mit den riesigen Informationsmengen im Internet konfrontiert, für die manuelle Codierungen kaum noch handhabbar sind. Die Verwendung automatischer Verfahren zur Textanalyse liegt nahe, da diese hohe Effizienz und Reliabilität versprechen. Schon lange vor der Verfügbarkeit von Online-Inhalten wurden automatische Verfahren für vielfältigste Fragestellungen, von Nachrichtenfaktorenforschung (SCHÖNBACH 1978) und Kampagnenberichterstattung (FAN 1997) über Analysen von Wahlprogrammen (LAVER/BENOIT/GARRY 2003) bis hin zur Auswertung offener Befragungsitems (KACZMIREK/BAIER/ZÜLL i.d.B.) in vielen Kontexten eingesetzt (vgl. ZÜLL/LANDMANN 2002). Trotzdem führt die computergestützte Inhaltsanalyse noch immer ein Schattendasein im Fach – die Methodenentwicklung scheint seit den 1980er-Jahren zu stagnieren. Die analytischen Möglichkeiten werden zudem häufig als begrenzt und wenig fruchtbar für den Forschungsalltag wahrgenommen (vgl. VAN CUILENBURG/KLEINNIJENHUIS/DE RIDDER 1988; FRÜH 2006). Die Entwicklung innovativer Verfahren der Textanalyse ist jedoch keineswegs stehen geblieben, sondern findet vermehrt in anderen Disziplinen statt, etwa der Politikwissenschaft und vor allem in der Informatik. In den letzten Jahren wurden sowohl computerlinguistische als auch statistische Verfahren für die Textklassifikation und Informationsextraktion rasant weiterentwickelt (vgl. MANNING/SCHÜTZE 1999; SEBASTIANI 2002). Dies wurde in den Sozialwissenschaften nur zum Teil wahrgenommen,

auch weil das Forschungsfeld stark fragmentiert und schwer überschaubar ist.

In diesem Beitrag soll der Versuch unternommen werden, die verschiedenen Ansätze zu systematisieren und zugleich Möglichkeiten und Grenzen der Verfahren zur automatischen Textanalyse[1] darzustellen. Nach einer kleinen Einführung in das Forschungsfeld wird in dieser Arbeit eine Typologie automatischer Textanalysen vorgestellt, an der sich die anschließende Vorstellung deskriptiver bzw. explorativer, deduktiver und induktiver Verfahren orientiert. Neben grundlegenden Begriffen soll dabei vor allem der Anwendungsbezug, d. h. mögliche Forschungsziele und Software für deren Umsetzung im Vordergrund stehen.

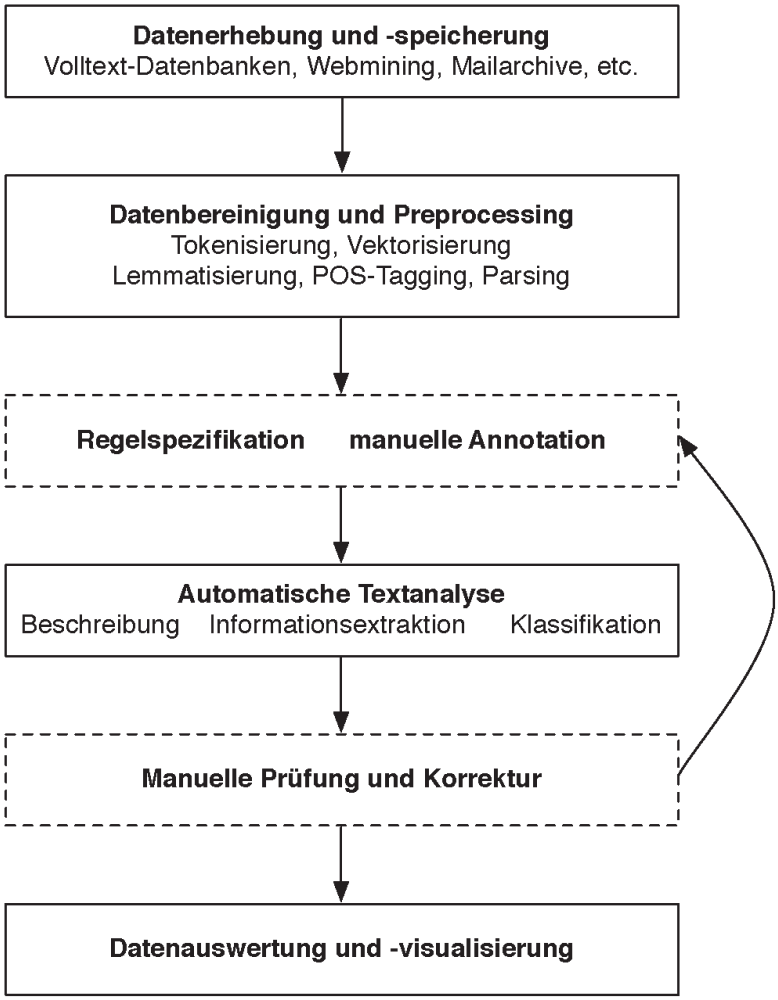
1. Grundlagen

1.1 *Begriffe und Forschungsprozess*

Seit der ersten Anwendung in den 1950er/1960er-Jahren hat sich das grundlegende Vorgehen bei automatischen Inhaltsanalysen nicht verändert, wohl aber die Rahmenbedingungen, unter denen die einzelnen Schritte der Datenerhebung und -codierung ablaufen (vgl. Abb. 1). Allein die Verfügbarkeit von Textmaterial hat sich in den letzten Jahrzehnten erheblich verbessert: sowohl offline als auch im Internet sind unüberschaubare Mengen digitaler und digitalisierter Texte vielfach kostenlos erhältlich, von klassischen Literaturkorpora (LEBERT 2005) über Volltext-Archiven von Zeitungen und Zeitschriften (SPIEGEL VERLAG 2007) bis zu Online-Nachrichten, Websites und E-Mail-Archiven. Aus dem Problem der Digitalisierung medialer Inhalte ist ein Problem des Datenmanagements, der Priorisierung und des Zugriffs auf relevante Informationen geworden, das sich nicht mehr nur mit manueller Arbeit, sondern mithilfe spezialisierter Software lösen lässt.

1 Der in der Litatur gängige Begriff der computergestützten Inhaltsanalyse (CUI) ist insofern unscharf, als so gut wie jede Inhaltsanalyse heute mithilfe von Computern durchgeführt wird, sei es bei der Datenerhebung durch CD-ROM- und Online-Datenbanken, bei der Codierung und Annotation von Texten, der Berechnung von Reliabilitätsmaßen und der Auswertung der Ergebnisse. Automatische Textanalyse liegt jedoch nur dann vor, wenn tatsächlich ein Computeralgorithmus zur Codierung verwendet wird.

ABBILDUNG 1
Ablaufschema automatischer Textanalyse



Nachdem der Text aus dem Internet oder anderen Quellen maschinenlesbar vorliegt, muss er zunächst bereinigt werden, etwa durch Entfernung von irrelevanten oder nicht-textuellen Inhalten. Zudem sollte er in ein standardisiertes Format, etwa ASCII oder Unicode-Text, HTML oder XML, umgewandelt werden (vgl. FEINERER/HORNIK/MEYER 2008). Anschließend werden die sogenannten *Features* aus dem Text extrahiert, in den meisten Fällen sind dies Wörter (Unigramme) bzw. Wortgruppen (N-Gramme) definierter Länge. Text ist in dieser Form nichts weiter als eine geordnete Abfolge von Zeichen bzw. ein *Bag-of-Words*[2]. Vor der

2 Die Verwendung von N-Grammen, also zwei- und mehrteiligen Wortgruppen, hat sich in vielen Studien als probates Mittel bewährt, die semantische Vielfalt natürlicher Sprache in

eigentlichen Analyse werden die Textdaten häufig mit linguistischen Verfahren vorbehandelt. So werden gebeugte Wörter durch Grundformen oder Wortstämme ersetzt, man spricht hierbei von Lemmatisierung bzw. *Stemming*. Weiterhin können Wortformen bestimmt (*Part-of-Speech-Tagging*), häufig vorkommende Wörter entfernt oder Synonyme und Anaphora aufgelöst werden (vgl. HOTHO/NÜRNBERGER/PAASS 2005). Der Nutzen einer solchen Vorbehandlung liegt in der Reduktion der Feature-Anzahl, die sich schon bei einfachen Texten im fünfstelligen Bereich bewegen kann. Sie ist aber nicht unumstritten, da die syntaktische und semantische Varianz des Textes unter Umständen erheblich verringert wird und darunter ggf. die Validität der automatischen Codierung leiden kann (BRASCHLER/RIPPLINGER 2004).

Als Ergebnis liegen die Texte zumeist als Dokument-Term-Matrix vor, d. h. einem Datensatz, in dem die Zeilen Dokumenten entsprechen, die Spalten den extrahierten Features und die einzelnen Zellen Informationen zum Vorkommen und ggf. der Häufigkeit eines Terms in einem Dokument enthalten. Dieses sogenannte Vektor-Raum-Modell für Texte erlaubt mit geringem Aufwand eine Vielzahl statistischer Analysen, die sich teilweise mit gängigen Programmen wie SPSS oder R auswerten lassen (vgl. MANNING/SCHÜTZE 1999; FEINERER et al. 2008).

Während sich die ersten beiden der in Abbildung 1 dargestellten Schritte für die verschiedenen Verfahren praktisch nicht unterscheiden, variieren die darauf folgenden Schritte des Forschungsprozesses je nach Textanalysemethode erheblich. Es lohnt sich daher, diese hinsichtlich ihrer Voraussetzungen, ihrer Zielsetzung und Analyselogik zu klassifizieren.

1.2 *Typologie der Verfahren automatischer Textanalyse*

Der Versuch, automatisierte inhaltsanalytische Verfahren zu systematisieren, ist keineswegs neu: So unterscheidet Roberts (1997) zwischen thematischen, semantischen und netzwerkbasierten Analysen von Texten, Shapiro (1997) hingegen betrachtet vor allem die Frage, ob automatische

relativ simple statistische Verfahren zu überführen (vgl. YERAZUNIS 2003). Der häufige Einwand gegenüber automatischer Textanalyse, es würde nur auf lexikalischen (Wort-)Ebene codiert, lässt sich so zumindest abschwächen (vgl. FRÜH 2006).

Inhaltsanalysen entweder instrumentell zur Prüfung spezifischer Hypothesen oder generisch zur systematischen Strukturierung von Texten dienen können. In dieser Arbeit soll eine weitere Typologie vorgestellt werden, die in Teilen den oben genannten ähnelt (vgl. auch MERTEN 1995), aber die Anforderungen und Einflussmöglichkeiten des Forschers und des Computers im Forschungsprozess als Maßstab nimmt.

ABBILDUNG 2
Typologie von Verfahren automatischer Textanalyse

	Unüberwachte Verfahren		Überwachte Verfahren	
	Deskriptiv	Explorativ	Deduktiv	Induktiv
Bag-of-Words-Ansätze	Textstatistik Stilometrie	Co-Occurrence/LSA Dokumentclusterung	Diktionärbasierte Verfahren	Überwachte Klassifikation
Syntaktisch-semantische Ansätze			Regelbasierte Verfahren	Induktive Informationsextraktion

Grundsätzlich lassen sich unüberwachte explorative und überwachte hypothesengeleitete Verfahren automatischer Textcodierung unterscheiden (vgl. Abb. 2).[3] Schon auf der Annenberg-Konferenz Ende der 1960er-Jahre, der ersten großen Tagung zu diesem Thema, war dies eine Grundsatzfrage: theoriefreie Exploration von Worthäufigkeiten versus hypothesengeleitete Kategorisierung von Dokumenten (STONE 1997: 41). Für erstere ist keine aufwändige und kostenintensive Regelspezifikation bzw. manuelle Annotation der Texte sowie Prüfung und Korrektur (vgl. Abb. 1) notwendig – es kann sofort mit der automatischen Analyse begonnen werden. Unüberwachte deskriptive bzw. explorative Textanalysen sind dementsprechend mit dem geringsten Aufwand verbunden, da sich der gesamte Forschungsprozess vollautomatisch durchführen lässt. Möglichkeiten und Grenzen dieser Verfahren werden nächsten Abschnitt näher vorgestellt.

Für die sozialwissenschaftliche Inhaltsanalyse ist die Vorgabe von Codierschemata durch den Forscher oft essentiell, da es in den meisten Fällen um spezifische Dimensionen eines Textinhaltes geht und diese für den Codierprozess operationalisiert werden müssen. Hier sind überwach-

3 Im Bereich des Maschinenlernens spricht man dann von unüberwachtem Lernen, wenn ein Algorithmus ohne Feedback von außen Klassifikationsentscheidungen treffen muss. Überwachtes Lernen liegt dann vor, wenn von außen Kriterien über richtig/falsch vorgegeben werden, anhand derer der Algorithmus ›lernen‹ kann.

te Verfahren notwendig, bei denen der Forscher der Software Regeln oder Beispiele vorgibt, nach denen dann die automatische Analyse durchgeführt wird. Die Kosten für eine solche (halb)automatische Lösung hängen dementsprechend von der Regelspezifikation oder manuellen Codierung von Texten ab, es muss also ein Kompromiss zwischen Aufwand und Umfang der manuellen Vorarbeiten gefunden werden.

Aufgrund ihrer großen Bedeutung für die angewandte Inhaltsanalyse sind überwachte Verfahren ein Schwerpunkt der Methodenentwicklung. Hier lassen sich sowohl die meisten ›klassischen‹ als auch neuere Ansätze verorten. Daher lohnt es sich, die vorhandenen Verfahren überwachter Codierung zu systematisieren, die auf der rechten Hälfte von Abbildung 2 dargestellt sind. Hierbei kann man sich an zwei Dimensionen orientieren: deduktives vs. induktives Vorgehen und rein wortbasierte vs. syntaktisch-semantische Ansätze.

Die horizontale Dimension der Typologie in Abbildung 2 differenziert zwischen einem deduktiven Vorgehen, d. h. der Forscher stellt explizite Regeln auf, nach denen klassifiziert wird, und den neueren induktiven Ansätzen, bei denen einem lernenden Algorithmus Beispieltexte und deren korrekte Codierung vorgegeben werden. Die Regeln, nach denen codiert wird, werden von der Software aus den Beispielen extrahiert.

Das erstgenannte deduktive Vorgehen ist bislang die dominante Praxis bei der automatischen Textanalyse, allerdings führt dies dazu, dass manuelle und automatische Codierung stark voneinander abgekoppelt sind. Eine manuelle Codiererschulung ist in den meisten Fällen beispielbasiert, da viele anspruchsvolle Konstrukte nach wiederholtem Üben richtig codiert werden können, aber selten explizite und umfassende Regeln (etwa, wann ein Presseartikel negativ über einen Wahlkandidaten berichtet) vorliegen. Grundsätzlich erfordern also deduktive Verfahren mehr konzeptionelle Vorarbeit vom Forscher, während induktive Verfahren vor allem auf viele und gut annotierte Beispieltexte angewiesen sind. Die Entwicklung inhaltsanalytischer Instrumente profitiert dabei von der gewachsenen Rechenleistung moderner Computer, indem inkrementell, d. h. nach jeder Änderung der Regeln oder zusätzlichen Annotationen, getestet werden kann, da die automatische Codierung selbst mit großen Mengen an Texten und/oder Kategorien nur Minuten dauert. Der Feedback-Prozess in Abbildung 1 ist daher heute erheblich kürzer als zu Zeiten, als Lochkarten in Kisten zum Großrechner gebracht werden mussten und die Codierung über Nacht lief (STONE 1997).

Die vertikale Dimension in Abbildung 2 spiegelt die syntaktisch-semantische Tiefe der Textanalyse wieder: Wird bei rein wortbasierten Verfahren jedes N-Gramm als eine einfache, isolierte Variable aufgefasst, versuchen eher aussagebasierte Verfahren Beziehungen zwischen semantischen Einheiten zu erfassen. Ein Satz wie »Peter sieht das Haus.« kann also einerseits als Menge von vier Unigrammen und drei Bigrammen betrachtet werden, andererseits als gerichteter Graph ›Peter‹ (Subjekt) – ›sehen‹ (Verb) – ›Haus‹ (Objekt). Traditionell versuchen wortstatistische Verfahren Dokumente zu klassifizieren (ein Dokument handelt von Peter, wenn das Wort ›Peter‹ vorkommt), während stärker linguistische Verfahren auf Aussageebene Beziehungen analysieren wollen (die Entität ›Peter‹ vollzieht die Handlung ›sehen‹ im Bezug auf die Entität ›Haus‹) und damit Antworten auf offene Fragen zum Text geben sollen. Man muss kein Linguist sein, um nachzuvollziehen, dass syntaktisch-semantische Ansätze deutlich anspruchsvoller, aber auch schwieriger bzw. gar nicht vollautomatisch umzusetzen sind, und die Ergebnisse bislang zumeist unbefriedigend blieben (vgl. VAN CUILENBURG et al. 1988; SHAPIRO 1997).

2. Unüberwachte deskriptive und explorative Verfahren

2.1 *Textstatistik*

Die Berechnung von Text-, Satz- und Wortstatistiken ist eines der ältesten und einfachsten Verfahren automatischer Textanalyse. Ein einfaches Beispiel ist die Auszählung der häufigsten Wörter eines Dokuments, mit der sich die wichtigsten Konzepte eines Textes zusammenfassen lassen, wie dies etwa beim automatischen *Tagging* von Blog-Einträgen geschieht (vgl. BROOKS/MONTANEZ 2006). Auch bei der Auswahl relevanter Wörter für weitere Analysen ist die gewichtete Worthäufigkeit *Term-Frequency-Inverse-Document-Frequency* (TF-IDF; SALTON 1989) ein zentrales Kriterium.

Obwohl die Beschreibung von Texten durch Häufigkeiten und Mittelwerte auf den ersten Blick trivial erscheint, können doch verschiedene interessante und wissenschaftlich relevante Konzepte mit textstatistischen Maßen operationalisiert werden. So lässt die durchschnittliche Länge (Wörterzahl) pro Artikel zum Beispiel Schlüsse auf Genre, Medi-

um oder journalistische Arbeitspraxis zu. Zudem ist die Beitragslänge ein entscheidender Indikator für den Nachrichtenwert einer Meldung (SCHULZ 1976). Im Bereich der Korpuslinguistik und Stilometrie werden darüber hinaus zahlreiche textstatistische Maße auf dem Gebiet der Autorschaftsforschung eingesetzt (GRIEVE 2007), derer sich in jüngster Zeit auch Sicherheitsbehörden bei der Überwachung von Online-Kommunikation bedienen (ABBASI/CHEN 2005).

Auch die Komplexität und Lesbarkeit eines Textes lässt sich durch Wort- und Satzlänge oder Umfang des benutzten Vokabulars erkennen (DUBAY 2004; BEST 2006). Lesbarkeitsmaße, die sich aus mehreren solcher Indikatoren zusammensetzen, werden häufig in der Bildungs- und Usability-Forschung eingesetzt. Obwohl die Validität und Genauigkeit solcher Indizes nicht unumstritten sind, eignen sie sich doch für längs- oder querschnittlich vergleichende Analysen von Texten. Online-Angebote, die auf unterschiedliche Alters- oder Bildungsgruppen zielen, sollten sich dementsprechend auch hinsichtlich ihrer Lesbarkeit unterscheiden (TAYLOR 2008).[4]

Viele grundlegende textstatistische Funktionen sind in gängigen Textverarbeitungsprogrammen und -editoren integriert. Einen größeren Funktionsumfang bieten die Website www.readability.info sowie das Freeware-Programm TEXTSTAT. Diese berechnen auch verschiedene Lesbarkeitsindizes wie den Flesch- oder Gunning-Fog-Index (TAYLOR 2008).

2.2 *Co-Occurrence und Latente Semantische Analyse*

Während die deskriptive Textstatistik den eigentlichen Inhalt eines Dokuments vollständig ignoriert, kann mit der *Co-Occurrence*-Analyse explorativ die syntaktische und semantische Vielfalt eines Textes reduziert und analysiert werden. Dabei wird davon ausgegangen, dass inhaltlich zusammengehörende Wörter oder Wortgruppen auch im Text nahe beieinander liegen. Betrachtet man die Wörter innerhalb eines spezifizierten Rahmens, etwa in kompletten Sätzen oder Absätzen, lässt sich das gemeinsame (Nicht-)Auftreten bestimmter Begriffe in eine Ähnlichkeits-

4 Dies zeigt sich tatsächlich, wenn man zum Beispiel die Lesbarkeitswerte der Startseiten von www.bravo.de (Flesch-Index 82) und www.heise.de (Flesch-Index 47) vergleicht, wobei nach der Flesch-Formel Werte von 0 (schwer) bis 100 (leicht) möglich sind (TAYLOR 2008).

matrix überführen (vgl. ZÜLL/ALEXA 2001). Anhand dieser Matrix lassen sich konventionelle hierarchische Clusteranalysen durchführen, die eine große Menge an Wörtern zu wenigen latenten semantischen Einheiten verdichten. Mit einer anschließenden multidimensionalen Skalierung lassen sich die semantischen Cluster sowie deren Positionierung zueinander visualisieren (LANDMANN/ZÜLL 2004).

Eng mit der Co-Occurence-Analyse verwandt ist eine Gruppe von faktoranalytischen Verfahren, die Latente Semantische Analyse oder Indizierung (LSI), die ebenfalls der Verdichtung von großen Feature-Vektoren dienen, um mit diesen reduzierten Faktoren analytisch weiterzuarbeiten. Sie basieren grundsätzlich auf der Logik der Einzelwertzerlegung von Matrizen, die auch der Hauptkomponentenanalyse zugrunde liegt (MANNING/SCHÜTZE 1999). Dieses Verfahren setzen beispielsweise Simon und Xenos (2005) für die Verdichtung und Interpretation verschiedener latenter Antwortkategorien bei offenen Fragen ein. Stephen (2006) analysiert mit einer Co-Occurrence-Analyse die Titel von kommunikationswissenschaftlichen Zeitschriftenartikeln und deckt dabei verschiedene Themencluster auf.

Der Nachteil dieser explorativen Verfahren liegt zweifellos in der Tatsache, dass die inhaltliche Ausrichtung der gemessenen latenten semantischen Einheiten a priori unbekannt und damit nicht aus theoretischen Überlegungen herleitbar ist. Wie bei einer explorativen Faktoren- oder Clusteranalyse kann höchstens die Anzahl der Cluster bzw. Faktoren vom Forscher bestimmt werden, und oft sind die rein statistisch gebildeten Textdimensionen inhaltlich nicht sinnvoll interpretierbar, wie Landmann und Züll (2004) bei ihrer Evaluation des Programms CATPAC feststellen. Grundsätzlich eignen sich daher Co-Occurrence-Verfahren ähnlich wie Häufigkeitszählungen vor allem dafür, schnell einen Überblick über eine potentiell unüberschaubare Menge von Textdaten zu gewinnen und ggf. das für spätere Analysen verwendete Feature-Set zu reduzieren. Als Variablen dienen dann nicht mehr einzelne Wörter oder Wortgruppen, sondern abstrakte semantische Einheiten, die im Prinzip latenten Konstrukten entsprechen.

2.3 *Dokumentclusterung*

Die unüberwachte Klassifikation von Dokumenten ist eine klassische Anwendung im Bereich der statistischen Textanalyse. Basierend auf einer

ggf. reduzierten und semantisch verdichteten Dokument-Term-Matrix, in denen jedes Dokument einen Fall darstellt, werden wiederum clusteranalytische Verfahren angewandt, um syntaktisch ähnliche Dokumente zu gruppieren, man spricht daher auch von *Document Clustering*.

Der am häufigsten verwendete Cluster-Algorithmus ist hierbei *k*-Means, bei dem im Gegensatz zur hierarchischen Clusteranalyse die Zahl der Cluster *k* a priori festgelegt wird (HOTHO et al. 2005; MANNING/SCHÜTZE 1999).[5] Alternativ zu partitionierenden Verfahren lassen sich auch hierarchische Clusteranalysen für die Textklassifikation einsetzen, insbesondere wenn die impliziten Zielkategorien ohnehin in einer entsprechenden thematischen Ontologie, d.h. zumeist einem Kategorienbaum, vorliegen. Auf diese Weise können Brooks und Montanez (2006) beispielsweise eine Themenhierarchie aus Blogeinträgen und deren Technorati-Tags[6] ableiten. Allerdings sind hierarchisch-agglomerative Verfahren sehr rechenintensiv, was ihre Anwendungsmöglichkeiten auf sehr große Datensätze, mit denen man es im Internet häufig zu tun hat, stark einschränkt.

Wie schon bei den zuvor dargestellten Verfahren unterliegen die Kategorienbildung und damit die Klassifikation der Dokumente nicht dem Einfluss des Forschers. Es ist dementsprechend notwendig, die entstandenen Cluster inhaltlich zu interpretieren und zu benennen. Häufig ergeben sich zumindest einige inhaltlich plausible Themencluster, die sich ggf. in einem zweiten Schritt für die hypothesengeleitete Klassifikation von Texten verwenden lassen. Trotz einiger Nachteile und Einschränkungen ist die automatische Dokumentklassifikation eines der meistgenutzten Verfahren bei der Kategorisierung von Online-Inhalten. Fast alle Suchmaschinen und Web-Verzeichnisse bedienen sich clusteranalytischer Verfahren, um die große Zahl an Web-Dokumenten sinnvoll zu strukturieren. Aus diesem Grund ist die Literatur zu diesem Thema sehr umfangreich und die Methodenentwicklung schreitet rasant voran (vgl. BRODER/GLASSMAN/MANASSE/ZWEIG 1997). Ähnlich wie bei der Co-Occurrence-Analyse können auch Dokumentencluster in einem zusätzli-

5 Als Ähnlichkeitsmaß wird zumeist der Kosinus oder Jaccard-Index zwischen zwei Dokumentvektoren verwandt. Weniger gut eignet sich die euklidische Distanz, da das gemeinsame Nichtauftreten von Wörtern bei einer großen Zahl an Features extrem häufig ist und daher kaum Informationswert hat (STREHL/GOSH/MOONEY 2000).

6 Für die Blog-Suchmaschine www.technorati.com können Nutzer beliebig Schlagwörter (Tags) vergeben, anhand derer dann Blog-Einträge gesucht und aggregiert werden.

chen Analyseschritt zueinander positioniert und visualisiert werden (DI GIACOMO/DIDIMO/GRILLI/LIOTTA 2007).

Automatische Dokumentclusterung wird von den meisten kommerziellen Daten- und Textmining-Programmen wie SPSS LEXIQUEST beherrscht, ist aber auch im Rahmen des tm-Paketes für das Textmining in R (Open Source) leicht zu realisieren. Einen sehr guten Überblick und Beispiele bieten Feinerer et al. (2008).

Insgesamt zeichnen sich die hier vorgestellten unüberwachten Verfahren durch eine perfekte Reliabilität aus, da sämtliche Codierprozesse automatisch ablaufen. Die einzige Gefahr liegt in dieser Hinsicht in Programmfehlern, die zu – allerdings reproduzierbaren – Codierfehlern führen. Die Frage nach der Validität der Codierung hängt davon ab, was eigentlich mit dem Verfahren gemessen werden soll: sicher sind hochgradig valide Inferenzen über die Textlänge oder die Reichhaltigkeit des Vokabulars möglich, doch auf der semantischen und pragmatischen Ebene sieht die Situation anders aus. Nicht jedes textuelle Verständlichkeitsmaß hat eine gleich hohe prädiktive Validität (KERCHER 2008), und inwiefern das gemeinsame Auftreten von Wörtern tatsächlich semantische Konzepte valide misst, hängt stark vom analysierten Textmaterial und der Wahl des Analyserasters (Satz, Absatz, Text) ab. Interpretationen durch den Forscher sind hier unerlässlich.

3. Deduktive Verfahren

3.1 *Diktionärsbasierte Verfahren und Freitextsuche*

Seit der Frühzeit computergestützter Inhaltsanalyse stellen diktionärsbasierte Verfahren das wichtigste und lange Zeit auch einzige Mittel hypothesengeleiteter, deduktiver Textcodierung dar (STONE 1997). Es verwundert daher nicht, dass dieser Ansatz den meisten empirischen Studien zugrunde liegt und bis heute fast synonym mit automatischer Textanalyse verwendet wird (vgl. ZÜLL/ALEXA 2001; ZÜLL/LANDMANN 2002; KACZMIREK/BAIER/ZÜLL i.d.B.). Die Grundlogik diktionärbasierter Verfahren ist dabei seit Jahrzehnten unverändert: Vor der eigentlichen Codierung wird vom Forscher ein Kategoriensystem entwickelt, bei dem jeder Klasse einzelne Wörter bzw. Wortstämme zugewiesen werden, die als Indikatoren für das interessierende Konstrukt dienen. Die Analysesoftware kann

dann problemlos nach den Wortstämmen suchen und die sie enthaltenden Dokumente entsprechend klassifizieren.

Auf dieselbe Logik stützt sich auch das Verfahren der Freitextrecherche in Datenbanken oder Online-Suchmaschinen, das mittlerweile in der Kommunikationswissenschaft recht weit verbreitet ist (HAGEN 2001; HOLLANDERS/VLIEGENTHART 2008). Hier werden spezielle Suchanfragen zu definierten Begriffen, oft mit Boolschen Operatoren verknüpft, an externe Dienste gestellt, die dann ein Ergebnisse mit passenden Dokumenten zurückliefern, etwa LexisNexis für digitalisierte Medieninhalte oder Google für das Internet allgemein. Die Freitextsuche hat vor allem den Vorteil, dass sie kostengünstig und schnell zu realisieren ist, allerdings mit dem Nachteil, dass die Grundgesamtheit an Dokumenten dem Forscher nicht vorliegt und oft keine Informationen zur Menge der *nicht* gefundenen Beiträge vorliegt, die Abfragevalidität mithin unklar ist (HAGEN 2001; WELKER/WERNER/SCHOLZ 2005: 51ff.). Eine weitere Einschränkung liegt in der Tatsache, dass man bei der Konstruktion der Suchanfragen auf die technischen Möglichkeiten der Betreiber angewiesen ist und meist nur UND/ODER-Verknüpfungen der Begriffe möglich sind.

Obwohl diese rein wortbasierte Codierung auch den meisten klassischen Diktionärstudien zugrunde liegt, lassen sich bei vorliegenden Volltexten auch komplexere Codierregeln entwickeln. Mittels sogenannter regulärer Ausdrücke (FRIEDL 2006), die in den meisten Programmiersprachen und in vielen Texteditoren implementiert sind, lassen sich beispielsweise auch Regeln definieren, wonach bestimmte Begriffe innerhalb eines Satzes oder in einer definierten Reihenfolge in einem Dokument vorkommen müssen, damit dieses codiert wird. Die Definition von komplexen Suchabfragen und regulären Ausdrücken ist eine Wissenschaft für sich, lohnt sich aber im Hinblick auf die deutlich größere Flexibilität bei der Codierung.

Da die Codierung von Dokumenten nach Stichwörtern ein deterministischer Prozess ist, kann für diktionärbasierte Verfahren vollständige Reliabilität angenommen werden. Der Nachteil einer solch simplen Analysestrategie liegt jedoch in der oftmals geringen Validität der Ergebnisse, wenn es um kommunikationswissenschaftlich interessante theoretische Konstrukte geht, die mit der Inhaltsanalyse gemessen werden sollen (vgl. FRÜH 2006). Während der diktionärbasierte Ansatz für spezielle Begriffe, etwa Eigen- oder Markennamen im Rahmen einer Medienresonanzanalyse, mit geringem Aufwand zu validen Ergebnissen kommt, gestaltet sich

die wortbasierte Codierung komplexer Konstrukte zunehmend schwierig. Da aber die meisten Fragestellungen nicht auf Wort-, sondern thematischer Ebene vorliegen, wird sie allein durch die Existenz von Rechtschreibfehlern und Homonymen weniger valide Ergebnisse produzieren. Die bekannten Harvard- und Lasswell-Diktionäre (DUNPHY/BULLARD/ELINOR 1974; LASSWELL/NAMENWIRTH 1969), mit denen auch verbalisierte Emotionen, Themen und Bewertungen codiert werden können, sind die Ergebnisse jahrelanger Arbeit, die sich nicht in jedem Fall lohnen wird. Zudem ist für viele Phänomene nicht ohne weiteres eine Wortliste ersichtlich, die tatsächlich zuverlässig und valide zwischen den Kategorien differenzieren kann. Als Beispiel sei nur die häufig codierte Personalisierung in der Medienberichterstattung genannt (vgl. WILKE/REINEMANN 2001). Der Aufwand für die Erstellung eines guten Diktionärs ist in vielen Fällen höher als bei der manuellen Codierung, sodass der Nutzen diktionärsbasierter Textanalyse erst mit steigender Dokumentenzahl zunimmt. Dann allerdings ist sie außerordentlich effektiv und effizient.

Obwohl wortlistenbasierte Verfahren sich auch ohne spezielle Software relativ leicht umsetzen lassen, etwa mit klassischen UNIX-Tools (vgl. SCHMITT/CHRISTIANSON/GUPTA 2007), gibt es eine große Auswahl an Textanalyse-Programmen wie TEXTPACK, GI oder DICTION, die diese Verfahren beherrschen und teilweise eigene Diktionäre mitbringen. Einen Überblick über kommerzielle Programme bieten Alexa/Züll (2000) und Lowe (2002). Besonders erwähnenswert ist zudem YOSHIKODER (LOWE 2006), das plattformunabhängig, Open Source und mehrsprachig ist, d. h. im Gegensatz zu den meisten Alternativen auch die Codierung chinesischer oder arabischer Texte ermöglicht.

3.2 *Regelbasierte Verfahren*

Während Diktionäre vor allem geeignet sind, Dokumente in zuvor definierte Kategorien einzuordnen, dienen regelbasierte Verfahren zumeist der Informationsextraktion. Auf Satz- oder Aussageebene soll so der syntaktische und semantische Gehalt eines relativ unstrukturierten natürlichsprachlichen Textes in eine Graphen- oder Baumstruktur von Subjekt-Objekt-Prädikat-Beziehungen umgewandelt werden, die sich dann automatisiert analysieren lässt (vgl. KING/LOWE 2003; ROBERTS 1997). Vereinfacht ausgedrückt: Regelbasierte Verfahren geben Auskunft darüber,

worum es in einer Aussage geht, sie beantworten quasi offene Fragen. Dazu ein stark vereinfachendes Beispiel: Man stelle sich eine lange Reihe von Agentur-Tickermeldungen vor, die immer die gleiche Form haben:

Merkel lädt SPD-Führung zum Gespräch. Bahn-Mitarbeiter fordern 10 Prozent mehr Lohn. Hertha BSC kauft brasilianischen Stürmer.

Mit der primitiven und sicher unzureichenden Regel, dass das erste kleingeschriebene Wort das Prädikat, alles davor das Subjekt und alles danach das Objekt ist, könnte zügig eine Liste mit Akteuren erstellt werden, deren Handlungen besonders oft berichtet werden, welche Handlungen dies sind, wen sie betreffen usw. Als Rahmen für weitere Analysen wird dann in vielen Fällen die relationale Inhaltsanalyse verwendet, die Beziehungen von Akteuren und deren Interaktionen auswertet (vgl. CARLEY 1997; ADAM 2008). Wie van Atteveldt (2008) treffend feststellt, ist eine solche semantische Analyse von Aussagen sehr viel schwerer zu automatisieren als thematische Klassifikationen. Lange Zeit überwog in den Sozialwissenschaften deshalb die Skepsis, ob linguistisch tiefergehende Analyseverfahren mit dem Computer überhaupt machbar seien, und man nicht mit der computergestützten Annotation von Aussagen zufrieden sein müsse (VAN CUILENBURG et al. 1988).

Trotz aller Schwierigkeiten wurde das Verfahren der Informationsextraktion in der Politikwissenschaft, und dort insbesondere bei der Analyse internationaler Ereignisse, seit Beginn der 1990er-Jahre erfolgreich angewandt und weiterentwickelt. So konnten mit dem diktionärs- und regelbasierten System KEDS/TABARI (SCHRODT/DAVIS/WEDDLE 1994) erfolgreich die Schlagzeilen des Reuters-Tickers hinsichtlich internationaler Konflikte und anderer Ereignisse codiert werden. Das von King und Lowe (2003) äußerst positiv evaluierte Tool VRA arbeitet auf Basis eines vollständigen Satz-*Parsings,* mit dem die syntaktische Struktur der Schlagzeilen zerlegt und durch zusätzliche Kontextinformationen auf Basis von Konkordanzen und Wortlisten ergänzt wird. Dadurch wird, so die Autoren, eine mit menschlichen Codierern vergleichbare oder langfristig sogar bessere Reliabilität erreicht. Einschränkend muss jedoch konstatiert werden, dass die genannten Verfahren sehr domänenspezifisch (internationale Ereignisse) und anhand einfacher und relativ stark strukturierter Texte (englische Schlagzeilen) eingesetzt werden. Zudem ist wie bei der Diktionärsentwicklung Expertenwissen für die Definition der Parsing-Regeln und die Erkennung von Akteuren erforderlich (vgl. SCHRODT et al. 1994).

Zwei Entwicklungen haben allerdings in jüngster Zeit dazu beigetragen, dass automatische semantisch-relationale Textanalysen einfacher und damit auch häufiger eingesetzt werden: Zum einen werden im Bereich des *Natural Language Processing* immer mächtigere Parser und andere Algorithmen entwickelt, die syntaktisch-semantische Analysen ermöglichen. Diese sind mittlerweile auch für andere Sprachen als Englisch verfügbar und liefern dort sehr ermutigende Ergebnisse, selbst bei umfangreicheren Texten (vgl. VAN ATTEVELDT 2008). Zum anderen liegen Inhalte im Internet in vielen Fällen sehr viel strukturierter vor als in Offline-Medien, da sie ja für die digitale Verarbeitung gedacht sind. So lässt sich das Ziel eines Hyperlinks mit einem HTML-Parser deutlich leichter extrahieren als das Objekt einer Aussage in einem Satz. Der Urheber einer Email ist durch einen festen Platz im *Mail-Header* ebenso problemlos identifizierbar wie die Nachricht, auf die sie sich bezieht. Die Strukturierung von Informationen in HTML/XML und anderen Formaten erleichtert syntaktische Analysen und besonders die relationale Inhaltsanalyse von Kommunikationsnetzwerken enorm (vgl. RUCHT/YANG/ZIMMERMANN 2009). Die Forschungsliteratur zu sozialen Online-Netzwerken ist in den letzten Jahren nicht zuletzt aufgrund der relativ unproblematischen automatisierten Datenerhebung und -analyse geradezu explodiert, es sei daher an dieser Stelle stellvertretend auf Stegbauer und Rausch (2006) sowie Berendt, Schlegel und Koch (2008) verwiesen. Letztere benutzen für ihre Analyse der deutschsprachigen Blogosphäre sowohl diktionär- als auch regelbasierte Verfahren.

Obwohl seit langem Annotationssoftware wie CETA (DE RIDDER/KLEINNIJENHUIS 2001) für relationale Inhaltsanalysen zur Verfügung steht, gibt es bislang keine vorgefertigten Tools für die vollautomatische Codierung der Texte, da diese oft sehr spezifisch von der Fragestellung abhängt. Jenseits ihrer angestammten Aufgabenfelder sind daher sowohl das freie TABARI als auch das für akademische Zwecke meist kostenlose VRA-Paket wenig sinnvoll nutzbar. Van Atteveldt (2008) hat jedoch bereits ein Open-Source-Paket für generische regelbasierte Textanalyse angekündigt.

Die Reliabiltität und Validität regelbasierter Verfahren hängt sehr stark von der Strukturiertheit des Textmaterials und der Qualität der Regeln ab. Bei sprachlich komplexen und reichhaltigen Texten wird es schwer sein, überhaupt zuverlässig Aussagen, Akteure und Handlungen zu erkennen. Wenn dies erfolgreich geschieht, sind auch valide Messungen zu erwarten, die allerdings wie bei Diktionären durch Homonyme und Rechtschreibfehler beeinträchtigt werden können.

4. Induktive Verfahren

4.1 *Überwachte Textklassifikation*

In der traditionellen manuellen Inhaltsanalyse werden Codieranweisungen in vielen Fällen nicht a priori mit einem umfassenden syntaktisch-semantischen Regelsatz definiert, sondern zumeist durch Beispiele dargestellt. Inhalte, die diesen Beispielen mehr oder minder ähneln, werden dann von den Codierern in die entsprechenden Kategorien eingeordnet. Die Ursache für dieses induktive Vorgehen liegt in der Schwierigkeit, komplexe sprachliche Inhalte gleichermaßen formal exakt und ausreichend allgemein in Codierregeln zu überführen. Oft ist man sich also bei der Codierung einig, ohne genau und übereinstimmend sagen zu können, warum ein Inhalt so und nicht anders kategorisiert wurde. Hohe Reliabilität und Validität, verstanden als Übereinstimmung mit den Vorgaben des Forschungsleiters (früh 2006), sind auch bei unscharfen, nicht-deterministischen Codieranweisungen möglich.

Diese Herangehensweise liegt auch sogenannten überwachten Textklassifikationsverfahren zugrunde, bei denen ein Algorithmus mit einigen Texten und deren korrekter Codierung trainiert wird und daraus mit statistischen Verfahren ein »probabilistisches Diktionär« (pennings/keman 2002) entwickelt, das dann für alle folgenden automatischen Klassifikationen genutzt wird.

Ein sozialwissenschaftlich geläufiges Verfahren für einen solchen Klassifikator wäre z. B. eine logistische Regression, bei der alle Wörter als Prädiktoren *x* verwendet werden, um eine binäre Klassenzugehörigkeit *y* (z. B. ob es sich um eine Sportmeldung handelt) zu erklären. Schätzt man ein solches Modell mit einem Trainings-Datensatz, bei dem alle *x* und *y* bekannt sind, erhält jedes Wort einen Regressions-Koeffizienten. Mit diesen lässt sich anschließend die Klassenzugehörigkeit eines Dokuments, von dem alle *x* bekannt sind, leicht schätzen. Es liegt die Vermutung nahe, dass in diesem Beispiel Wörter wie ›Olympia‹ oder ›Doping‹ einen hohen positiven Koeffizienten und ›Bundesrat‹ oder ›Aktienkurs‹ eher negative Koeffizienten aufweisen sollten. Schließlich werden viele Wörter wie ›Reaktion‹ oder ›grün‹ wenig zwischen den Klassen diskriminieren und daher Koeffizienten nahe Null haben. Der Vorteil induktiver automatischer Klassifikation liegt darin, dass der Forscher kein umfassendes Diktionär für komplexe Kategorien entwerfen muss, sondern dieses vom

Algorithmus aus den vorgegebenen Beispielcodierungen extrahiert wird. Es ist daher möglich, dass dem Forscher wenig bewusste Wörter oder Wendungen am besten zwischen den Kategorien trennen können.

Überwachte Klassifikation ist eines der derzeit meisterforschten Gebiete des Maschinenlernens. Es stehen sehr viele Klassifikationsalgorithmen zu Verfügung, von denen Naive Bayes und Support-Vektor-Maschinen am häufigsten und erfolgreichsten eingesetzt werden (SEBASTIANI 2002; JOACHIMS 1998). Obwohl die Verfahren in vielen verschiedenen Kontexten Verwendung finden, etwa bei der schon erwähnten Klassifikation von Reuters-Tickermeldungen oder der Verarbeitung von Film-Bewertungen, werden sie in den Sozialwissenschaften noch vergleichsweise wenig angewandt bzw. mit neuen Begrifflichkeiten neu erfunden. Ein erstes Beispiel stellt die Klassifikation von Wahlprogrammen durch Laver, Benoit und Garry (2003) dar, die zuerst mit einem ›Kalibrations-Set‹ die Feature-Gewichte errechnen und dann auf ›jungfräuliche‹, also uncodierte, Texte anwenden. Obwohl bei diesem Verfahren zumeist von sehr wenigen bzw. einem einzigen Beispiel auf weitere Texte geschlossen wird, ist die Logik doch dieselbe: Wenn in einem eher rechtskonservativen Wahlprogramm bestimmte Begriffe häufig vorkommen, ist ein Programm mit ähnlichen Wörtern mit hoher Wahrscheinlichkeit auch eher rechtskonservativ. Auch wenn die Ergebnisse von der simplen WORDSCORES-Methode Probleme bei der Interpretation und Replikation aufweisen (vgl. BUDGE/PENNINGS 2007; MARTIN/VANBERG 2008), ist das Verfahren doch ein wichtiger Schritt bei der Verwendung induktiver Textklassifikation in der Politikwissenschaft.

Hopkins und King (2010) verwenden in ihrem quelloffenen README-Paket für R ebenfalls statistische Maschinenlerntechniken, um große Mengen an Blog-Einträgen zuverlässig zu klassifizieren. Das Verfahren hat allerdings den Nachteil (bzw. nach den Autoren den Vorteil), keine individuellen Dokumente zu identifizieren, sondern nur die relativen Häufigkeiten der einzelnen Klassen inferenzstatistisch exakt schätzen zu können. Daher eignet sich das Tool nicht für Forschungsfragen, in denen Zusammenhänge auf Individualebene, etwa zwischen formalen und inhaltlichen Merkmalen der Texte, analysiert werden sollen.

Weitere Anwendungen im sozialwissenschaftlichen Kontext finden sich bei der Klassifikation von Gesetzestexten (HILLARD/PURPURA/WILKERSON 2007) und politischen Blog-Postings (DURANT/SMITH 2006). Angesichts der Anschlussfähigkeit an bestehende manuelle Codierprak-

tiken und der offenbar recht guten Validität der Ergebnisse eignen sich die hier vorgestellten induktiven Verfahren für eine Vielzahl an Fragestellungen, auch wenn noch Evaluationsstudien notwendig sind, um die Qualität der Klassifikationen beurteilen zu können. Da bei induktiven Verfahren menschliche Codierer beteiligt sind, um Trainingsmaterial zu erstellen, hängt die Reliabilität der Klassifikation von der Reliabilität der manuellen Codierung und der Auswahl des Trainingsmaterials ab. Dies lässt sich allerdings problemlos überprüfen, indem etwa aus dem vorcodierten Material eine Stichprobe zum Training, der Rest zum Test der Klassifikation eingesetzt wird (vgl. MANNING/SCHÜTZE 1999). Viele Autoren sprechen dann von einer Validitätsprüfung, wenn angenommen werden kann, die manuelle Codierung sei ein fehlerfreier ›Goldstandard‹. Wenn man sich dieser Argumentation nicht anschließt, kann das Vorgehen auch als einfacher Intercoder-Reliabilitätstest zwischen Mensch und Maschine angesehen werden, wobei auch mehrere Algorithmen gegeneinander antreten können. Ein Test auf konvergente, diskriminante und prädiktive Validität kann aber nur anhand anderer Messungen vorgenommen werden.

4.2 *Informationsextraktion*

Da induktiv-probabilistische Klassifikationsverfahren ebenso wie ihre deduktiv-deterministischen Gegenstücke auf relativ einfachen Wort- bzw. N-Gramm-Strukturen aufbauen, können damit lediglich komplette Dokumente oder Ausschnitte klassifiziert werden. Die Königsdisziplin für die automatische Codierung von Textinhalten wäre aber die Informationextraktion aus natürlichsprachlichen Texten, d. h. das ›Verstehen‹ eines Satzes durch den Computer. Wie in den vorangegangen Abschnitten gezeigt wurde, ist allein die Spezifikation linguistisch tiefgehender Analyseregeln mit erheblichem Aufwand und ungewissem Ausgang verbunden. Selbst Expertensysteme zu themenspezifischen, vorstrukturieren Texten sind selten in der Lage, Inhalte zuverlässig aus komplexen Formulierungen zu destillieren. Doch selbst wenn die computerlinguistischen Tools für die Vorbehandlung der Texte perfekt funktionieren würden, bliebe noch immer die Schwierigkeit, dem Computer mitzuteilen, *was* genau aus einer Aussage zu extrahieren ist. Bei den Verfahren, die in Abschnitt 3.2 vorgestellt wurden, war das Vorgehen deduktiv, d. h.

der Forscher muss einen vollständigen Regelsatz definieren, wie z. B. der Urheber einer Aussage zu erkennen ist (vgl. VAN ATTEVELDT 2008). Eleganter und näher an der manuellen Codierpraxis wäre jedoch auch bei syntaktisch-semantischen Verfahren ein induktives Vorgehen über die Verknüpfung von Rohtext und beispielhaft extrahierten Objekten. Während es bei menschlichen Codierern sehr leicht ist, ihnen die betreffenden Satzbestandteile zu markieren und wichtige Objekte zu identifizieren, ist es momentan noch sehr schwer, dem Computer per Beispiel beizubringen, wo in einem Satz die relevanten Informationen liegen, die dann im nächsten Satz scheinbar völlig anders angeordnet sind. Zwar ist eine Benutzerschnittstelle mit digitalem Markieren der Objekte bereits in bestehender Software für qualitative Inhaltsanalyse, z. B. MAXQDA oder Atlas.Ti, vorhanden, allerdings werden die manuellen Annotationen nicht automatisch in Regeln überführt, da die Entwicklung passender Algorithmen noch auf sich warten lässt.

Erste Fortschritte gibt es wie bei den anderen regelbasierten Verfahren vor allem dort, wo Texte stark strukturiert und mit einem kleinen Wortschatz ausgestattet sind. Siefkes (2008) Trainable Information Extractor (TIE) konnte relativ gute Ergebnisse bei der Extraktion von Terminen und Sprechern aus Veranstaltungsankündigungen erzielen. Van Atteveldt (2008) erreicht eine mittlere Genauigkeit bei der Identifikation von Sprechern im niederländischen politischen Diskurs, muss aber noch immer Diktionäre für die Objektidentifikation zur Hilfe nehmen. Noch scheinen syntaktisch-semantische Verfahren für ein induktives Vorgehen nicht weit genug entwickelt zu sein, obwohl die Lage weniger düster ist als van Cuilenburg et al. (1988) einst prognostizierten. Es lohnt sich aber, die aktuellen Entwicklungen in der Informatik auf diesem Gebiet im Auge zu behalten und frühzeitig für die sozialwissenschaftliche Anwendung zu evaluieren.

5. Diskussion und Ausblick

Automatisierte Verfahren der Inhaltsanalyse sind heute in zweierlei Hinsicht attraktiver denn je: Erstens sind riesige Mengen von Text im Internet gratis oder mit geringem Aufwand verfügbar. Der schiere Umfang an Nachrichten, Blog-Einträgen, Kommentaren und anderen nutzergenerierten Inhalten lässt abseits von speziellen Fragestellungen und winzigen Stichproben kaum noch manuelle Analysen zu. Hier amortisiert sich

der Aufwand selbst für umfangreiche Diktionäre oder Annotationen von Trainingsdatensätzen sehr schnell. Die eigentliche Codierung kann fast in Echtzeit erfolgen, was gerade für die Medienresonanzanalyse oder andere Formen von Themenmonitoring (GLANCE/HURST/TOMOKIYO 2004) im höchsten Maße attraktiv ist.

Zweitens ist die Methodenentwicklung für die automatisierte quantitative Analyse medialer Inhalte in vollem Gange, auch außerhalb der Sozialwissenschaften. Fortschritte werden dabei nicht nur auf dem Gebiet des statistischen Maschinenlernens, z.B. von Support-Vektor-Maschinen (JOACHIMS 1998), gemacht, sondern auch auf der Ebene der Feature-Extraktion und eigentlichen Messung von multimedialen Inhalten. Obwohl bisher größtenteils auf Textmaterial zurückgegriffen wird, sind die hier vorgestellten Verfahren genauso auf Audio- und Videomaterial anwendbar, wenn die entscheidenden Merkmale automatisch extrahierbar sind. Dabei sind nicht nur Transkriptionen von wörtlicher Rede durch Spracherkennung denkbar (SAHUGUET/BEZMAN 2008), sondern auch die Erhebung und Analyse genuin audiovisueller Stimuluseigenschaften wie Farben, Schnitte, Geräusche und Bewegungsmuster.

Besonders viel versprechend für die Verknüpfung manueller und automatischer Codierpraktiken scheinen die vorgestellten induktiven Verfahren. Diese sind jedoch bislang nicht in einfach bedienbare Software-Pakete integriert worden, zudem fehlt es schlicht an praktischen Erfahrungen mit verschiedenen Extraktionsmethoden und Algorithmen im sozialwissenschaftlichen Kontext. Für diktionärsbasierte Forschung sind die Bedingungen dagegen so gut wie nie zuvor: Sowohl umfangreiche Textkorpora als auch einfach bedienbare Programme stehen gratis im Internet zur Verfügung, während Suchmaschinen immer gründlicher und benutzerfreundlicher werden. Dadurch eignen sich automatische Textanalysen nicht nur für umfangreiche Forschungsprojekte, sondern auch für den Einsatz in Unternehmen, etwa in der kommerziellen Medienresonanzanalyse (RAUPP/VOGELGESANG 2009).

Zu einer vollautomatischen Analyse beliebiger medialer Inhalte, die genauso valide ist wie eine gute manuelle Codierung, wird es wohl trotz der hier vorgestellten Entwicklungen in absehbarer Zukunft nicht kommen. Es muss also weiterhin ein Kompromiss zwischen Validität bzw. Reliabilität der Messung und dem realisierbaren Stichproben- und Indikatorenumfang gefunden werden. Mit modernen Verfahren automatischer Textanalyse lässt sich diese Diskrepanz aber erheblich verringern.

Literatur

ABBASI, A.; CHEN, H.: Applying authorship analysis to extremist-group Web forum messages. In: *Intelligent Systems*, 20 (5), 2005, S. 67-75

ADAM, S.: Medieninhalte aus der Netzwerkperspektive. In: *Publizistik*, 53 (2), 2008, S. 180-199

ALEXA, M.; ZUELL, C.: Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review. In: *Quality and Quantity*, 34 (3), 2000, S. 299-321

BERENDT, B.; SCHLEGEL, M.; KOCH, R.: Die deutschsprachige Blogosphäre: Reifegrad, Politisierung, Themen und Bezug zu Nachrichtenmedien. In: ZERFASS, A.; WELKER, M.; SCHMIDT, J. (Hrsg.): *Kommunikation, Partizipation und Wirkungen im Social Web*. Köln 2008, S. 72-96

BEST, K.-H.: Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In: WICHTER, S; BUSCH, A. (Hrsg.): *Wissenstransfer – Erfolgskontrolle und Rückmeldungen aus der Praxis*. Frankfurt/M. u. a. 2006, S. 21-31

BRASCHLER, M.; RIPPLINGER, B.: How Effective is Stemming and Decompounding for German Text Retrieval? In: *Information Retrieval*, 7, 2004, S. 291-316

BRODER, A.; GLASSMAN, S.; MANASSE, M.; ZWEIG, G.: Syntactic clustering of the Web. In: *Computer Networks and ISDN Systems*, 29 (8-13), 1997, S. 1157-1166

BROOKS, C.; MONTANEZ, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: *Proceedings of the 15th international conference on World Wide Web*, 2006, S. 625-632

BUDGE, I.; PENNINGS, P.: Do they work? Validating computerised word frequency estimates against policy series. In: *Electoral Studies*, 26 (1), 2007, S. 121-129

CARLEY, K.: Network text analysis: The network position of concepts. In: ROBERTS, C. (Hrsg.): *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Mahwah 1997, S. 79-100

DE RIDDER, J.; KLEINNIJENHUIS, J.: Media Monitoring Using CETA: The Stock-Exchange Launches of KPN and WOL. In: WEST, M. (Hrsg.): *Applications of Computer Content Analysis*, Westport 2001, S. 165-184

DI GIACOMO, E.; DIDIMO, W.; GRILLI, L.; LIOTTA, G.: Graph Visualization Techniques for Web Clustering Engines. In: *IEEE Transactions on Visualization and Computer Graphics*, 2007, S. 294-304

DUBAY, W.: *The Principles of Readability*. Costa Mesa 2004

DUNPHY, D.; BULLARD, C.; ELINOR, E.: *Validation of the General Inquirer Harvard IV Dictionary*. Harvard 1974

DURANT, K.; SMITH, M.: Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection. In: *Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web,* 2006

FAN, D.: Computer Content Analysis of Press Coverage and Prediction of Public Opinion for the 1995 Sovereignty Referendum in Quebec. In: *Social Science Computer Review*, *15*(4), 1997, 351-366

FEINERER, I.; HORNIK, K.; MEYER, D.: Text Mining Infrastructure in R. In: *Journal of Statistical Software*, 25 (1), 2008, S. 1-54

FRIEDL, J.: *Mastering Regular Expressions*. Sebastopol 2006

FRÜH, W.: *Inhaltsanalyse: Theorie und Praxis*, 6. Auflage. Konstanz 2006

GLANCE, N.; HURST, M.; TOMOKIYO, T.: BlogPulse. Automated Trend Discovery for Weblogs. In: *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004

GRIEVE, J.: Quantitative Authorship Attribution: An Evaluation of Techniques. In: *Literary and Linguistic Computing*, 22(3), 2007, S. 251-270

HAGEN, L.: Freitextrecherche in Mediendatenbanken als Verfahren zur computerunterstützten Inhaltsanalyse. Beschreibung, theoretische und praktische Überlegungen zur Validität und ein Anwendungsbeispiel. In: WIRTH, W./LAUF, E. (Hrsg.): *Inhaltsanalyse. Perspektiven, Probleme, Potentiale.* Köln 2001, S. 337-352

HILLARD, D.; PURPURA, S.; WILKERSON, J.: Computer Assisted Topic Classification for Mixed Methods Social Science Research. *In: Journal of Information Technology & Politics*, 4 (4) 2007

HOLLANDERS, D.; VLIEGENTHART, R.: Telling what yesterday's news might be tomorrow: Modeling media dynamics. In: *Communications*, 33 (1), 2008, S. 47-68

HOPKINS, D.; KING, G.: A Method of Automated Nonparametric Content Analysis for Social Science. In*: American Journal of Political Science*, 54 (1), 2010 http://gking. harvard. edu/files/words.pdf

HOTHO, A.; NÜRNBERGER, A.; PAASS, G.: A Brief Survey of Text Mining. In: *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology*, 20 (1), 2005, S. 19-62

JOACHIMS, T.: Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. In: *Computational Linguistics*, 29 (4), 1998, S. 655-664

KERCHER, J.: Die Verständlichkeit deutscher Spitzenpolitiker – Eine Untersuchung zur Messung und Erklärung einer bislang unerforschten Thematik. Vortrag auf Jahrestagung des DVPW-Arbeitskreises »Wahlen und politische Einstellungen« an der Universität Duisburg-Essen, 2008

KING, G.; LOWE, W.: An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. In: *International Organization*, 57 (3), 2003, S. 617-642

LANDMANN, J.;ZÜLL, C.: Computerunterstützte Inhaltsanalyse ohne Diktionär? In: *ZUMA Nachrichten* 54, 2004, S. 117-140

LASSWELL, H.; NAMENWIRTH, J.: *The Lasswell Value Dictionary*. New Haven 1969

LAVER, M.; BENOIT, K.; GARRY, J.: Extracting Policy Positions from Political Texts Using Words as Data. In: *American Political Science Review*, 97, 2003, S. 311-331

LEBERT, M.: *Project Gutenberg, from 1971 to 2005*, 2005, http://www.etudes-francaises.net/dossiers/gutenberg_eng.htm [20.08.2008]

LOWE, W.: *Software for content analysis: A review*, 2002, http://people.iq.harvard.edu/~wlowe/Publications/rev.pdf [20.08.2008]

LOWE, W.: *Yoshikoder: An open source multilingual content analysis tool for social scientists*. Paper presented at the 2006 APSA Conference, 2006

MANNING, C.; SCHÜTZE, H.: *Foundations of Statistical Natural Language Processing*. Cambridge 1999

MARTIN, L.; VANBERG, G.: A Robust Transformation Procedure for Interpreting Political Text. In: *Political Analysis*, 16 (1), 2008, S. 93-100

MERTEN, K.: *Inhaltsanalyse. Einführung in Theorie und Methode*. Wiesbaden 1995

PENNINGS, P.; KEMAN, H.: Towards a New Methodology of Estimating Party Policy Positions. In: *Quality and Quantity*, 36 (1), 2002, S. 55-79

RAUPP, J.; VOGELGESANG, J.: *Medienresonanzanalysen. Ein Lehr- und Arbeitsbuch*. Wiesbaden 2009

ROBERTS, C.: Semantic Text Analysis: On the Structure of Linguistic Ambiguity in Ordinary Discourse. In: ROBERTS, C. (Hrsg.): *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Mahwah 1997, S. 55-78

RUCHT, D.; YANG, M.; ZIMMERMANN, A.: *Politische Diskurse im Internet und in Zeitungen: Das Beispiel Genfood*. Wiesbaden 2008

SAHUGUET, A.; BEZMAN, A.: *In their own words: political videos meet Google speech-to-text technology*, 2008, http://googleblog.blogspotcom/2008/07/in-their-own-words-political-videos.html [20.08.2008]

SALTON, G.: *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston 1989

SCHMITT, L.; CHRISTIANSON, K.; GUPTA, R.: Linguistic Computing with UNIX Tools. In: KAO, A.; POTEET, S. (Hrsg.): *Natural Language Processing and Text Mining*, London 2007, S. 221-258

SCHÖNBACH, K.: Nachrichtenwerte und computerunterstützte Inhaltsanalyse. In: ZUMA Nachrichten, (2), 1978. S. 3-11

SCHRODT, P.; DAVIS, S.; WEDDLE, J.: Political Science: KEDS – A Program for the Machine Coding of Event Data. In: *Social Science Computer Review*, 12(4), 1994, S. 561-588

SCHULZ, W.: *Die Konstruktion von Realität in den Massenmedien*. Freiburg/München 1976

SEBASTIANI, F.: Machine learning in automated text categorization. In: *ACM Computing Surveys*, 34, 2002, S. 1-47

SHAPIRO, G.: The Future of Coders: Human Judgments in a World of Sophisticated Software. In: ROBERTS, C. (Hrsg.): *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Mahwah 1997, 225-238

SIEFKES, C.: *An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models*. Saarbrücken 2008

SIMON, A.; XENOS, M.: Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis. In: *Political Analysis*, 12 (1), 2004, S. 63-75

SPIEGEL VERLAG: *SPIEGELnet und Wissen Media starten SPIEGEL Wissen*, 2007, http://www.spiegelgruppe.de/spiegelgruppe/home.nsf/pmwebaktuell/1619A963C27E7741C12573B4003468B9 [20.08.2008]

STEGBAUER, C.; RAUSCH, A.: *Strukturalistische Internetforschung*, Wiesbaden 2006

STONE, P.: Thematic text analysis: New agendas for analyzing text content. In: ROBERTS, C. (Hrsg.): *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Mahwah 1997, S. 35-54

STREHL, A.; GHOSH, J.; MOONEY, R.: Impact of similarity measures on webpage clustering. In: *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000, S. 58-64

TAYLOR, D.: Readability.info, 2008, http://www.readability.info/info.shtml [20.08.2008]

VAN ATTEVELDT, W.: *Parsing, Semantic Networks, and Political Authority*. Paper presented at the 2008 ICA conference, 2008

VAN CUILENBURG, J.; KLEINNIJENHUIS, J.; DE RIDDER, J.: Artificial intelligence and content analysis. In: *Quality and Quantity*, 22 (1), 1988, S. 65-97

WELKER, M.; WERNER, A.; SCHOLZ, J.: Online-Research. Heidelberg 2005

WILKE, J.; REINEMANN, C.: Do the Candidates Matter? Long-Term Trends of Campaign Coverage – A Study of the German Press Since 1949. In: *European Journal of Communication*, 16 (3), 2001, S. 291-314

YERAZUNIS, W.: Sparse Binary Polynomial Hashing and the CRM114 Discriminator. In: *Proceedings of the 2003 Cambridge Spam Conference*, 2003

ZÜLL, C.; LANDMANN, J.: *Computerunterstützte Inhaltsanalyse: Literaturbericht zu neueren Anwendungen*, ZUMA, ZUMA Methodenbericht 2002/02, 2002