Automatische Inhaltsanalyse

Michael Scharkow

Abstract Nachdem es viele Jahre vergleichsweise wenig Entwicklung auf dem Gebiet automatischer Inhaltsanalysen gegeben hat, erleben diese Verfahren in jüngster Zeit geradezu einen Boom in Wissenschaft und Marktforschung. Angesichts des Umfangs an leicht verfügbaren digitalen Dokumenten, sowohl aus (halb-)öffentlicher interpersonaler Kommunikation als auch aus Medienangeboten, scheinen automatische Analyseverfahren nicht nur sehr attraktiv, sondern geradezu alternativlos. In diesem Beitrag werden traditionelle und neuere Ansätze automatischer Inhaltsanalyse vorgestellt sowie deren Vor- und Nachteile gegenüber der manuellen Codierung diskutiert. Dabei zeigt sich, dass man sich als Forschende keineswegs für einen Weg entscheiden muss, sondern sich in vielen Fällen manuelle und automatische Verfahren kombinieren lassen. Allerdings gibt es für viele neuere Ansätze weiterhin kaum fertige Lösungen für die angewandte Forschung.

Schlagwörter computerunterstützte Inhaltsanalyse (CUI), Computerlinguistik, maschinelles Lernen, Co-Occurrence-Analyse

1 Einführung

Seitdem Computer in der Lage sind, digitale Texte zu speichern und zu verarbeiten, haben sich Sozialwissenschaftlerinnen und Sozialwissenschaftler um die Nutzung dieser Technologie für die quantitative Inhaltsanalyse bemüht. Schon in den 1950er Jahren hatten die Pioniere der Inhaltsanalyse festgestellt, dass die Aufbereitung und Codierung großer Textmengen, wie sie die Massenmedien rund um die Uhr produzieren, enorme Ressourcen benötigten, die nicht selten in einem ungünstigen Verhältnis zum (erwarteten oder erhaltenen) Erkenntnisgewinn standen (Pool 1959). Die Nutzung von Computern sollte dafür sorgen, dass Inhaltsanalysen effektiver und letztlich auch effizienter

W. Möhring, D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft,*
DOI 10.1007/978-3-531-18776-1_16,

durchzuführen waren. Von Anfang an standen jedoch nicht forschungsökonomische Überlegungen im Vordergrund der Anwendung computergestützter Analyseverfahren, sondern deren Reliabilität. Computer können zuverlässiger Wörter zählen, Texte sortieren, Stichproben generieren, umfangreiche Dokumente archivieren und schließlich auch Ergebnisse berechnen als Menschen. Dies machte sie für viele Teile des inhaltsanalytischen Forschungsprozesses interessant, selbst zu einer Zeit, als eine Stunde Computernutzung noch teurer als das Monatsgehalt einer Hilfskraft war (Stone 1997). Heute ist fast jede quantitative und qualitative Inhaltsanalyse computerunterstützt: Dokumente werden per Computer gespeichert und verteilt, die Dateneingabe geschieht zumeist direkt am Computer, Reliabilitätststest und auch die Endergebnisse werden mit entsprechender Software ausgewertet. Man kann sagen, dass die rein manuelle Inhaltsanalyse praktisch nicht mehr vorkommt.

Das zentrale Anliegen seit den Ursprüngen der computergestützten Inhaltsanalyse war jedoch von jeher, auch die eigentliche Codierung zu automatisieren, so dass man letztlich auf die Unterstützung von teuren, immer wieder neu zu schulenden, Fehler machenden und insgesamt schwer kontrollierbaren menschlichen Codiererinnen und Codierern verzichten kann. Obwohl mittlerweile viele Forschergenerationen an der Lösung dieses Problems gearbeitet haben, ist die automatische Inhaltsanalyse noch immer eine Randerscheinung in der Kommunikationswissenschaft. Die Argumente, die gegen den Einsatz von Software zur Codierung hervorgebracht wurden und werden, sind dabei fast unverändert: Die Verfahren sind entweder zu einfach, um sinnvolle Inferenzen zu gestatten, zu aufwändig für den Forschungsalltag oder zu wenig valide im Vergleich zu dem, was die „gelenkte Rezeption" (Wirth 2001: 157) der Codierenden leisten kann (vgl. Früh 2007, Rössler 2010). Das Misstrauen, das automatischen Inhaltsanalysen noch heute häufig entgegengebracht wird, liegt aber auch darin begründet, dass es in den letzten Jahrzehnten eher ein Auseinanderdriften von manuellen und automatischen Analyseverfahren gegeben hat: In dem Maße, wie sich beide Forschungsrichtungen ausdifferenziert und weiterentwickelt haben, wurde es schwieriger, Gemeinsamkeiten und Anknüpfungspunkte zu erkennen und praktisch nutzbar zu machen.

Grundsätzlich unterscheiden sich manuelle und automatische Inhaltsanalyse in nur wenigen, aber zentralen Punkten des Forschungsprozesses (vgl. Scharkow 2012: 49 ff.): Datenaufbereitung, Operationalisierung, Codierung und Qualitätssicherung. Wie bei manuellen Analysen hängt die Wahl des konkreten Verfahrens eng mit der Fragestellung zusammen: Explorative Ansätze eignen sich vor allem bei der Beschreibung großer Textmengen und dem Entdecken von semantischen oder thematischen Zusammenhängen, hypothesenprüfende Verfahren dagegen zur reproduzierbaren Zuordnung von Texten zu vordefinierten Kategorien. Dies hat auch Konsequenzen für die Qualitätskontrolle: Während auf rein technischer Ebene jede automatische Inhaltsanalyse vollständig zuverlässig und reproduzierbar ist, muss auf inhaltlicher Ebene stets geprüft werden, ob die Codierung zuverlässig und gültig im Sinne des eigenen Forschungsinteresses ist. Diese Qualitätssicherung ist ebenso wie klassische Intercoder-Reliabilitätstests mit nicht

unerheblichem Aufwand verbunden, stellt aber den einzigen Weg dar, die Validität des Verfahrens und damit der Ergebnisse intersubjektiv überprüfbar zu machen. Ein zentraler Vorteil automatischer Verfahren gegenüber dem Arbeiten mit menschlichen Codierenden liegt in der Tatsache, dass die Dokumentation der Operationalisierung, Reliabilitätsprüfungen und der eigentlichen Codierung deutlich transparenter ist. Da die Codierregeln ohnehin in maschinell lesbarer Form vorliegen müssen, sind sie vollständig dokumentiert und für andere replizierbar.

2 Begriffe und Grundlagen

In der deutschsprachigen Literatur werden die hier dargestellten Verfahren unter verschiedenen Begriffen zusammengefasst. Am häufigsten findet sich – nicht zuletzt wegen entsprechend benannter Kapitel in den Lehrbüchern von Früh (2007) oder Rössler (2010) – der Begriff Computerunterstützte Inhaltsanalyse (CUI). Da praktisch kaum eine Inhaltsanalyse und insgesamt kaum ein empirisches Forschungsprojekt heute noch gänzlich ohne Computerunterstützung durchgeführt wird, ist dieser Begriff jedoch wenig aussagekräftig. In diesem Beitrag wird stattdessen der Begriff ‚Automatische Inhaltsanalyse' verwendet, welcher Verfahren bezeichnet, bei denen die eigentliche Codierentscheidung von einem Computeralgorithmus getroffen wird und damit nur mittelbar dem Einfluss der Forschenden unterliegt (vgl. Monroe & Schrodt 2008).

Grundsätzlich lassen sich automatische Verfahren der Inhaltsanalyse nach verschiedenen Kriterien klassifizieren (vgl. Roberts 2000; West 2001; Züll & Alexa 2001), die sich entweder auf das gewählte Textmodell oder die Operationalisierungsstrategie beziehen. Scharkow (2010) folgt in einem ersten Schritt der aus der Informatik übernommenen Einteilung der Verfahren in unüberwachte, d. h. vollautomatische, und überwachte Verfahren (vgl. auch Hillard, Purpura & Wilkerson 2007). Unüberwachte Verfahren laufen vollständig autonom im Computer ab und benötigen keinerlei text- oder kategorienspezifischen Input der Forschenden. Diese können lediglich generell die Parameter des Analysemodells festlegen und versuchen, die Ergebnisse der vollautomatischen Analyse hinsichtlich der Forschungsfrage zu interpretieren. Traditionell spielen in der sozialwissenschaftlichen Inhaltsanalyse fast nur überwachte Verfahren eine Rolle, da nur mit diesen gezielt Kategorien entwickelt und Hypothesen geprüft werden können. Da die Codierregeln bei überwachten Verfahren stets manuell definiert bzw. angepasst werden müssen, sind diese deutlich aufwändiger in der Umsetzung, versprechen aber auch eine höhere Validität und leichtere Interpretation der Ergebnisse.

Eine zweite grundlegende Unterscheidung betrifft die Frage, wie eine Mitteilung in ein maschinenlesbares Format transformiert wird und welches Textmodell man an die Analyse heranträgt. Hier lassen sich rein statistische Verfahren von (computer-)linguistischen Ansätzen unterscheiden. Der statistische Ansatz basiert auf der Annahme, dass die für die Forschungsfrage relevanten Merkmale einer Mitteilung (eines Textes, Bildes

oder einer Tonfolge) sich als ungeordnete Menge verschiedener Wörter, Bildpunkte oder Töne verstehen lassen, deren Binnenstruktur jedoch keine Rolle spielt. Bei Textanalysen spricht man deshalb von *Bag-of-Words*-Ansätzen, in denen im Prinzip jedes Wort als isolierte Variable aufgefasst wird und die Sätze „Die Regierung kritisiert die Opposition" und „Die Opposition kritisiert die Regierung" äquivalent sind. Im Gegensatz dazu berücksichtigen computerlinguistische Verfahren explizit die syntaktische und semantische Struktur von Aussagen. Hierfür werden die Texte in eine Graphen- oder Baumstruktur überführt, die dann nach verschiedenen Transformationsregeln ausgewertet werden kann (vgl. Carstensen et al. 2009).

Die Entscheidung für Bag-of-Words oder linguistische Verfahren hat weitreichende Konsequenzen, auch wenn in vielen Anwendungskontexten beide Ansätze verknüpft werden (vgl. van Atteveldt 2008). Statistische Ansätze benötigen naturgemäß eine größere Anzahl an Untersuchungseinheiten, um effektiv arbeiten zu können, während linguistische Verfahren auch bei kleinen Stichproben gut funktionieren. Zugleich erfordert die computerlinguistische Inhaltsanalyse deutlich mehr Vorarbeiten und vor allem projektspezifische Anpassungen, weil die Codierregeln je nach Sprache, Textsorte und Forschungsinteresse erheblich variieren können. Rein statistische Verfahren sind größtenteils sprach- und themenunabhängig, d. h. zumindest die grundlegenden Algorithmen können unverändert für verschiedene Fragestellungen eingesetzt werden. Im Gegensatz zu syntaktisch-semantischen Analysen sind statistische Verfahren auch nicht an die Textform gebunden. Da die Algorithmen intern ohnehin mit arbiträren Zahlenwerten arbeiten, können die gleichen statistischen Verfahren für die Analyse von Texten, Bildern, Videos oder Tönen eingesetzt werden, solange deren Merkmale in eine entsprechende Datenmatrix überführbar sind. Aus Sicht der Software ist es dann unerheblich, ob nun Musikstücke, Porträtfotos oder Zeitungsbeiträge automatisch codiert werden (Scharkow 2012: 210 f.).

Im Zusammenhang mit der Unterscheidung zwischen rein statistischen und syntaktisch-semantischen Ansätzen ist auch die Frage nach den für die Analyse relevanten Merkmalen *(Features)* der Untersuchungseinheiten zu beantworten. Die sog. Feature-Extraktion wird bei automatischen Textanalysen recht selten explizit thematisiert, weil traditionell die meisten automatischen Verfahren Einzelwörter als Features verwenden. Dies ist aber nicht zwingend notwendiges Merkmal automatischer Inhaltsanalysen: Neben einzelnen Wörtern *(Unigrammen)* lassen sich auch kleinteiligere (z. B. einzelne Zeichen) oder komplexere Features (z. B. Wortgruppen der Länge *n,* sog. *n-Gramme*) für die Analyse verwenden. Die Analyse von n-Grammen hat gegenüber traditionellen Unigramm-Analysen den Vorteil, dass sich die Semantik von Negationen oder bestimmten Idiomen zu einem gewissen Grad berücksichtigen lässt, ohne gleich eine volle syntaktisch-semantische Analyse *(Parsing)* durchführen zu müssen.

Jenseits der Analyse von digitalen Texten ist gerade die Definition relevanter Features und deren Extraktion aus dem Codiermaterial von entscheidender Bedeutung: Hier ist zunächst einmal eine umfassende theoretische Auseinandersetzung nötig, wo-

durch und wie Bedeutung in Bildern, Tönen oder Videos transportiert wird. Bei komplexen medialen Inhalten ist allein diese Aufgabe nicht zu unterschätzen, zumal diese Texttheorie auch soweit formalisiert werden muss, dass daraus konkrete Extraktionsregeln für die Software definiert werden können. Eine für menschliche Codierende vergleichsweise einfache Codieranweisung nach dem Schema „Codieren Sie 1, wenn in dem Bild eine Person abgebildet ist" erfordert ggf. einen enormen Operationalisierungsaufwand. Andererseits verspricht gerade die Auseinandersetzung mit der Frage, was nun die für die Analyse relevanten Features einer Mitteilung sind, einen großen theoretischen und empirischen Mehrwert (vgl. z. B. Cutler & Davis 2002; Xu, Maddage, Xu, Kankanhalli & Tian 2003). Anders formuliert zwingt die Nutzung automatischer Analyseverfahren die Inhaltsanalytikerinnen und -analytiker dazu, selten oder zumeist nur implizit formulierte Annahmen zur Messung explizit zu machen. Dies führt häufig dazu, dass die Instrumentenentwicklung bei automatischen Inhaltsanalysen viel mehr Aufwand verursacht als die eigentliche Codierung. Rössler (2010: 191) weist daher zu Recht darauf hin, dass automatische Analysen nicht zwangsläufig weniger aufwändig sind als manuelle.

3 Datenerhebung und -aufbereitung

Eine grundlegende Voraussetzung für jede automatische Inhaltsanalyse ist maschinenlesbares Codiermaterial. Während die Digitalisierung von Dokumenten vor einigen Jahrzehnten noch ein großes Problem für die Forschung war, ist die Verfügbarkeit digitaler Medieninhalte heute nur noch ein Randproblem. Seitdem de Weese (1977) die Möglichkeiten digitaler Satzsysteme für die Erhebung tagesaktueller Medieninhalte erstmals praktisch demonstrieren konnte, ist zumindest die automatische Codierung von Print-Medien denkbar einfach. Schon seit den 1980er Jahren stehen Volltexte vieler Zeitungen als Jahrgangs-Archive digital zur Verfügung. Zusätzlich ist über kommerzielle Anbieter wie LexisNexis oder Factiva eine sehr große Anzahl an Beiträgen einfach über das Internet abzurufen. Trotzdem sind diese digitalisierten Archive nicht ohne ihre Probleme: Wie Deacon (2007) zeigen konnte, werden nicht alle Beiträge der Druckfassung elektronisch archiviert. Zudem gehen durch die Nutzung einfacher Textformate potentiell wichtige Informationen zum Layout der Beiträge verloren, zumeist sind auch Abbildungen nicht im digitalen Archiv vorhanden. Auch wenn sowohl die Archive der Medienangebote als auch kommerzieller Dienstleister auf standardisierten Datenbanken basieren, sind die exportierbaren Textdaten oft nicht unmittelbar verwendbar, sondern müssen vor der Analyse nochmals bereinigt werden. Ein weiteres Problem bei der Verwendung von LexisNexis und anderen Anbietern liegt in deren Nutzungsbedingungen, die das Herunterladen und dauerhafte Archivieren von Dokumenten untersagen. Dies macht die Durchführung intersubjektiv überprüfbarer Inhaltsanalysen zumindest schwierig. Schließlich liegen auch heute noch nicht von allen Medien digitale Archive

vor – wer Rundfunkbeiträge, Fernsehnachrichten oder die BILD-Zeitung automatisch codieren möchte, muss diese selbst transkribieren und digitalisieren.

Eine Alternative zu digitalisierten Inhalten von Print- und Rundfunkmedien stellen genuine Online-Inhalte dar, die oftmals mit geringem Aufwand und Kosten erhoben und archiviert werden können. An dieser Stelle kann die Erhebung von Online-Inhalten nicht ausführlich diskutiert werden, es sei daher auf Rössler und Wirth (2001) sowie Scharkow (2012) verwiesen. Neben der Stichprobenziehung ist das zentrale Problem von Online-Inhalten, dass diese im Gegensatz zu Print- oder Rundfunkbeiträgen keine kanonische Form haben, sondern je nach Web-Browser oder Nutzereinstellungen zumindest unterschiedlich dargestellt werden bzw. sogar inhaltlich variieren. Wie bei den digitalen Printarchiven gibt es zudem das Problem, dass Layoutinformationen und zusätzliche multimediale Inhalte für die automatische Verarbeitung gar nicht oder nur mit großem Aufwand nutzbar gemacht werden können. Kurz: Die Inhalte, die die einzelne Nutzerin bzw. der einzelne Nutzer sieht, lassen sich kaum in dieser Form archivieren und codieren, während die archivierbaren Beiträge ggf. in einer Form verarbeitet werden, die niemals so rezipiert wurde. Dies betrifft sowohl Inhalte, die direkt von einer Webseite heruntergeladen werden als auch solche, die über spezielle Programmschnittstellen (API) von Plattformen wie Facebook, Twitter oder Youtube abgerufen werden. Letztere Möglichkeit bietet aber zumindest den Vorteil, dass die Daten in einem stark standardisierten Format vorliegen und sich so leichter weiterverarbeiten lassen als konventionelle Webseiten.

Unabhängig davon, ob das Untersuchungsmaterial nun aus digitalisierten und genuinen Online-Inhalten besteht, ist es notwendig, die Texte in einem Format zu speichern, das möglichst leicht zu verarbeiten ist. In der Regel ist dies heute wie vor 50 Jahren einfacher Klartext im ASCII oder UTF-Format. Während bei manuellen Analysen häufig Dateiformate verwendet werden, die Layoutinformationen oder eingebettete Abbildungen enthalten, wie etwa das PDF-Format oder sogar Microsoft Word-Dateien, unterstützen die meisten Analysetools lediglich Klartext und ggf. HTML als Datenformat. Dies bedeutet, dass häufig zunächst eine große Anzahl an Dokumenten konvertiert werden muss, bevor automatisch codiert werden kann. Wenn sich diese Umwandlung nicht zuverlässig automatisieren lässt, entfällt auch der Vorteil computergestützter Analysen, große Datenmengen schnell verarbeiten zu können.

Bei Inhaltsanalysen, in denen Auswahl- und Analyseeinheit nicht identisch sind, muss vor der eigentlichen Codierarbeit zunächst das Untersuchungsmaterial zerlegt werden *(Unitizing)*. Bei manuellen Inhaltsanalysen tun dies zumeist die Codiererinnen und Codierer während der Feldarbeit. Typische Fälle sind die Segmentierung von Rundfunknachrichten in Einzelbeiträge oder die Zerlegung von Einzelbeiträgen in Aussageeinheiten, z. B. Claims oder Frames. Während menschliche Codierende sich zumeist inhaltlicher, d. h. syntaktischer oder semantischer Kriterien bedienen, kann das Unitizing bei automatischen Verfahren häufig nur auf formale Kriterien, d. h. Satzzeichen oder Absatzmarken, zurückgreifen. Praktisch muss man sich dann entscheiden,

die Codiereinheiten entweder manuell zu identifizieren – was gerade bei komplexen Claim-oder Frame-Analysen häufig Reliabilitätsprobleme nach sich zieht – oder sich damit begnügen, auf Absatz- oder Satzebene zu arbeiten. In der bisherigen Forschung zeigt sich, dass eher statistische Analysen häufig auf Beitrags- oder Absatzebene angesiedelt sind, während computerlinguistische Verfahren praktisch immer auf Satzebene arbeiten. Hierbei besteht das Problem, dass der Kontext aus den umliegenden Sätzen bei der Codierung berücksichtigt werden muss, z. B. um Anaphern aufzulösen, d. h. Pronomen durch ihre Referenten zu ersetzen.

Der nächste Schritt einer automatischen Inhaltsanalyse ist die Definition von Regeln für die Feature-Extraktion, also die Aufsplittung in die für die Analyse relevanten Merkmale. Bei der Verwendung fertiger Software-Lösungen für die Textanalyse ist dieser Schritt meist implizit, zumindest wenn Einwort-Features verwendet werden. Sollen für die Analyse andere Merkmale verwendet werden, muss ein entsprechender Extraktionsalgorithmus verwendet werden. Bei multimedialen Inhalten gibt es eine Vielzahl möglicher Features auf visueller und auditiver Ebene (Cutler & Davis 2002), jedoch kaum fertige Lösungen.

Da selbst bei einer kleinen Stichprobe von Untersuchungseinheiten die Zahl extrahierter Features sehr groß werden kann, werden gerade bei automatischen Textanalysen häufig mehrere Bereinigungsschritte vor der eigentlichen Analyse vorgenommen. Dieses *Preprocessing* dient einerseits dazu, die Anzahl an Features zu reduzieren, andererseits diese mit zusätzlichen Kontextinformationen anzureichern. Zu den gebräuchlichsten Feature-reduzierenden Preprocessing-Verfahren zählt die Entfernung von besonders häufig vorkommenden, sog. Stoppwörtern, die Ersetzung von gebeugten Wortformen durch Grundformen *(Lemmatisierung* oder *Stemming)* und die Auflösung von Anaphern oder Synonymen (vgl. Hotho, Nürnberger & Paaß 2005). Für alle linguistischen und einige statistische Auswertungsverfahren werden zudem alle Wortformen hinsichtlich ihrer Satzfunktion annotiert *(Part-of-Speech-Tagging).* Dies ermöglicht es, bei der Analyse nur bestimmte Wortarten (z. B. Substantive oder Verben) zu berücksichtigen oder Homonyme aufzulösen (van Atteveldt 2008: 45). Während einige Preprocessing-Schritte relativ leicht umzusetzen sind (z. B. Stoppwortentfernung oder Stemming), erfordern andere sprachspezifische Anpassungen oder manuelle Annotation. Zudem ist der Nutzen verschiedener Preprocessing-Schritte keineswegs unumstritten. Zwar ist der Aufwand der automatischen Codierung geringer und die Interpretation der Ergebnisse vielfach einfacher, wenn die Zahl der Features reduziert wird. Gleichzeitig sinkt die Varianz in den Textdaten, so dass ggf. sogar die korrekte Codierung schwieriger wird, weil wichtige semantische Informationen durch starkes Preprocessing verloren gehen (vgl. Scharkow 2012).

Für die eigentliche Analyse werden die extrahierten, bereinigten und ggf. weiter vorbehandelten Features in ein für die Codierung möglichst effektives Format gebracht. Die meisten automatischen Verfahren nutzen dabei entweder die Form der Term-Dokument-Matrix (TDM), in der jede Codiereinheit eine Datenzeile und jedes Feature

eine Datenspalte besetzt (bzw. umgekehrt), oder eine gerichtete Graphenstruktur, in der die Syntax und Semantik einer Aussage abgebildet wird. Während die TDM sich vor allem für komplexe statistische Analysen von großen Datenmengen eignet, benötigen linguistische Analysen zunächst eine flexiblere Datenstruktur, auch wenn für deren Auswertung letztlich ebenfalls ausgezählt wird (van Atteveldt 2008). Obwohl im Folgenden zumeist von automatischen Verfahren der Textanalyse die Rede ist, lassen sich viele statistische Techniken auch auf audiovisuelles Untersuchungsmaterial anwenden. Verallgemeinernd muss dann von einer Objekt-Feature-Matrix gesprochen werden. Bei der automatischen Bilderkennung wären dann etwa Fotos die Objekte und einzelne Bildpunkte oder Farben die zu analysierenden Features.

Wie dieser Abschnitt gezeigt hat, hängt die Effektivität von automatischen Verfahren nicht nur von der eigentlichen Codierung ab, sondern vor allem auch von der Frage, ob sich Datenerhebung und -bereinigung zuverlässig automatisieren lassen. In der Vergangenheit wurden die wenigsten automatischen Inhaltsanalysen tatsächlich von der Datenerhebung bis hin zur Auswertung ohne manuelle Teilschritte durchgeführt. Wenn jedoch Untersuchungseinheiten manuell gesichert oder bereinigt werden müssen, geht damit nicht nur ein Verlust an Reliabilität und Transparenz einher, auch die realisierbare Stichprobengröße liegt zumeist deutlich unter dem, was mit automatischen Verfahren möglich wäre.

4 Automatische Codierung

Auch wenn die Erhebung und Vorbehandlung maschinenlesbarer Inhalte häufig genauso aufwändig ist, bleibt doch die automatische Codierung das Herzstück automatischer Inhaltsanalysen. Mit der Wahl des Analyseverfahrens, der Kategorien und Parameter der Codierung steht und fällt die Validität der Ergebnisse. Stärker noch als bei manuellen Inhaltsanalysen, wo die Codiererinnen und Codierer ggf. eigenständig Fehler im Codebuch korrigieren oder unvollständige Codieranweisungen erweitern, gilt bei der automatischen Codierung die alte Maxime des „garbage in, garbage out". Die Validität einzelner Verfahren kann nur durch umfangreiche Prüfung vor, während und nach der Codierung gewährleistet werden. Zudem müssen vor allem die Ergebnisse vollautomatischer Analyseverfahren sorgfältig und kritisch interpretiert werden, weil der Computer weder die Inhalte der Analyse noch die Überlegungen bei der Operationalisierung berücksichtigt. Es bleibt Aufgabe der Forschenden nachzuweisen, dass sich aus der uni- oder bivariaten Analyse von Worthäufigkeiten, die vielen vollautomatischen Verfahren der Textanalyse zugrunde liegt, Inferenzen auf Autorenschaft, Stil, Genre oder Lesbarkeit von Dokumenten ziehen lassen. Gerade dies sind traditionelle Anwendungsfelder textstatistischer Verfahren, wie sie schon seit den 50er Jahren in den Sozialwissenschaften verwendet werden.

Die Textstatistik geht davon aus, dass sich aus der Häufigkeit der Verwendung bestimmter Wörter Aussagen über die Kommunikate, deren Urheber und deren Rezeption treffen lassen. Da Computer nachweislich schneller und zuverlässiger zählen können als Menschen, bedient sich die textstatistische Forschung seit Jahrzehnten praktisch ausschließlich automatischer Verfahren. Besonders häufig werden Textstatistiken in der Stilometrie und Autorenschaftsforschung eingesetzt (Grieve 2007; Holmes 1998), da sich über die Verwendung bestimmter Wörter ein relativ klarer „Fingerabdruck" einer Autorin oder eines Autors generieren lässt. Auch in den Sozialwissenschaften wird dieses Prinzip genutzt, etwa von Landmann und Züll (2008), die aktuelle Ereignisse in der Berichterstattung automatisch identifizieren können, weil bestimmte Schlüsselwörter häufiger als erwartet vorkommen. Ein zweites Anwendungsfeld für textstatistische Verfahren ist die Lesbarkeits- und Verständlichkeitsforschung, die davon ausgeht, dass bestimmte Textindikatoren (z. B. Wort- und Satzlänge, Umfang des Vokabulars) die Komplexität und damit Verständlichkeit eines Textes gut vorhersagen können. Kercher (2010) nutzt beispielsweise textstatistische Verfahren zur Analyse politischer Kommunikation und deren potenzielle Wirkung auf die Rezipientinnen und Rezipienten.

Bei der explorativen Analyse von Texten ist oft nicht nur die einfache Häufigkeit einzelner Wörter von Interesse, sondern das gemeinsame Auftreten, d. h. *Co-Occurrence* bestimmter Begriffe. Im Prinzip handelt es sich also um die bivariate Erweiterung der einfachen Wortstatistik. Die Co-Occurrence-Analyse basiert auf der Annahme, dass kognitiv bzw. semantisch zusammenhängende Begriffe auch räumlich nahe beieinander stehen. Betrachtet man die Wörter innerhalb eines spezifizierten Rahmens, etwa in kompletten Sätzen oder Dokumenten, lässt sich das gemeinsame Auftreten (Kollokation) bestimmter Begriffe in eine Kontingenztabelle oder eine Ähnlichkeitsmatrix überführen (Galliker & Herman 2003). Diese Ähnlichkeitsmatrizen lassen sich wiederum mit statistischen Verfahren wie Clusteranalyse oder multidimensionaler Skalierung verdichten und visualisieren (Landmann & Züll 2004). Wie bei jeder explorativen Faktoren- oder Clusteranalyse besteht bei Co-Occurrence-Analysen das Problem, dass lediglich die Anzahl der Cluster bzw. Faktoren ex ante von den Forschenden bestimmt werden kann und die rein statistisch gebildeten Textdimensionen inhaltlich nicht immer interpretierbar sind. Obwohl das Verfahren selbst vollautomatisch abläuft, waren in vielen publizierten Co-Occurrence-Studien zudem so viele manuelle Vorbereitungen nötig, z. B. in der Auswahl der relevanten Wörter, dass man kaum von einem vollautomatischen Ansatz sprechen kann. Nichtsdestotrotz ist die Co-Occurrence-Analyse ein Standardwerkzeug in der sozialwissenschaftlichen Forschung, für das es eine große Zahl fertiger Software-Lösungen gibt und das bei der Exploration von großen Textmengen wertvolle Dienste leisten kann.

Die Informationen, die in der Term-Dokument-Matrix enthalten sind, lassen sich nicht nur nutzen, um das gemeinsame Auftreten von einzelnen (Text-)Merkmalen zu analysieren, sondern auch, um die Ähnlichkeiten zwischen Codiereinheiten zu bestim-

men. Dies lässt sich für die vollautomatische Kategorisierung von Dokumenten nutzen, die in der Literatur als *Document Clustering* bezeichnet wird (Grimmer & King 2011). Dem Document Clustering liegt die Annahme zugrunde, dass Dokumente, in denen die gleichen Features vorkommen, thematisch ähnlich sind. Um die Distanz zwischen zwei Dokumenten zu bestimmen, werden deren Term-Vektoren, d. h. die Zeilen der Term-Dokument-Matrix, miteinander in Beziehung gesetzt. Als Distanzmaß wird dabei häufig der Kosinus oder der Jaccard-Koeffizient eingesetzt, da diese relativ unabhängig von der Textlänge, d. h. der Anzahl relevanter Features, funktionieren. Die dabei entstehende Distanzmatrix der Dokumente kann anschließend als Ausgangspunkt für verschiedene clusteranalytische Verfahren eingesetzt werden. Wie bei der Co-Occurrence-Analyse lässt sich auch beim Document Clustering nur die Anzahl der Cluster festsetzen, entweder vor der Analyse (bei partitionierenden Verfahren) oder im Nachhinein (bei hierarchisch-agglomerativen Verfahren).

Eine Erweiterung der vollautomatischen Dokumentklassifikation stellen sog. *Topic Models* dar, bei denen jedes Dokument nicht genau einem Cluster zugeordnet wird, sondern in verschiedenen Anteilen zu mehreren Themengruppen gehören kann. Dies ist in vielen Anwendungsgebieten der Inhaltsanalyse der Fall: Eine Nachricht kann sich auf die Themen Sport *und* Wirtschaft beziehen, ein Film kann in mehrere Genres passen, in einer Parlamentssitzung können viele Politikfelder behandelt werden (Quinn, Monroe, Colaresi, Crespin & Radev 2010). Da sowohl beim Document Clustering als auch bei Topic Models die Gruppen vollautomatisch generiert werden, müssen diese nicht nur interpretiert, sondern ggf. auch nachträglich validiert werden, um auszuschließen, dass lediglich die Stichprobenkomposition oder einige Parameterwerte für die Ergebnisse verantwortlich sind. Grimmer und Stewart (2012) stellen in ihrem Überblicksbeitrag verschiedene Varianten der unüberwachten Dokumentklassifikation sowie entsprechende Validierungsverfahren vor.

Viele der oben genannten vollautomatischen Verfahren lassen sich bei der Analyse kombinieren: Mittels textstatistischer Verfahren können z. B. überzufällig häufige Wörter aus dem Untersuchungsmaterial herausgefiltert werden, die dann den Ausgangspunkt für eine Co-Occurrence-Analyse bilden. Ein anderer häufiger Anwendungsfall in der Forschung ist die Verdichtung von verschiedenen zusammengehörigen Begriffen zu sog. Latenten Semantischen Indizes (LSI), die dann in einem zweiten Schritt für die Berechnung von Dokumentähnlichkeiten genutzt werden. Dieses Vorgehen reduziert die Anzahl der Features von vielen tausend auf vielleicht wenige hundert Indizes, was gleichzeitig den Speicherbedarf und die Laufzeit des Document Clusterings erheblich verringert.

Trotz der technischen Fortschritte der letzten Jahrzehnte sind alle vollautomatischen Verfahren bis heute ausschließlich zur Exploration einzusetzen, sozialwissenschaftliche Inhaltsanalysen sind jedoch in der Regel hypothesengeleitet. Grimmer und Stewart (2012) weisen in diesem Zusammenhang ausdrücklich darauf hin, dass auch eine elaborierte Interpretation und Validierung unüberwachter Klassifikationsergebnisse nicht

mit einer hypothesengeleiteten Kategorisierung von Dokumenten vergleichbar ist. Schon seit den 1960er Jahren hat es daher in den Sozialwissenschaften eine Präferenz für Verfahren gegeben, bei denen ähnlich wie bei der manuellen Inhaltsanalyse codiert wird, d. h. unterschiedliche Kategorien regelgeleitet den Untersuchungseinheiten zugeordnet werden. Hier sind überwachte Verfahren notwendig, bei denen die Forschenden der Software Regeln oder Beispiele vorgeben, nach denen dann die automatische Analyse durchgeführt wird. Die Anwendungsmöglichkeiten für eine solche (halb-)automatische Lösung hängen dementsprechend vom Aufwand für die Regelspezifikation oder die manuelle Codierung von Texten ab. Es muss daher stets ein Kompromiss zwischen dem Operationalisierungsaufwand und Umfang der manuellen Vorarbeiten gefunden werden.

Das älteste und bekannteste überwachte Analyseverfahren, jahrzehntelang praktisch synonym mit automatischer Inhaltsanalyse verwendet, ist die Wörterbuch- bzw. diktionärbasierte Codierung. Seit über 50 Jahren ist das Grundprinzip dieses Ansatzes, der erstmals in der Software *General Inquirer* (Stone, Dunphy, Smith & Ogilvie 1966) umgesetzt wurde, praktisch unverändert: Vor der eigentlichen Codierung wird von den Forschenden ein Kategoriensystem entwickelt, in dem jeder Kategorie im Codebuch einzelne Wörter (oder andere Features) zugewiesen werden, die als Indikatoren für das interessierende Konstrukt dienen. Mit der Analysesoftware kann dann problemlos nach den entsprechenden Features in der Term-Dokument-Matrix gesucht werden, die sie enthaltenden Dokumente bzw. Textabschnitte werden entsprechend klassifiziert. Auf diese Weise ist es möglich, eine sehr große Anzahl an Dokumenten schnell und zuverlässig zuvor festgelegten Kategorien zuzuordnen. Aufgrund des vollständig deterministischen Codiervorgangs ist aber bei diktionärbasierten Verfahren kein Raum für unscharfe Kategorien, Doppeldeutigkeiten und Kontextfaktoren, die jedoch natürliche Sprache erst auszeichnen. Bei jeder Kategorie im Diktionär muss im Rahmen der Instrumententwicklung ausführlich geprüft werden, ob die darin verzeichneten Wörter (n-Gramme höherer Ordnung werden fast nie berücksichtigt) gleichzeitig trennscharf und vollständig sind. Während der diktionärbasierte Ansatz für spezielle Begriffe, etwa Eigen- oder Markennamen im Rahmen einer Medienresonanzanalyse (Raupp & Vogelgesang 2009), mit geringem Aufwand zu validen Ergebnissen führt, gestaltet sich die wortbasierte Codierung komplexer Konstrukte zunehmend schwierig. Da aber die meisten Fragestellungen nicht auf Wort-, sondern auf thematischer Ebene vorliegen, wird ein diktionärbasiertes Verfahren allein durch das Vorhandensein von Rechtschreibfehlern und Homonymen weniger valide Ergebnisse produzieren als eine gute manuelle Codierung. Zudem ist für viele theoretische relevante Konstrukte nicht ohne weiteres eine Wortliste ersichtlich, die tatsächlich zuverlässig und valide ist. Die Erwartung, dass sich durch die Verwendung standardisierter Wörterbücher automatische Inhaltsanalysen effektiver und effizienter durchführen lassen, hat sich letztlich nicht erfüllt. Es gibt nur wenige allgemein verwendbare Diktionäre, erst recht in deutscher Sprache, so dass letztlich viele Forschende projektspe-

zifische Wörterbücher entwerfen müssen (Scharkow 2012: 80). Dieser Aufwand ist mitunter größer als bei manuellen oder auch anderen automatischen Verfahren (s. u.), so dass diktionärbasierte Verfahren oft erst dann ihre Vorteile ausspielen können, wenn wirklich große Datenmengen, d. h. zumindest einige hundert oder tausend Beiträge, einheitlich zu verarbeiten sind.

Eng mit der Diktionärcodierung verwandt sind regelbasierte Verfahren, bei denen neben der reinen Wortebene auch zusätzliche syntaktische und/oder semantische Informationen genutzt werden. Im Gegensatz zu rein thematischen Analysen, wie sie zumeist mit Wörterbuchverfahren durchgeführt werden, lassen sich mit linguistisch-regelbasierten Ansätzen nicht nur Dokumente danach klassifizieren, ob in ihnen ein Akteur oder ein Thema vorkommt, sondern auch, in welcher Beziehung verschiedene Akteure zu anderen stehen. Viele regelbasierte Verfahren nutzen weiterhin Diktionäre, etwa Akteurs- oder Ortslisten, zur Codierung und nutzen diese Informationen dann in Kombination mit POS-Tagging (Markierung von Wortarten) oder syntaktischem Parsing (Zerlegung von Aussagen) für die Detailanalyse. Seit vielen Jahrzehnten wird diese Methodenkombination mit der Software KEDS/TABARI für die automatische Verarbeitung von Tickermeldungen eingesetzt, aus denen Informationen über internationale Ereignisse extrahiert werden (Schrodt 2011). Die verwendeten computerlinguistischen Algorithmen sind allerdings vergleichsweise primitiv *(shallow parsing),* hochgradig themenspezifisch und nur für englische Texte zu verwenden. Einen deutlich anspruchsvolleren Ansatz verfolgen seit den 1980er Jahren die Entwickler der CETA-Systems (van Cuilenburg, Kleinnijenhuis & de Ridder 1988), das im Prinzip Osgoods (1959) Evaluative Assertion Analysis mit Verfahren der Computerlinguistik automatisieren soll. Obwohl es hier in den letzten Jahren große Fortschritte vor allem im Bereich des syntaktischen Parsings gegeben hat, ist das Verfahren nach wie vor mit manueller Vorcodierung, der Definition umfangreicher Akteurs- und Handlungslisten sowie Programmierarbeit verbunden. Die Ergebnisse der Studien von van Atteveldt (2008) zeigen das große Potential der semantischen Netzwerkanalyse – kein anderes automatisches Verfahren kann bislang für detaillierte Frame- oder Claimanalysen eingesetzt werden –, aber auch den damit verbundenen Aufwand. Dabei wird ein generelles Problem der diktionär- und regelbasierten Codierung deutlich: Da sowohl das Wörterbuch als auch die Parsing-Regeln sprachspezifisch sind, ist der Aufwand für mehrsprachige Analysen zumeist deutlich höher als bei manuellen Analysen. Zudem erfordert sowohl die Regelspezifikation als auch die Definition der Wortlisten umfangreiche Anpassungen an die zu analysierende Textart und den Themenkontext der Analyse. Ein Transfer in andere Themengebiete oder Sprachen ist daher äußerst schwierig.

Ein grundsätzlicher Nachteil der bislang vorgestellten automatischen Verfahren liegt in der Tatsache, dass gerade der für die Inhaltsanalyse zentrale Prozess der Operationalisierung und Codierung sich deutlich von der manuellen Analyse unterscheidet, d. h. man wenig auf bewährte Entscheidungsregeln oder Vorerfahrungen aus manuellen Analysen aufbauen kann. Der Prozess der Diktionärentwicklung oder der Implementie-

rung komplexer Regeln für syntaktisch-semantische Parser hat nur wenig mit der traditionellen Codebuchentwicklung, der Codierschulung und der manuellen Feldarbeit gemeinsam. Dies mag einer der Gründe sein, warum sich automatische und konventionelle Inhaltsanalyse in den letzten Jahrzehnten eher auseinanderentwickelt als gegenseitig befruchtet haben. Wer sich für die automatische Codierung entscheidet, muss dies relativ frühzeitig im Forschungsdesign berücksichtigen. Umgekehrt ist es bei den o. g. Verfahren nur selten möglich, das bei der manuellen Analyse gewonnene Wissen für die automatische Codierung nutzbar zu machen. Neuere Ansätze aus dem maschinellen Lernen *(supervised learning)* versprechen genau dies möglich zu machen, indem klassisch manuell codierte Inhalte als Trainingsmaterial für einen statistischen Lernalgorithmus verwendet werden. Im Gegensatz zu den streng deduktiv-deterministischen Regel- und Wörterbuchverfahren werden mit der Software hier nicht fertige Entscheidungsregeln umgesetzt, sondern diese aus den korrekt codierten Daten gewonnen. Dies hat gleich mehrere Vorteile (vgl. ausführlich Scharkow 2012): (1) Das Verfahren ist grundsätzlich sprach- und themenunabhängig, solange die Trainingsdaten konsistent sind. (2) Es ist so gut wie kein Preprocessing und keine manuelle Anpassung der Codiersoftware nötig. (3) Wenn manuell vorcodierte digitale Inhalte vorliegen, kann mit den Dokumenten und Codes der maschinelle Klassifikationsalgorithmus trainiert und evaluiert werden, ohne dass weiterer Aufwand entsteht. (4) Der Klassifikationsalgorithmus kann ähnlich wie menschliche Codierende geschult und mittels Reliabilitätstests geprüft werden. Die Qualitätsmaßstäbe und Evaluationsverfahren sind dabei größtenteils identisch mit denen der manuellen Inhaltsanalyse.

Selbstverständlich ist die überwachte Klassifikation nicht in jeder Situation besser als andere automatische Verfahren: Das Verfahren ist rein statistisch, d. h. die Codierung erfolgt ohne Rücksicht auf Syntax und Semantik. Inhaltsanalysen, bei denen vor allem das Kontextwissen der Codierenden und weniger der tatsächliche Textgehalt die Codierung beeinflussen, etwa bei den Nachrichtenfaktoren Überraschung oder Prominenz, lassen sich auch mit maschinellem Lernen nicht umsetzen. Zudem kann die maschinelle Codierung nie zuverlässiger und valider sein als die manuell codierten Trainingsdaten. Da jedoch auch heute noch klassische Themenfrequenzanalysen die Grundlage vieler Forschungsprojekte sind, und diese sich mit geringem Aufwand und hoher Erfolgschance per maschinellem Lernen automatisieren lassen (Scharkow 2012), besitzt das Verfahren gerade für die angewandte Kommunikationsforschung ein großes Potential. Hinzu kommt die Tatsache, dass in der Informatik die Entwicklung neuer Lern- und Klassifikationsalgorithmen eines der wichtigsten Forschungsfelder darstellt, von der die überwachte Textklassifikation enorm profitieren kann. Da die statistischen Klassifikatoren ohnehin nur auf Basis von Zahlenwerten operieren, lassen sich die verschiedenen Algorithmen leicht austauschen, kombinieren *(Ensemble Classification)* und auf unterschiedliches Untersuchungsmaterial anwenden. Letztlich basiert die Gesichtserkennung auf Facebook, die automatische Mimik-Codierung oder die Klassifikation von Musikstücken auf denselben statistischen Algorithmen wie die überwachte Textco-

dierung. Sie unterscheiden sich lediglich in der Verwendung unterschiedlicher Features, d. h. statt Wörtern etwa Tonfolgen oder Farbmuster.

Dieser Abschnitt sollte zeigen, dass die automatische Inhaltsanalyse deutlich variantenreicher ist, als man nach der Lektüre der klassischen Lehrbücher annehmen sollte. Auch wenn einige Verfahren seit Jahrzehnten kaum methodisch weiterentwickelt wurden, bedeutet das nicht, dass diese durch neuere Ansätze verdrängt wurden oder gänzlich obsolet sind. Vollautomatische Verfahren sind vor allem für die Exploration von großen Mengen an Untersuchungseinheiten von Nutzen, die man nicht einzeln inspizieren und nach Mustern durchsuchen kann. Sie sind aber für sich genommen nur in den seltensten Fällen der Kern einer Inhaltsanalyse. Überwachte Verfahren können vielseitig als Alternative und auch Ergänzung zu manuellen Inhaltsanalysen eingesetzt werden, da sie zumindest den grundsätzlich gleichen Anspruch erheben, nämlich eine große Anzahl an Dokumenten hinsichtlich spezifischer inhaltlicher Aspekte zu codieren. Jedes Verfahren hat dabei individuelle Stärken und Schwächen: Diktionärbasierte Ansätze eignen sich insbesondere für die zuverlässige und schnelle Codierung von einfachen Kategorien, die mittels einer überschaubaren Wortliste definiert werden können. Gerade im Bereich der kommerziellen Medienresonanzforschung ist dies häufig der Fall, weshalb dort Wörterbuchverfahren einen großen Stellenwert haben. Die semantische Netzwerkanalyse ist ein vergleichsweise komplexes Verfahren, das grundsätzlich den Anspruch verfolgt, Aussagen weitestgehend maschinell zu codieren und weiterzuverarbeiten. Damit ist sie am ehesten mit konventionellen Ansätzen wie der Argumentationsanalyse oder Frühs Semantischer Struktur- und Inhaltsanalyse verwandt, die bislang in der Kommunikationswissenschaft eher randständig sind (Früh 2007: 270 ff.). Wenn jedoch das Interesse an aussagebasierten Inhaltsanalysen unterhalb der Beitragsebene weiter zunimmt, wie dies gerade bei der Frame-Analyse der Fall zu sein scheint, ist die semantische Netzwerkanalyse eine vielversprechende Alternative zur manuellen Codierung. Die relativ neuen Verfahren aus dem maschinellen Lernen sind besonders als Ergänzung und ggf. Weiterführung manueller Inhaltsanalysen geeignet. Zudem erfordern sie in der Anwendung nur wenig computerlinguistische und statistische Kenntnisse. Neben der Themenanalyse eignet sich das Verfahren auch für die Analyse von Bewertungen (*Sentiment Analysis,* Pang & Lee 2008), wobei in diesem Anwendungsfeld auch diktionärbasierte Verfahren nach wie vor relevant sind.

5 Fazit und Ausblick

Im kommunikationswissenschaftlichen Forschungsalltag kommen automatische Verfahren der Inhaltsanalyse noch immer recht selten vor. Ein Grund dafür sind sicher die relativ schlechten Erfahrungen, die in der Vergangenheit mit traditionellen Einwortanalysen gemacht wurden. Obwohl diese Skepsis, die sich in fast allen Lehrbüchern zur Inhaltsanalyse findet, durchaus nachvollziehbar ist, zeigen neuere Arbeiten zu über-

wachten und unüberwachten Verfahren, dass es durchaus möglich ist, mit vertretbarem Aufwand zu validen Ergebnissen zu kommen. Dies bedeutet nicht, dass automatische Verfahren immer schneller und preiswerter als manuelle Analysen sind – und das müssen sie auch nicht sein, da sie auch andere Vorzüge haben. So sind automatische Inhaltsanalysen in aller Regel deutlich leichter zu dokumentieren und zu replizieren, da ein Computer unter gleichen Bedingungen immer zu gleichen Ergebnissen kommt. Dies ist auch bei der Datenerhebung und -bereinigung ein zentraler Vorteil automatischer Verfahren: Wer in größerem Umfang Mitteilungen aus dem Internet herunterladen und codieren will, wird früher oder später zu Softwarelösungen greifen, zumal der Reliabilitätsvorteil automatischer Datenerhebung mit steigendem Stichprobenumfang noch zunimmt. Schon jetzt werden viele automatische Verfahren ganz selbstverständlich für Arbeitsschritte eingesetzt, die zuvor manuell erledigt wurden, vor allem im Bereich der Datenerhebung und Auswertung.

Bei der eigentlichen Codierarbeit ist es angesichts der Vielzahl möglicher Ansätze und Fragestellungen schwierig, generelle Empfehlungen für oder gegen automatische Codierverfahren zu geben. Unstrittig ist, dass die Quantität der Codierungen sich nur mit Hilfe des Computers steigern lässt, und es gibt gute Gründe davon auszugehen, dass die Codierquantität letztlich auch die Codierqualität positiv beeinflusst (vgl. ausführlich Scharkow 2012: Kap. 2). In jedem Fall ist es nötig, die Validität jedes Ansatzes für die eigene Forschung selbst zu überprüfen.

Um die Validität automatischer Verfahren bewerten zu können, muss man zudem ihren Einsatzzweck berücksichtigen: einerseits als Ersatz für manuelle Arbeit, andererseits als genuin neues Verfahren, das bislang gar nicht verwendet wurde. Im ersten Fall muss man kritisch betrachten, ob und wie stark operationale Veränderungen gegenüber der manuellen Referenz notwendig sind, um eine Codierung automatisch durchzuführen. Hier hat sich in der Vergangenheit gezeigt, dass die Validität der Messung tendenziell sinkt, je weiter sich ein automatisches Verfahren von der Logik der manuellen Referenzcodierung entfernt. Man geht daher zumeist von einer relativ schlechteren Validität automatischer Verfahren aus (Früh 2007; Rössler 2010). In diesen Fällen ist stets abzuwägen, ob die ggf. größere Skalierbarkeit der Analyse den Verlust an Validität ausgleichen kann. Bei genuin automatischen Verfahren, etwa im Bereich der explorativen Textanalyse, fehlt zumeist ein klassischer Vergleichsmaßstab. Hier ist vor allem die Kriteriums- oder prognostische Validität gefragt, d. h. ob das Verfahren Ergebnisse hervorbringt, die mit textexternen Merkmalen oder Expertenurteilen in Einklang zu bringen sind.

Wenn sich nachweisen lässt, dass ein automatisches Verfahren mit einiger Sicherheit ähnlich zuverlässig und valide funktioniert wie ein manuelles, wird zukünftig kaum jemand an der manuellen Arbeit festhalten. Dies gilt umso mehr, wenn das automatische Verfahren zuverlässiger als Handarbeit ist, was bei den vielen eher handwerklichen Aufgaben im Rahmen einer Inhaltsanalyse nicht selten vorkommt. Ob die automatische Codierung von Texten, aber auch Bildern und audiovisuellen Medien, in naher Zukunft

zu einem Standardwerkzeug der Kommunikationswissenschaft werden wird, ist bislang ungewiss. Obwohl die technischen Möglichkeiten stetig weiterentwickelt werden, wird der Erfolg automatischer Verfahren davon abhängen, ob sie gewinnbringend in den Forschungsalltag zu integrieren sind und einen eigenständigen methodischen Beitrag zur Inhaltsanalyse erbringen können.

Literaturtipps

Manning, Christopher D. & Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Popping, Roel (2000). *Computer-Assisted Text Analysis.* London: Sage.

Scharkow, Michael (2012). *Automatische Inhaltsanalyse und maschinelles Lernen.* Berlin: epubli.

Literatur

Carstensen, Kai-Uwe, Ebert, Christian, Ebert, Cornelia, Jekat, Susanne, Langer, Hagen & Klabunde, Ralf (2009). *Computerlinguistik und Sprachtechnologie: Eine Einführung.* Heidelberg: Springer.

Cutler, Ross & Davis, Larry (2002). Look who's talking: Speaker detection using video and audio correlation. *Proceedings of the 2000 IEEE International Conference on Multimedia, vol. 3* (S. 1589–1592). IEEE.

Deacon, David (2007). Yesterday's Papers and Today's Technology: Digital Newspaper Archives and ‚Push Button' Content Analysis. *European Journal of Communication, 22*(1), 5–25.

DeWeese, L. Carroll (1977). Computer Content Analysis of „Day-Old" Newspapers: A Feasibility Study. *Public Opinion Quarterly, 41*(1), 91–94.

Früh, Werner (2007). *Inhaltsanalyse: Theorie und Praxis.* Konstanz: UVK.

Galliker, Mark & Herman, Jan (2003). Inhaltsanalyse elektronisch gespeicherter Massendaten der internationalen Presse. *Zeitschrift für Medienpsychologie, 15*(3), 98–105.

Grieve, Jack (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22*(3), 251–270.

Grimmer, Justin & King, Garry (2011). General Purpose Computer-assisted Clustering and Conceptualization. *Proceedings of the National Academy of Sciences, 108*(7), 2643–2650.

Grimmer, Justin & Stewart, Brandon (2012). *Text as data: The promise and pitfalls of automatic content analysis methods for political texts*. http://www.stanford.edu/~jgrimmer/tad2.pdf

Hillard, Dustin, Purpura, Stephen & Wilkerson, John (2007). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics, 4*(4), 31–46.

Holmes, David (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*(3), 111–117.

Hotho, Andreas, Nürnberger, Andreas & Paaß, Gerhard (2005). A Brief Survey of Text Mining. *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, 20*(1), 19–62.

Kercher, Jan (2010). Zur Messung der Verständlichkeit deutscher Spitzenpolitiker anhand quantitativer Textmerkmale. In Thorsten Faas, Kai Arzheimer & Sigrid Roßteutscher (Hrsg.), *Information – Wahrnehmung – Emotion: Politische Psychologie in der Wahl- und Einstellungsforschung* (S. 97–121). Wiesbaden: VS.

Landmann, Juliane & Züll, Cornelia (2004). Computerunterstützte Inhaltsanalyse ohne Diktionär? Ein Praxistest. *ZUMA-Nachrichten, 54,* 117–140.

Landmann, Juliane & Züll, Cornelia (2008). Identifying Events Using Computer-Assisted Text Analysis. *Social Science Computer Review, 26*(4), 483–497.

Monroe, Burd L. & Schrodt, Philip A. (2008). Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis, 16*(4), 351– 355.

Osgood, Charles (1959). The representational model and relevant research methods. In Iithiel de Sola Pool (Hrsg.) *Trends in content analysis* (S. 33–88). Urbana: University of Illinois Press.

Pang, Bo & Lee, Lilian (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1–135.

Quinn, Kevin M., Monroe, Burd L., Colaresi, Michael, Crespin, Michael H. & Radev, Dragomir R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science, 54*(1), 209–228.

Raupp, Juliana & Vogelgesang, Jens (2009). *Medienresonanzanalyse: Eine Einführung in Theorie und Praxis*. Wiesbaden: VS.

Roberts, Carl W. (2000). A Conceptual Framework for Quantitative Text Analysis. *Quality and Quantity, 34*(3), 259–274.

Rössler, Patrick (2010). Das Medium ist nicht die Botschaft. In Martin Welker & Carsten Wünsch (Hrsg.), *Die Online-Inhaltsanalyse* (S. 31–43). Köln: von Halem.

Rössler, Patrick & Wirth, Werner (2001). Inhaltsanalysen im World Wide Web. In Werner Wirth & Edmund Lauf (Hrsg.), *Inhaltsanalyse. Perspektiven, Probleme, Potentiale* (S. 280–302). Köln: von Halem.

Scharkow, Michael (2010). Lesen und lesen lassen. Zum State of the Art automatischer Textanalyse. In Martin Welker & Carsten Wünsch (Hrsg.), *Die Online-Inhaltsanalyse* (S. 340–364). Köln: von Halem.

Scharkow, Michael (2012). *Automatische Inhaltsanalyse und maschinelles Lernen.* Berlin: epubli.

Schrodt, Philip (2011). *Automated high-volume production of near-real time event data.* Paper presented at the New Methodologies Conference, Princeton University, February 2011.

De Sola Pool, Iithiel (1959). *Trends in content analysis: Papers of the Work Conference on Content Analysis of the Committee on Linguistics and Psychology.* Urbana: University of Illinois Press.

Stone, Philip (1997). Thematic text analysis: New agendas for analyzing text content. In Carl W. Roberts (Hrsg.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (S. 35–54). Mahwah: Lawrence Erlbaum Associates.

Stone, Philip, Dunphy, Dexter, Smith, Marshall & Ogilvie, Daniel (1966). *The general inquirer: A computer approach to content analysis.* Cambridge: The MIT Press.

van Atteveldt, Wouter (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content.* Charleston: BookSurge Publishers.

van Cuilenburg, Jan J., Kleinnijenhuis, Jan & de Ridder, Jan A. (1988). Artificial Intelligence and Content Analysis. *Quality and Quantity, 22*(1), 65–97.

West, Mark (2001). The future of computer content analysis: trends, unexplored lands, and speculations. In Mark West (Hrsg.), *Theory, method, and practice in computer content analysis, vol. 16,* (S. 159–75). Westport: Greenwood.

Wirth, Werner (2001). Der Codierprozeß als gelenkte Rezeption. Bausteine für eine Theorie des Codierens. In Werner Wirth & Edmund Lauf (Hrsg.), *Inhaltsanalyse: Perspektiven, Probleme, Potentiale* (S. 157–182). Köln: von Halem.

Xu, Min, Maddage, Namunu, Xu, Changsheng, Kankanhalli, Mohan & Tian, Qi (2003). Creating audio keywords for event detection in soccer video. *Proceedings of the 2003 International Conference on Multimedia, vol. 2*. IEEE.

Züll, Cornelia & Alexa, Melina (2001). Automatisches Codieren von Textdaten. Ein Überblick über neue Entwicklungen. In Werner Wirth & Edmund Lauf (Hrsg.), *Inhaltsanalyse – Perspektiven, Probleme, Potenziale* (S. 303–317). Köln: von Halem.